1 2 9 0

UNIVERSIDADE Đ
COIMBRA

Rita Maria Vieira Carvalho Marques

Automatic Segmentation of the Optic Nerve Head in Optical Coherence Tomography Data

Thesis submitted to the University of Coimbra in fulfillment of the requirements of the Master's Degree in Biomedical Engineering under the scientific supervision of Ph.D. Professor João Manuel Rendeiro Cardoso, Ph.D. Pedro Guilherme da Cunha Leitão Dias Vaz and Ph.D. Danilo Andrade de Jesus

October of 2021

University of Coimbra

Integrated Master in Biomedical Engineering

Automatic Segmentation of the optic nerve head in Optical Coherence Tomography data

Rita Maria Vieira Carvalho Marques

Thesis submitted to the Faculty of Sciences and Technology of the University of Coimbra in fulfillment of the requirements for the Master's Degree in Biomedical Engineering

Supervisors:
Ph.D. Danilo Andrade De Jesus
Professor João Manuel Rendeiro Cardoso
Ph.D. Pedro Guilherme Vaz

· U C ·

FCTUC FACULDADE DE CIÊNCIAS
E TECNOLOGIA
UNIVERSIDADE DE COIMBRA

Erasmus MC

Coimbra, 2021

Acknowledgements

First of all, I would like to thank Pedro Vaz, Danilo Jesus and Luisa Brea for all the guidance, patience and help, for always being present and pushing me to be and do better. Thank you for this challenging but rewarding past year of constant learning and growth.

I want to thank LIBPhys-UC, particularly professor João Cardoso, and the BIGR from Erasmus MC, notably Stefan Klein and Theo van Walsun, for all the availability and for the opportunity to develop this project and integrate such a rich, diverse and supportive team. I also thank the LCA from the University of Coimbra for providing use of the Navigator Cluster, crucial for the success of this project.

To João Breda, Ingeborg Stalmans and Jan Van Eijgen for sharing your profound knowledge and expertise on ophthalmology, particularly on glaucoma, with me. Thank you for your time and advice. And I would also like to thank Mariana, for the hours spent revising my manual segmentations.

To all my friends, the best I could ask for, the ones I met during these incredible past five years and the ones that have been with me ever since I can remember. Thank you for always being there, motivating me, and being part of memories I will forever cherish. And a very special thanks to Vânia; your friendship was precious during the past year.

To Francisco. Growing, learning and sharing with you has been the most incredible journey. Thank you for all the love and for being who you are.

And last but certainly not least, I want to thank my family, parents, aunts, uncles and cousins. Thank you for always sharing my joys and minimising my sorrows. To my parents, in particular, Maria dos Anjos and Carlos, for being my biggest supporters. Thank you for always believing in me, loving me and for all the sacrifices. I am so grateful for you.

Abstract

Glaucoma is an irreversible but preventable disease, and one of the main causes of blindness worldwide. The ONH represents the intraocular section of the optic nerve, which is prone to damage by increases in the IOP. The advent of OCT has enabled the evaluation of ONH parameters (biomarkers), which have achieved promising results for diagnosis and monitoring of retinal diseases, including glaucoma. Nonetheless, these OCT derived biomarkers are mostly extracted through manual segmentation of the ONH tissues, a time-consuming and prone to bias task that limits their usability in clinical practice. Therefore, the automatic segmentation of ONH in OCT scans could further improve the current clinical management of glaucoma and other diseases.

The work presented in this thesis comprises the development of a deep learning based segmentation model for five structures of the ONH (RNFL, RPE/BM complex, other retinal layers, choroid and LC) in OCT data. The available dataset comprised 23 raster volumes from 13 healthy subjects, and 213 radial volumes from 20 healthy subjects and 46 glaucoma patients. In total, 300 images were manually segmented by two graders for generating a segmentation ground-truth. Different models based on the U-net architecture were trained and evaluated. Cross-validation grid search over a parameter grid was used to choose the best model based on four metrics: overall accuracy and Dice coefficient, sensitivity, and specificity of each structure individually. Six features (optic disc diameter, BMO-MRW, RNFL thickness, RNFL area, LC depth and LCCI) were obtained from the best model predictions. The inter-grader variability was analysed to assess the quality and consistency of the ground truth. A statistical analysis of the biomarkers extracted in both the manual and automatic segmented images was performed to further evaluate the model.

The results showed that the proposed model was able to separate the five ONH tissues, with all metrics above 70% except for the LC for which performs above 63%. The biomarkers extracted from manual and automatic segmentations showed correlations above 0.85, except for the LCCI. The main results of this project suggest that it is possible to extract reliable clinical parameters from automatic segmentations after training an optimized deep learning model based on a simple U-net architecture. Among all clinically relevant structures of the ONH, LC has shown to be the most challenging to automatically segment.

Future work should focus on adding contextual information of this tissue into the model. Larger datasets and more reliable manually segmented data should also be considered for improving the current model. Lastly, differences between the predictions of radial and raster B-scans were observed highlighting the need for standard imaging protocols when evaluating the ONH.

Keywords: Optical Coherence Tomography, Optic Nerve Head, Glaucoma, Automatic Segmentation, Machine Learning, Lamina Cribrosa

Resumo

O glaucoma, uma das principais causas de cegueira no mundo, é uma doença irreversível, mas que pode ser prevenida. A secção intraocular do nervo ótico está particularmente sujeita a danos provocados pelo aumento da pressão intraocular (IOP). O advento da tomografia de coerência ótica (OCT) possibilitou a avaliação de parâmetros do disco ótico (biomarcadores), que se têm vindo a provar promissores para o diagnóstico e monitorização de doenças da retina, tal como o glaucoma. No entanto, estes biomarcadores são obtidos maioritariamente através da segmentação manual das estruturas do disco ótico, um processo demorado e sujeito a erros humanos, limitando o seu uso na prática clínica. Deste modo, a segmentação automática do disco ótico em OCT poderia ajudar a gestão clínica do glaucoma e outras doenças.

O trabalho apresentado nesta dissertação envolve o desenvolvimento de um modelo automático baseado em deep learning para segmentação de cinco estruturas do disco ótico (camada das fibras nervosas da retina (RNFL), complexo epitélio segmentado da retina (RPE)/membrana de Bruch (BM), outras camadas da retina, coróide e lâmina cribrosa (LC)) em dados de OCT. Os dados disponíveis continham 23 volumes de varredura de 13 indivíduos saudáveis e 213 volumes radiais de 20 indivíduos saudáveis e de 46 pacientes com glaucoma. No total, 300 imagens foram segmentadas manualmente por dois especialistas de forma a gerar um *ground truth.* Diferentes modelos baseados na arquitetura U-net foram treinados e avaliados de acordo com uma *grid search.* O melhor modelo foi escolhido com base em quatro métricas: precisão média de todas as segmentações, e coeficiente Dice, sensitividade e especificidade de cada estrutura individualmente. Seis biomarcadores (diâmetro do disco ótico, a espessura da rima neural a partir da abertura da membrana de Bruch (BMO-MRW), espessura da RNFL, área da RNFL, profundidade da LC e índice de curvatura da LC (LCCI)) foram obtidos a partir do melhor modelo. A diferença entre imagens segmentadas por diferentes especialistas foi avaliada para analisar a consistência do *ground truth*. A análise estatística dos biomarcadores extraídos a partir de segmentações manuais e automáticas permitiu uma avaliação mais aprofundada do modelo e das diferenças entre olhos saudáveis e com glaucoma.

Os resultados mostraram que o modelo proposto é capaz de separar os cinco tecidos do disco ótico com métricas acima dos 70% exceto para a LC, a qual obteve métricas acima dos 63%. Os biomarcadores apresentaram correlações acima dos 0.85 entre segmentações manuais e automáticas, exceto para o LCCI. Os principais resultados deste projeto sugerem que é possível extrair parâmetros clínicos de confiança a partir de segmentações automáticas obtidas com um modelo otimizado de *deep learning* baseado numa *U-net*. De entre todas as estruturas do disco ótico, a LC mostrou-se a mais desafiante para segmentar, tanto manualmente como automaticamente.

Para trabalho futuro, deve-se considerar adicionar informações contextuais da LC ao modelo. O modelo também beneficiaria de um maior conjunto de dados e do desenvolvimento de métodos que garantissem a qualidade e consistência das segmentações manuais. Finalmente, as diferenças observadas entre B-scans de varredura e radiais, destacam a necessidade de definir protocolos para a avaliação do disco ótico.

Keywords: Tomografia de Coerência Ótica, Disco Ótico, Glaucoma, Segmentação Automática, Machine Learning, Lâmina Cribrosa

List of Acronyms

BIGR Biomedical Imaging Group Rotterdam

BM Bruch's membrane

BMO Bruch's membrane opening

BMO-MRW Bruch's membrane minimum-rim-width

CDR cup-to-disc ratio

CHUSJ Centro Hospitalar e Universitário São João

CNN convolutional neural network

CNS central nervous system

CSFP cerebrospinal fluid pressure

DRUNET dilated-residual U-Net

EDI enhanced depth imaging

FD-OCT Fourier-domain optical coherence tomography

GAN generative adversarial network

IIH idiopathic intracranial hypertension

ILM inner limiting membrane

IOP intraocular pressure

LC lamina cribrosa

LCA Laboratório de Computação Avançada

LCCD lamina cribrosa curvature depth

LCCI lamina cribrosa curvature index

LCD lamina cribrosa depth

LCT lamina cribrosa thickness

LIBPhys-UC Laboratory for Instrumentation Biomedical Engineering and Radiation Physics - University of Coimbra

LUT look up table

MRF Markov random field

MS multiple sclerosis

MSE mean square error

NMOSD neuromyelitis optica spectrum disorders

NTG normal tension glaucoma

OCT optical coherence tomography

ON optic neuritis

ONH optic nerve head

PCA principal component analysis

POAG primary open angle glaucoma

ReLU rectified linear unit

RGC retinal ganglion cell

RMSE root mean square error

RNFL retinal nerve fiber layer

RPE retinal pigment epithelium

SD-OCT spectral-domain optical coherence tomography

SS-OCT swept-source optical coherence tomography

SVM support vector machine

VF MD visual field mean deviation

List of Figures

List of Tables

Contents

Chapter 1

Introduction

In this chapter, the context and main goals of this thesis project are explained as well as the outline for the rest of the document. Moreover, the research team involved in the project and the scientific contribution resulting from it are presented.

1.1 Motivation

Glaucoma is the leading cause of irreversible blindness worldwide [1]. This disease causes progressive peripheral vision loss which cannot be recovered. Nevertheless, its progression is preventable [2], highlighting the important role an early diagnostic and close monitoring of the disease progression may play in the patients vision preservation and life quality. However, glaucoma can remain asymptomatic until a late stage of the disease and the clinical practice still lacks a standard and automatic diagnostic tool. Additionally, the heterogeneity of the glaucoma portrayal and human subjectivity associated with clinical exams performance and evaluation [3] further difficult the current diagnostic and monitoring.

The primary risk factor of glaucoma development is the increase on intraocular pressure (IOP) [4]. While subtle fluctuations of IOP are normal over the day, drastic increases in this pressure may translate into damage to the retinal ganglion cell (RGC) axons as they exit the eye at the optic nerve head (ONH), and into complex 3D structural modifications in the ONH [5]. Evidence suggests that ONH changes are the first ones to occur in glaucoma disease, making them more relevant for an early diagnosis that could prevent its progression [6].

The lamina cribrosa (LC), a mesh-like structure where the nerve fibers exit the eye, plays an important role on the ONH biomechanics. However, it has been rather inaccessible for a long time due to its deeper location in the ONH, where the optical coherence tomography (OCT) signal suffers more attenuation. Nevertheless, advances in imaging technologies such as enhanced depth imaging (EDI) [7], swept-source optical coherence tomography (SS-OCT) [8] and adaptive compensation [9] have enabled the inclusion of the LC in automatic segmentation algorithms and the evaluation of new ONH parameters, particularly, the LC depth and curvature.

While commercial OCT devices have in-built segmentation software, they segment some, but not all ONH tissues [10], and still require a lot of clinical supervision. Moreover, the only way to get LC parameters in clinical practice at the moment is from manual segmentation. However, manual segmentation is time-consuming and prone to bias, thus limiting its usability in clinical practice. Therefore, a reliable automatic ONH segmentation of OCT scans could further improve the current clinical management of glaucoma and other diseases.

1.2 Objectives

The goal of this dissertation was to develop an algorithm that can automatically segment the ONH in OCT data, and extract ONH features that can be used in clinical practice for glaucoma management. To achieve this goal, the following objectives were set:

1. Gather OCT data from both healthy subjects and glaucoma patients and generate a ground truth of manual segmentations;

2. Develop an automatic segmentation model based on artificial neural networks;

3. Use the ground truth generated on 1. to train and validate the segmentation model

4. Develop methods to compute clinically relevant biomarkers automatically from the segmentation of the ONH on OCT B-scans;

5. Examine differences between biomarkers extracted from manual segmentations and automatic segmentations predicted using 2.;

6. Infer differences between healthy subjects and glaucoma patients from ONH biomarkers.

1.3 Thesis content

This document is divided in six chapters:

- Chapter 1 - Introduction: comprises the main goals and motivation of this project, as well as the research team involved in its development, and the list of scientific contributions.
- Chapter 2 - Theoretical background: contains a brief presentation of the eye anatomy, glaucoma disease, and OCT imaging, followed by the current state of the art in automatic segmentation of the ONH in OCT data, and the definitions and reference values of the clinically relevant biomarkers that can be extracted from these segmentations.
- Chapter 3 - Methods: the development of the automatic segmentation algorithm is presented, starting with the description of the dataset and the manual segmentation, followed by a description of the segmentation model. Finally, the computation methods for automatic biomarkers extraction are described.
- Chapter 4 - Results: includes the different experiments performed to achieve the best segmentation model configuration. The biomarkers' results are analysed and their accuracy and discriminating power evaluated.
- Chapter 5 - Discussion: analysis and discussion of the main finding presented on Chapter 4. The results are associated with the main objectives of the project and with the state of the art.
- Chapter 6 - Conclusion: comprises the general conclusions of this work and the future possible improvements.

1.4 Project team

The research team behind the development of this masters thesis project is presented in Table 1.1. This project arose from a partnership between the Laboratory for Instrumentation Biomedical Engineering and Radiation Physics - University of Coimbra (LIBPhys-UC) , the Biomedical Imaging Group Rotterdam (BIGR) from Erasmus MC, the Department of Ophthalmology from Centro Hospitalar e Universitário São João (CHUSJ) and the Research Group of Ophthalmology from the

Department of Neurosciences in KU Leuven. This partnership gathered a multidisciplinary team from which the project benefited.

This project also granted the possibility to integrate the Erasmus + program, for a two month remote internship at Erasmus University Medical Center.

Name	**Role**	**Institution**
Danilo Andrade De Jesus	Technical supervisor	BIGR, Erasmus MC, Netherlands
Luisa Sánchez Brea	Technical supervisor	BIGR, Erasmus MC, Netherlands
Stefan Klein	Technical advisor	BIGR, Erasmus MC, Netherlands
Theo van Walsum	Technical advisor	BIGR, Erasmus MC, Netherlands
João Breda	Clinical supervisor	CHUSJ, Portugal
Mariana Dias	Clinical advisor	CHUSJ, Portugal
João Manuel Rendeiro Cardoso	Technical supervisor	LIBPhys-UC, Portugal
Pedro Vaz	Technical supervisor	LIBPhys-UC, Portugal
Ingeborg Stalmans	Clinical advisor	Research Group Ophthalmology, KU Leuven, Belgium
Jan Van Eijgen	Clinical advisor	Research Group Ophthalmology, KU Leuven, Belgium

Table 1.1: Research team involved in this project.

1.5 Scientific dissemination

The scientific contributions resulting from this project are:

- R. Marques, D. A. Jesus, J. B. Breda, J. Eijgen, I. Stalmans, T. Walsum, S. Klein, P. Vaz, L. S. Brea. (2021) *"Automatic Segmentation of the Optic Nerve Head Region in Optical Coherence Tomography: A Methodological Review"*. Submitted for publication. [11]

- R. Marques, L. S. Brea, T. Walsum, S. Klein, J. Cardoso, M. Dias, J. B. Breda, J. Eijgen, I. Stalmans, D. A. Jesus, P. G. Vaz. *"Automated Lamina Cribrosa Segmentation in Optical Coherence Tomography Scans of Healthy and Glaucomatous Eyes "* [Conference session]. 8th Dutch Bio-Medical Engineering Conference. The Netherlands, January 28-29, 2021.

Chapter 2

Theoretical background

The theoretical background presented in this work comprises two sections. The first section begins with the basics of the eye anatomy and physiology, particularly the ONH and the LC, and how they can be affected by glaucoma. The OCT imaging procedure and technique evolution is also described.

The second section is dedicated to the state of the art of the automatic segmentation of the ONH in OCT imaging. The existing methods are described, and their best results presented. Finally, the most relevant biomarkers that can be extracted from ONH segmentation in OCT are presented along with their normative values.

2.1 Eye anatomy

The eye is one of the main sensory organs of the human body. The eye can be divided in three layers. The outer layer is made by the cornea, that refracts and transmits the light to the retina and protects against infection and structural damage, and the sclera, a connective tissue that helps maintain the shape of the eye. The middle layer is made by the iris, that controls the amount of light reaching the retina, the ciliary body, that controls the shape of the lens, and the choroid, a vascular layer that provides nutrients to the outer layer [12, 13]. Finally, the inner layer is the retina, a layered structure of neurons that capture and process light. These three layers surround the aqueous, the vitreous and the lens as schematized in Figure 2.1 (a). The lens also refract the light and allow the eye to focus on objects at varying distances.

The neurons in the retina are divided into six major classes: photoreceptors, bipolar cells, horizontal cells, amacrine cells, ganglion cells, and the Müllerian glia [12]. These cells are arranged in the retina in several parallel layers as shown in Figure 2.1 (b). The human eye has also two types of photoreceptors, responsible for the conversion of light into an electrical signal: rods and cones.

Therefore, the light pathway goes as follows: the light enters through the clear cornea and then successively crosses the aqueous-filled anterior chamber, the papillary aperture, the lens and finally crosses the clear vitreous gel and reaches the retina [13].

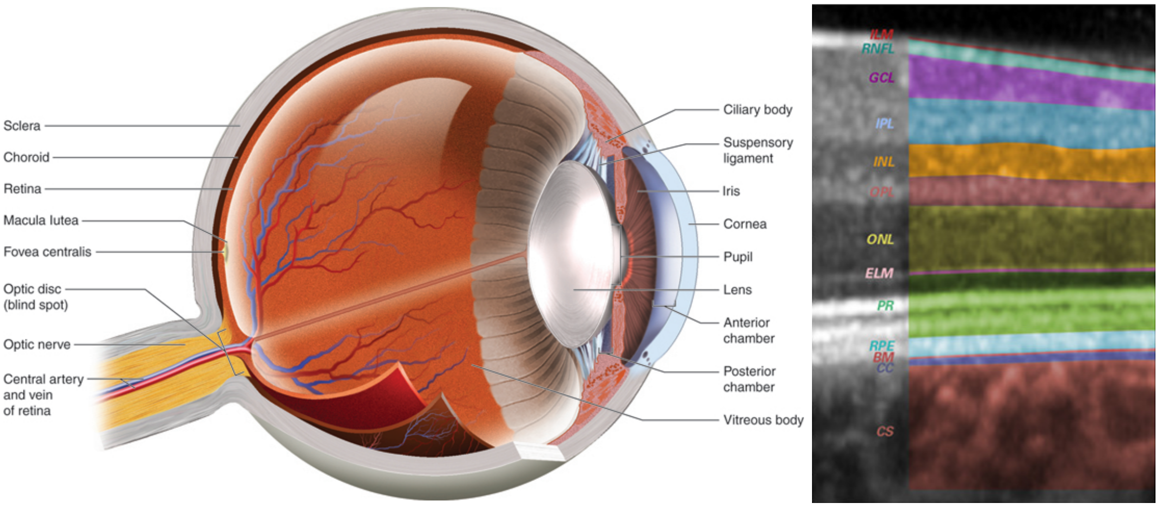
The optic nerve is the largest sensory nerve of the eye and it carries visual signals from the retina to the brain exiting the eyeball on the ONH through ganglion cell axons. These axons pass through an active mesh like structure before converging in the optic nerve. This structure is the LC. The LC is localized in the posterior scleral canal and provides structural and functional support to the ganglion cell axons. Moreover, the LC also accommodates vessels that nourish the retina and stabilizes the IOP by forming a barrier between the intraocular and extraocular space [14].

The eye can be seen as an extension of the brain, displaying similarities with the central nervous system (CNS) in terms of tissue structure and interactions with the immune system. Therefore, it is common for major brain diseases to manifest within the eye and for some ocular diseases to display characteristics of neurodegenerative disorders [15, 16]. Since the eye is easier to access, knowledge acquired from its studying has been hypothesized as a window to the brain functionalities and to the cardiovascular system, thus assisting diagnosis of CNS and cardiovascular pathologies.

2.2 Optical coherence tomography

There are different imaging technologies that show the eye with high resolution and have diagnostic capabilities for glaucoma. However, OCT is the most used due to its high fidelity and ability to observe inner regions of the retina in 3D non-invasively.

OCT uses low coherence interferometry and is a non-invasive optical technique that allows for high resolution in-vivo cross-sectional images of the retina [18]. OCT is a technology that was presented more than 20 years ago, but keeps evolving/improving due to hardware and software advances. These improvements lead to a better view/better fidelity in the retinal structures, especially in the deeper layers and the choroid. Current technologies include spectral-domain optical



(a)

(b)

Figure 2.1: (a) Eye anatomy. (b) Retinal layers: ILM (Internal Limiting Membrane); RNFL (Retinal Nerve Fiber Layer); GCL (Ganglion Cell Layer); IPL (Inner Plexiform Layer); INL (Inner Nuclear Layer); OPL (Outer Nuclear Layer); ELM (External Limiting Membrane); PR (Photoreceptor Layers); RPE (Retinal Pigment Epithelium); BM (Bruch's Membrane); CC (Choriocapillaris) and CS (Choroidal Stroma). Addapted from [17].

coherence tomography (SD-OCT) and SS-OCT (time-domain OCT was the first to appear but is now obsolete).

In OCT, many one-dimensional scans (A-scans) are performed at several depths to create a two-dimensional image (B-scan). If acquired closely and rapidly, these B-scans can be translated into a volume [19].

SS-OCT was introduced as a third generation OCT modality. Both SD-OCT and SS-OCT are categorized as Fourier-domain optical coherence tomography (FD-OCT). FD-OCT, unlike time-domain OCT, does not employ a moving reference arm and uses the Fourier transformation to convert measurements of interfered light into physical delays or distances [20].

SD-OCT uses a super luminescent diode, a continuous light source that emits a broad range of wavelengths, while in SS-OCT, the light source is a tunable laser, which emits a single wavelength at any instant in time and sweeps across a broad range of wavelengths as a function of time [20]. In SD-OCT, diffraction divides the broad wavelength light in a spectrum that is projected into a spectroscope to achieve light interference. SS-OCT, on the other hand, does not require a spectroscope to achieve light interference since the tunable laser already divides the light in a spectrum.

By simplifying the process, SS-OCT enables a faster acquisition, twice as fast as SD-OCT, resulting in less imaging artefacts due to eye movements [19, 21].

Two diagrams showing a typical SD-OCT and SS-OCT system can be found in Figure 2.2 (a) and (b), respectively.

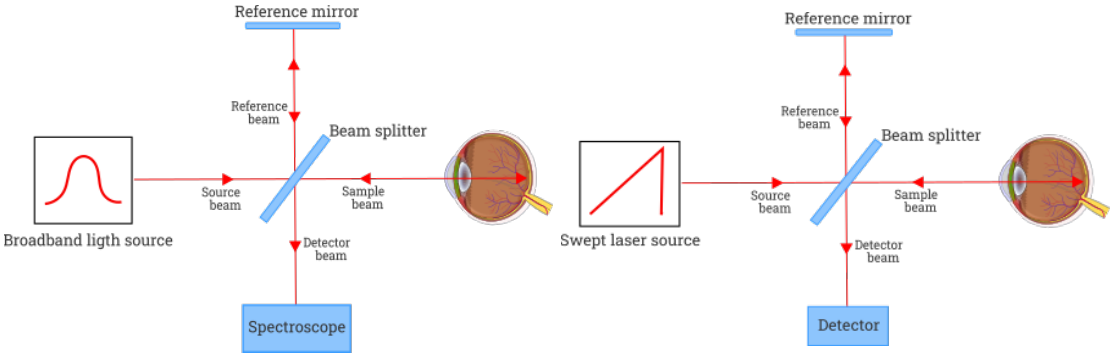


Figure 2.2: Diagram of a (a) SD-OCT and a (b) SS-OCT.

Moreover, the SS-OCT uses a longer wavelength than the earlier OCT systems, which allows for deeper tissue penetration and, consequently, visualization of deeper structures without requiring averaging of multiple B-scans, partly reducing the attenuation and scattering faced by SD-OCT [5, 20]. However, since axial resolution in OCT imaging is directly proportional to the bandwidth of the light source used and longer wavelengths are more readily absorbed by water, SS-OCT can have a lower image resolution compared to SD-OCT when water is present (such as in the vitreous body of the eye) [20].

Despite all advancements, SD-OCT and SS-OCT alone still present limitations. The OCT signal is highly attenuated when reaching deeper structures, and the shadow of the blood vessels, which merge at the ONH, can limit the correct identification of the LC and other ONH structures [22].

Considering this, some improvements in OCT hardware, such as EDI, and light attenuation correction software, such as adaptive compensation, were able to improve the visibility of such structures as the LC without compromising acquisition time. EDI OCT was originally developed in order to improve the visualization of the choroid, although it has also been adopted to improve cross-sectional images of the LC. Adaptive compensation is a post-processing technique developed to remove blood vessel shadows, enhance tissue contrast, and reduce noise over-amplification [5].

Figure 2.3 presents an OCT B-scan where ONH structures mentioned in Section 2.1 can be identified.

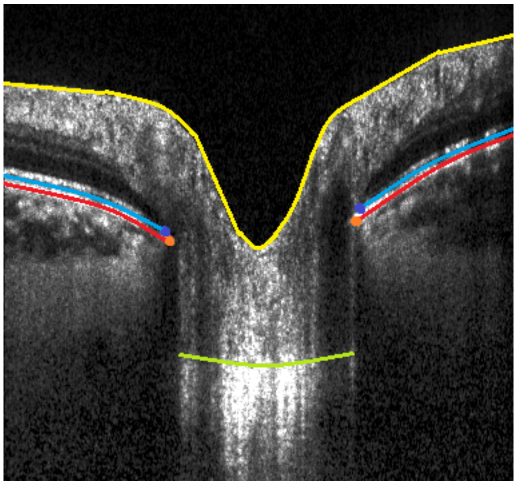


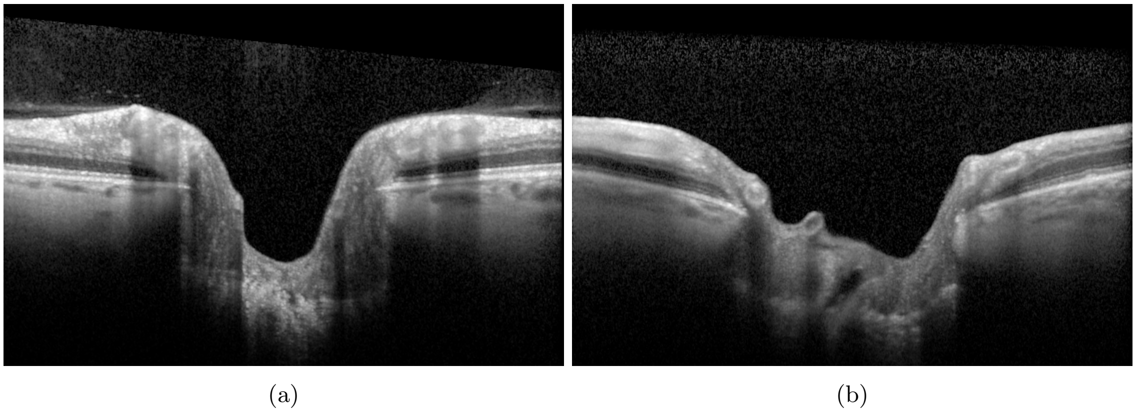
Figure 2.3: OCT B-scan from Spectralis OCT2 (Heidelberg Engineering, Germany) showing a manual boundary-based segmentation of the ONH. In yellow, the ILM anterior surface. In light blue, the RPE layer. In dark blue, the RPE endpoints. In red, the BM. In orange, the BM opening points. In green, the LC anterior surface.

2.3 Glaucoma disease

Glaucoma is one of the main causes of irreversible blindness worldwide [23]. It is a chronic optic neuropathy characterized by progressive degeneration of retinal ganglion cells, resulting in a cupping of the ONH and visual loss [14,24]. High IOP is the main risk factor of the glaucoma disease. Normal IOP values range between 11 and 21 mmHg [25]. Variance of the IOP is expected throughout the day but for values higher than 21 mmHg there is an increased incidence of development of glaucoma. Current management of glaucoma also include an assessment of the visual field through the visual field mean deviation (VF MD) test. The lower the VF MD, the more damaged the visual field is and the more likely it is for glaucoma to develop.

The ONH cupping is associated with complex 3D structural modifications, such as thinning of the RNFL, changes in the Bruch's membrane minimum-rim-width (BMO-MRW) and LC depth, thickness and curvature [5, 26–29]. Since glaucoma is an irreversible but preventable disease, its early detection is crucial, making the understanding of ONH biomechanics increasingly important.

In Figure 2.4 is possible to see the differences between the ONH of healthy and glaucomatous eyes. Corroborating existing studies [30], the prelaminar tissue, the accumulated RNFL tissue on the ONH right before the LC [30], is thicker in glaucoma patients, and the RNFL thinner, resulting from a loss of optic nerve fibers. Moreover, the visible increase of the LC depth, and its backward bowing and thinning, are also associated with axonal death, which leads to visual field loss with the progression of the disease.


(a)

(b)

Figure 2.4: (a) Radial OCT scan from a healthy subject. (b) Radial OCT scan from a glaucoma patient. Both images have been acquired using an Spectralis device, as part of the Leuven Eye Study [31].

Evidence suggests that the ONH surface depression occurs before the RNFL thinning [6], making it more relevant for early diagnosis. The load bearing tissues of the ONH are the peripapillary sclera, the LC and the scleral canal [32]. These connective tissues bear the forces generated by the IOP. However, when these tissues fail under high IOP, it can affect the blood supply to the laminar segments of the axons, limit the diffusion of nutrients, and weaken the LC, making it more prone to damage. [12]

The LC works as a barrier between two differentially pressurized compartments: the intraocular space and the retrobulbar space. The pressure on the LC tissue is reduced from a higher pressure level, the IOP, to a lower pressure level, the cerebrospinal fluid pressure (CSFP), creating a pressure gradient that leads to a backward bowing of the LC. The material properties and geometry of the LC, and remaining connective tissues, determine the amount of pressure it can withstand without being deformed [14, 33]. Glaucoma is characterized by an enlargement of the ONH and alterations to the LC thickness and depth [14, 34]. Studies have shown that LC deformations occur in a very early stage of glaucoma [35] and Paulo *et al.* has found lamina cribrosa thickness (LCT) and lamina cribrosa depth (LCD) as being the most used and discriminative features. Moreover, LC features are seen as promising for patient follow-up as well, since they are significantly different between healthy patients and ocular and systemic pathologies while being patient-specific [34]. An illustration of the measurements differences between healthy subjects and different ophthalmic diseases for LCT and LCD can be found in Figures 2.5 and 2.6, respectively.

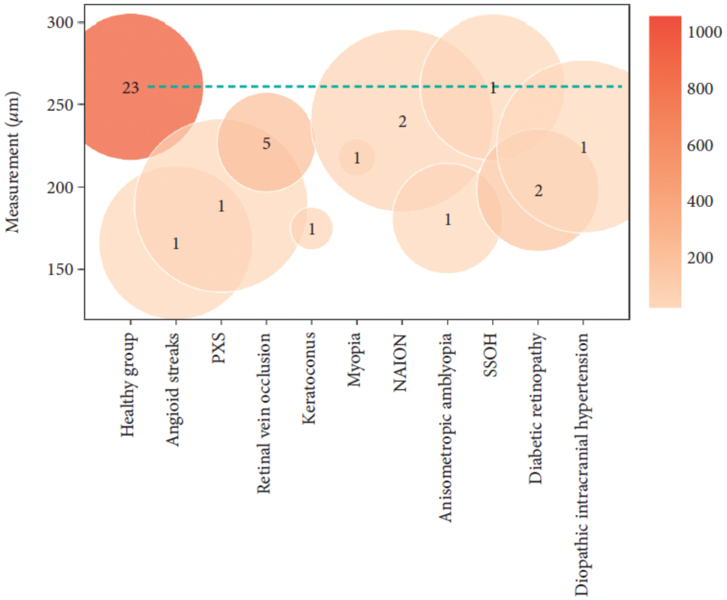


Figure 2.5: Comparison of (a) LCT measurements in healthy subjects and ophthalmic diseases patients. The dashed green line represents the mean for the healthy population. The number in the circle represents the amount of studies used for calculating the averaged measurements for each disease. The color scale shows the number of eyes comprised in the studied groups, and the radius of each circle denotes the standard deviation of the averaged values. Reproduced from [34].

Although these ONH structural changes have been mostly studied in a context of glaucoma diagnosis, they are also widely representative of non ophthalmic diseases such as idiopathic intracranial hypertension (IIH), optic neuritis (ON), multiple sclerosis (MS) or neuromyelitis optica spectrum disorders (NMOSD) [22], Alzheimer [16, 36], and Parkinson's disease [37]. An illustration of these differences for LCT measurements can be found in Figure 2.7.

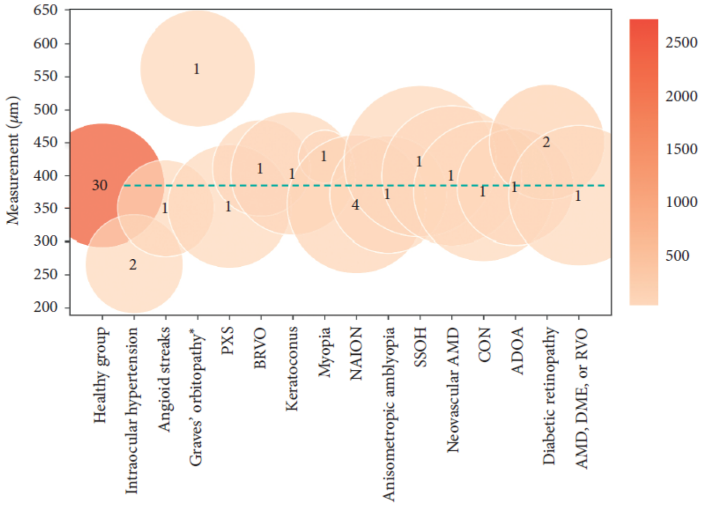


Figure 2.6: Comparison of (a) LCD measurements in healthy subjects and ophthalmic diseases patients. The dashed green line represents the mean for the healthy population. The number in the circle represents the amount of studies used for calculating the averaged measurements for each disease. The color scale shows the number of eyes comprised in the studied groups, and the radius of each circle denotes the standard deviation of the averaged values. Reproduced from [34].

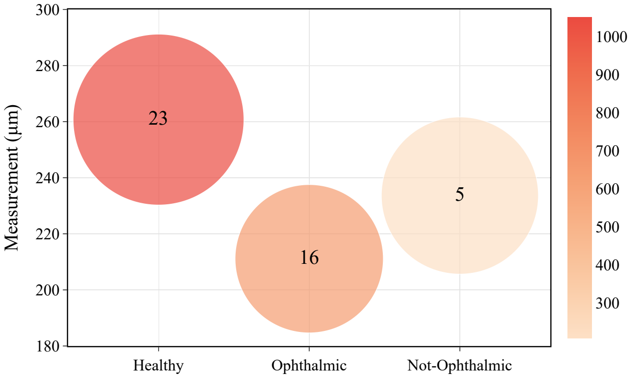


Figure 2.7: Comparison of LCT measurements between healthy subjects and ophthalmic and non-ophthalmic patients. The number in the circle represents the amount of studies used for calculating the averaged measurements for each group. The color scale shows the number of eyes comprised in the studied groups, and the radius of each circle denotes the standard deviation of the averaged values. Reproduced from [34].

2.4 State of the art

2.4.1 Automatic segmentation of the optic nerve head in optical coherence tomography

In this section, a description of existing algorithms to segment the ONH in OCT and their main results are provided. Special attention is paid to the methods aiming to include the LC in their segmentation.

The studies were separated in three categories based on their learning strategy and complexity: conventional methods, that use non-learning based image processing techniques only, machine-learning methods (alone or as a refinement/post-process step after conventional methods), and deep learning methods.

Conventional methods

Conventional methods are unsupervised segmentation techniques that rely on image processing methods such as thresholding, edge detection and morphological operations.

Belghith *et al.* [38] proposed a novel shape constrained surface evolution method to segment the anterior surface of the LC. For this purpose the group used the Markov random field (MRF) class of Bayesian methods. The MRF is a statistical image model that introduces a boundary cost to the segmentation by including a constraint that neighboring pixels are likely to share the same classification. With Bayesian models, prior knowledge about the shape, position and distribution of the LC voxels can be incorporated in the model so that interface creation is penalized with a model of the intensity of the region for an improved pixel classification [39]. To overcome vascular and other reflective structures artifacts, they considered a non-local framework for the MRF energy function [40]. The non-local approach exploits repetitive structures in the image to create a multi model from a single observation. The LC surface is iteratively refined following a perturbation-based approach inspired by the biased and filtered point sampling [41] according to the non-local MRF energy function. By applying this method in a longitudinal study using 21 healthy and 21 glaucomatous eyes, they were also able to show a correlation ($R^2 = 0.68$) between the IOP variation and the anterior LC surface depth variation over time.

Mao *et al.* [42] started by addressing the common appearance of noise and artifacts in the LC region by training a deep learning model to denoise the B-scans. Since the group intended to explore a 3D segmentation, they felt the need to remove or compensate the shadow artifacts from all directions of the volume to recover the lost information in the shadowed areas. Moreover, since the existing adaptive contrast enhancement methods [9,43] to improve the visibility of deeper structures such as the LC often yield a bright band bellow the compensation depth limit, Mao *et al.* set a linear contrast adjustment step after the denoising and before applying adaptive contrast enhancement. This way, the noise floor at the deep layers is minimized. Finally, a two-round segmentation method to find the 3D anterior surface of the LC was proposed. They apply 2D and 3D canny edge detectors to the interpolated 3D volume (selected from an *en-face* image to include the ONH region only). Different weights, based on prior knowledge of the LC anatomy, are applied to the vertical and horizontal gradients of the canny edge detectors. In the first round the confidence level of the LC border is calculated with the accumulated cost map along the shortest path to each location. The second round removes candidate points based on their distance to the neighboring LC border, updating it in the end. A total of 180 individual B-scans from 36 subjects were reviewed, achieving a segmentation accuracy of 91%.

Machine-learning methods

Machine-learning methods find patterns and features in large amounts of data. They are usually applied after, or in combination with, conventional methods, in order to improve performance.

The hole structure of the ONH can be challenging to segment since it is embedded with multiple surrounding surfaces that cease to exits on the location where the optic nerve exits the eye. Therefore, existing methods to segment the retinal layers without considering the ONH are prone to failure [44,45].

Addressing this, Antony *et al.* [46] adapted an existing graph theoretic approach [47] proposed for simultaneous segmentation of multiple continuous surfaces in order to make it able to identify the ONH boundary in 3D. They proposed an iterative approach to correct the z-axis values until the convergence of the segmented ONH boundary column. As part of the iterative process, a random forest classifier was trained to find the boundary of the optic disc in volumetric OCT data, based on the previously learned textural features.

However, the presence of externally oblique border tissue, which attaches to the end of the Bruch's membrane (BM) surface and looks very similar to the ending point of the BM surface, can sometimes mislead iterative processes. This may lead to a wrong placement of the Bruch's membrane opening (BMO) and, consequently, of the borders of the ONH.

Therefore, Miri *et al.* [48] eliminated the iteration phase, and presented an automatic machine-learning graph-theoretic approach to segment the BMO points. The BMO points were first identified in 2D with a graph theoretic approach. Then, they used a random forest classifier to compute a cost function that was able to identify the BMO points in 3D, based on a principal component analysis (PCA) intensity model of the BMO points. Finally the BMO points were segmented in 3D using a shortest path method while refining the path in the z-direction. Twenty B-scans were used per subject, from a dataset that included suspects and confirmed cases of glaucoma. The unsigned mean error measurements of the BMO-MRW were 26.65 ± 13.27 μm and 22.22 ± 5.99 μm for Antony *et al.* and Miri *et al.*, respectively. The signed mean error measurements of the BMO-MRW were 6.61 ± 18.59 μm and $-$ 0.30 ± 12.44 μm for Antony *et al.* and Miri *et al.*, respectively. And the root mean square error (RMSE) values were 17.99 ± 8.15 μm and 11.62 ± 4.63 μm for the previous methods, respectively. These results corroborate the superiority of the second method in segmenting the BMO points.

Yu *et al.* [49] also modified a graph search algorithm with a random forest classifier. They used locally adaptive constraints obtained from previously detected surfaces as constraints for the subsequent layers smoothness. Yu *et al.* achieved a dice similarity coefficient of 0.925 ± 0.030 for the ONH boundary detection and an overall mean unsigned border position error of 7.3 ± 5.4 μm.

Wu *et al.* [50] combined a graph search approach with with a patch search using a support vector machine (SVM) to segment the ONH boundaries and calculate the cup-to-disc ratio (CDR). The features extracted for patch description were the local binary pattern and histogram of gradient. The unsigned border error for ONH segmentation, and the evaluation error for CDR comparing with manual segmentation, were 67 ± 42 μm and 0.045 ± 0.033, respectively.

Deep learning methods

Deep learning methods are an advanced type of machine-learning algorithms that have been gaining visibility in the last decade. They are able to extract and classify features automatically when a large amount of training data is given [51].

Chen *et al.* [52] outperformed Wu *et al.* [50] (67.00 μm) and Miri *et al.* [48] (49.28 μm) with a mean error of 42.38 μm for BMO detection.

They proposed a two stage segmentation method. First, in the coarse detection, they find the region of interest, indicative of the BMO location, with a registration between color fundus images and the 2D projection and segment the retinal pigment epithelium (RPE) by training a SVM to construct the energy of a conditional random field. A fixed detection based on a U-Net [53] convolutional neural network (CNN) was used to improve the accuracy of the coarse detection.

Since the dataset had an imbalanced class problem, the loss function of the network is based on the Dice loss. However, since the area of the RPE is too small compared with the entire image to give an high Dice coefficient, they also added area bias and mean square error (MSE) to the Dice loss. Their loss function is given by:

$$Loss = Dice\ Loss + Area\ Bias + MSE =$$
$$2 - \frac{2\sum_x^{\Omega} p_x g_x}{\sum_x^{\Omega} p_x^2 \sum_x^{\Omega} g_x^2} - \frac{min(\sum_x^{\Omega} p_x \sum_x^{\Omega} g_x)}{max(\sum_x^{\Omega} p_x \sum_x^{\Omega} g_x)} + \frac{1}{n}\sum_x^{\Omega}(p_x - g_x)^2$$

where p_x and g_x are the prediction and the ground truth with the pixel x in the patch Ω and n is the number of total pixels in region Ω. They used the Stochastic Gradient Descent optimizer with a 0.9 momentum, a learning rate that started at 0.001 and is gradually decreased and a dropout rate of 0.5.

Heisler *et al.* [54] verified that for the segmentation of the inner limiting membrane (ILM), posterior surface of the RNFL, BM and choroid–sclera boundary a Pix2Pix generative adversarial network (GAN) [55] performed better than a U-Net [53]. The network was trained with an Adam optimizer, learning rate of 0.0001 and cross entropy loss function. Even though the U-Net had a good performance, better results were achieved with a semi supervised GAN and fine-tuning using pseudo-labels when using a small dataset. Moreover, when using the semi-supervised GAN with 10x more data, the performance improved only in 1-2%. For BMO segmentation, the same group used a Faster Region-CNN [56]. They used Adam optimizer with a 0.00001 learning rate and the loss function was the sum of the classification (log loss) and bounding box regression (smooth loss). Most of the incorrect BMO segmentations were successfully eliminated during post-processing, leading to a strong correlation for both glaucoma ($R^2 = 0.93$) and control eyes ($R^2 = 0.99$) when comparing the BMO area parameter in manual and automatic methods.

Devalla *et al.* [57], on the other hand, aimed to segment six neural and connective tissue structures in OCT images of the ONH: the RNFL and the prelamina, the RPE, the remaining retinal layers (as a whole), the choroid, the sclera, and the LC (as shown in figure 2.8).

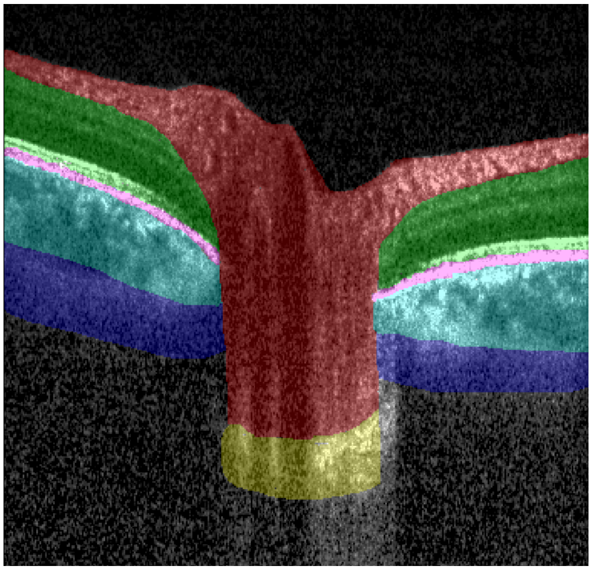


Figure 2.8: OCT B-scan from Spectralis OCT2 showing a manual segmentation of the ONH tissues. In red, the RNFL and prelamina. In green, the other retinal layers. In pink, RPE. In light blue, the choroid. In dark blue, the sclera. In yellow, the LC.

They used a CNN as detailed in Figure 2.9. The output layer had 6 neurons, one for each class of tissues, and a softmax function was applied to obtain the probability of each patch to belong to a certain class. Every patch was labeled with the highest probability class in its center. They used Adam optimizer, a learning rate of 0.001 and a 35% dropout in the last layer before softmax activation.

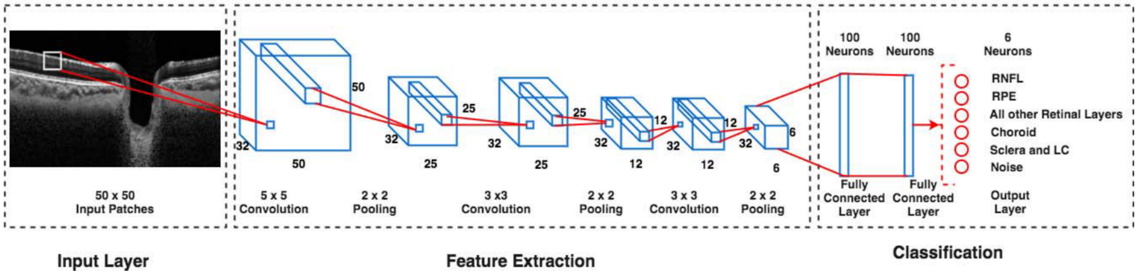


Figure 2.9: Architecture of the network used in Devalla *et al.* [57].

For all the tissues the averaged sensitivities varied between 0.89 and 0.97, the averaged specificities were always higher than 0.98 and no significant differences were observed in the performance metrics between healthy subjects and glaucoma patients. However, these metrics were not used to the peripapillary sclera and LC, since their visibility varied considerably across images. Moreover, they found the algorithm performed better when trained with adaptive compensation pre-processing of the images [9]. Visually, it was possible to verify that this methods still fails in separating the LC from the sclera, has some artificial LC-scleral insertions and does not offer reliable tissue boundaries.

Therefore, Devalla *et al.* [58] later introduces the dilated-residual U-Net (DRUNET). The DRUNET is based on the combination of a U-net and residual blocks. To improve the accuracy of the simultaneous segmentation of the 6 tissues, they used a CNN that exploits the advantages of U-Net skip connections, residual learning and dilated convolutions in order to add contextual information to the local information. The architecture of the network can be found in Figure 2.10.

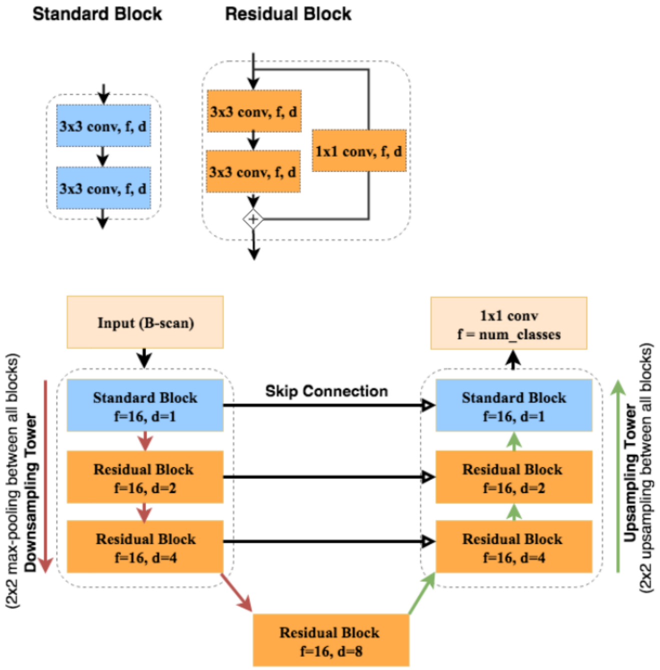


Figure 2.10: Architecture of the DRUNET used in Devalla *et al.* [58] where *f* is the number of feature maps, *d* is the dilation rate and *conv* is a convolution layer.

All layers, except the output, were batch normalized and activated by an exponential linear unit function. They used Stochastic Gradient Descent with a Nesterov momentum of 0.9, an initial learning rate of 0.1 that is gradually decreased and a loss function based on the mean Jaccard

Index, calculated for each tissue:

$$Jaccard\ Index_i = \sum_{i=1}^{N} \frac{|P_i \cap T_i|}{|P_i \cup T_i|}$$
$$Loss = 1 - \frac{Jaccard\ Index}{N}$$

where i is the index of each tissue, N is the number of classes, P_i the pixels predicted as part of class i and T_i the pixels labeled as part of class i in the manual segmentation.

When compared to the previous method, the results showed that the DRUNET performed overall significantly better for all the tissues and it was able to dissociate the LC from the peripapillary sclera providing an advantage against existing techniques. However, the segmentation of the peripapillary sclera and of the LC was only qualitatively assessed. The mean sensitivities for all the tissues for glaucoma and healthy eyes were 0.92 ± 0.04.

Devalla *et al.* [10] also addressed the lack of device-independent algorithms for the automatic segmentation of the ONH in OCT. This is one of the main reasons why these methods have not been adopted in clinical practice yet. There are already some major commercial manufacturers of OCT devices and considering the fast technological developments and increased search in the market, the next-generation of devices will soon start to release.

Given the variability in the images each device provides, due the proprietary processing software, it is becoming increasingly unfeasible to train the algorithms with labeled data from each existing device. Since patients are normally imaged by different OCT devices during their care, the adoption of automatic segmentation to aid clinical diagnosis and follow-up has been limited. Moreover, it also limits validation of the algorithms in research.

To overcome these differences, Devalla *et al.* [10] proposed a deep learning based enhancer to reduce speckle noise, enhance contrast and equalize the histogram of images from three different SD-OCT devices (Spectralis, Cirrus and RTVue) while reducing the differences between their device specific characteristics. For this purpose they used an existing network [59] and added a sigmoid activation function to the output layer.

For ONH segmentation they used 3D CNNs since it can further improve the reliability of the automatic segmentation, by not only harnessing the information from each image, but also effectively combining it with the depth-wise spatial information from adjacent images. Each CNN has four micro U-Nets and a latent space. Each of the three CNNs gives a different but equally plausible segmentation. The feature maps of each segmentation are concatenated and fed to an ensembler. The ensembler has three sets of 3D convolutional layers separated by a dropout layer to fine-tuning the segmentation. The architecture can be found in Figure 2.11.

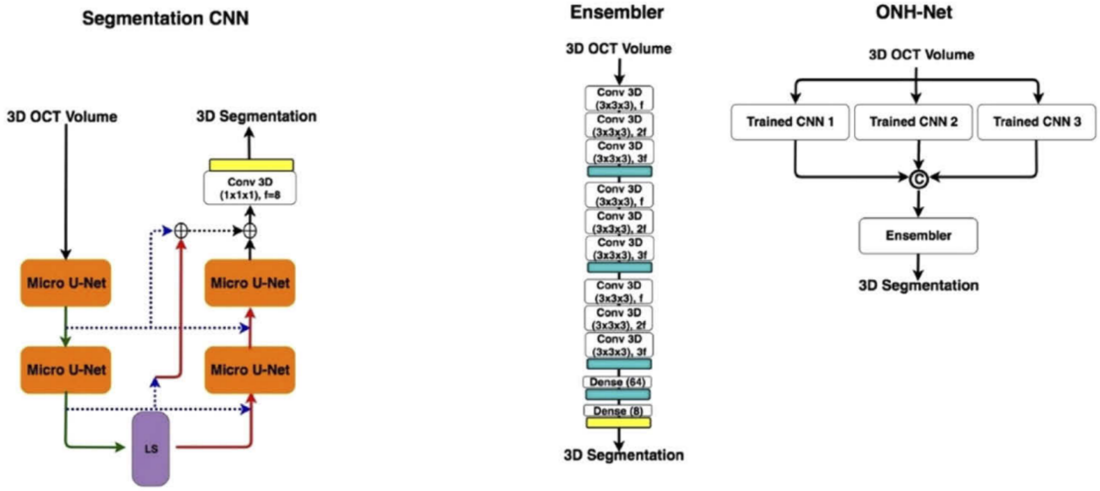


Figure 2.11: Architecture of the network used in Devalla *et al.* [10]. The yellow blocks represent a softmax activation, the blue blocks represent a dropout layer, the red arrows are transpose 3D convolutions (stride=2), the green arrows are 3D max pooling (stride=2), the blue arrows are skip connections and f is the number of feature maps.

The networks were trained with the stochastic gradient descent (Nesterov momentum of 0.05), learning rate of 0.01 and the mean Jaccard index based loss function used with the DRUNET [58].

The method was tested using 20 volumes per device from both healthy and glaucoma patients. In all cases, the mean structural similarity index for the deep learning enhanced B-scans (compared to digitally- enhanced B-scans) were: 0.95 ± 0.02, 0.91 ± 0.02, and 0.93 ± 0.03, for Spectralis, Cirrus, and RTVue, respectively. The mean sensitivities / specificities (mean of all tissues; mean±SD)were: 0.94 ± 0.02 / 0.99 ± 0.00, 0.93 ± 0.02 / 0.99 ± 0.00, and 0.93 ± 0.02 / 0.99 ± 0.00 for Spectralis, Cirrus, and RTVue, respectively. Given the subjectivity in the visibility of the posterior LC boundary, it was excluded from quantitative assessment. Furthermore, the results from deep learning enhanced OCT volumes, regardless of the device used for training, showed no significant differences in the segmentation performance for all tissues, except for the LC.

2.4.2 Optic nerve head derived biomarkers

Biomarker is defined as a characteristic that is objectively measured and evaluated as an indicator of normal biological processes, the presence or progress of a disease, or the effects of treatment [60].

For this project in particular, biomarkers will be referred to as the measurements taken from the structures of the ONH that will quantify its changes. The comparison between biomarkers measurements in different pathological groups allows for more insight into the biological and pathogenic processes that differentiate the groups.

Next, some of the most common biomarkers are described and their reference values presented. Note that reference values are not considered normative, since the quality of the images varies throughout devices and acquisition protocols. Moreover, there is not a standard way to compute most LC biomarkers, as highlighted by Paulo *et al.* [34], which limits the comparison of reference values between studies.

Optic disc diameter

The optic dis diameter assessment is an important part of the ONH examination since its variation is deeply connected with the neuroretinal rim and optic cup size [61]. There is not an agreement on its use for glaucoma diagnosis since the many available measurement techniques provides estimates that are not comparable between studies and there is a large variation within and among populations. Hoffmann *et al.* [61] reported that small discs range between 1.1-1.3mm while bigger discs range between 1.8 and 2.0mm.

Retinal nerve fiber layer thickness

The RNFL is the expansion of the fibers from the optic nerve and is made of unmyelinated ganglion cell axons. The RNFL been identified as one of the primary sites of glaucoma damage, and its thinning can be measured before visual loss starts to occur. There is wide variability for normal RNFL thickness values and the influence of several factors, such as age, sex, and race, has been widely studied. One of these studies [62] showed, for healthy eyes, a mean RNFL thickness of 97.3 ± 9.6 μm. The RNFL thickness is represented in Figure 2.12 with the red arrows. Bowd *et al.* [63] verified that RNFL thickness in normal eyes varied from 40 to 105 μm with a standard deviation of 6.2 μm and in glaucoma eyes it varied from 4 to 85 μm with a standard deviation of 6.4 μm.

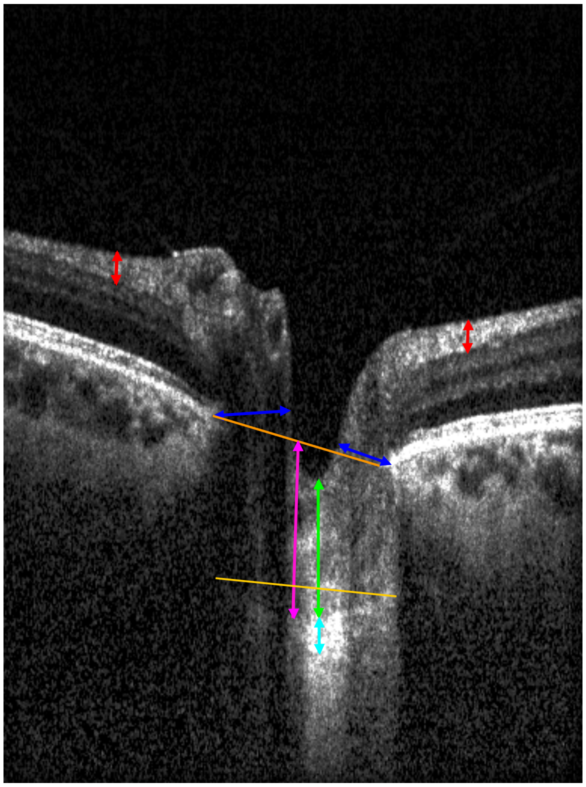


Figure 2.12: OCT B-scan from a healthy subject showing the biomarkers referred in this section. In red, the RNFL thickness. In dark blue, the BMO-MRW. In orange, the BMO reference line. In green, the prelaminar tissue thickness. In yellow, the curvature reference plan. In pink, the LC depth. In light blue, the LC thickness.

Statistical tests proved that the RNFL is significantly thinner in glaucomatous eyes ($p<0.05$).

Prelaminar tissue thickness

The prelaminar tissue thickness is defined as the distance from the optic cup to the anterior surface of the LC [30] (the green arrow in Figure 2.12).

In the study by Wu *et al.* [30], a total of 91 eyes were examined with SS-OCT and divided into IOP≥30 mmHg, 21 mmHg<IOP<30 mmHg, and normal IOP control group. The baseline values of the prelaminar tissue thickness for patients with IOP > 30 mmHg, 21 mmHg < IOP < 30 mmHg and normal were 107.5 ± 42.6 μm, 171.1 ± 131.2 μm and 242.9 ± 142.3 μm, respectively. The group showed that the prelaminar tissue was significantly thinner in glaucomatous groups compared with controls (p = 0.045).

Bruch's membrane opening - minimum rim width

The BMO-MRW is defined as the shortest distance between the BM termination and ILM anterior surface [64] (see dark blue arrows in Figure 2.12).

In Wu *et al.* [30], the baseline values of BMO-MRW for patients with IOP $>$ 30 mmHg, 21 mmHg $<$ IOP $<$ 30 mmHg and normal are 129.6 ± 38.0 μm, 154.6 ± 67.3 μm and 195.1 ± 99.6 μm, respectively. They showed that the BMO-MRW decreased with increasing IOP among the three groups and Gmeiner *et al.* [64] verified that BMO-MRW and RNFL thickness have similar glaucoma diagnostic potential.

Lamina cribrosa depth

The LCD (pink arrow in Figure 2.12) is defined as the perpendicular distance from the BMO plane (orange line in Figure 2.12) to the maximum depth point of the anterior surface of the LC [34].

In Wu *et al.* [30] the baseline values of the LCD for patients with IOP $>$ 30 mmHg, 21 mmHg $<$ IOP $<$ 30 mmHg and normal are 655.3 ± 173.9 μm, 513.9 ± 154.8 μm and 404.5 ± 102.9 μm, respectively.

Lamina cribrosa curvature index

Posterior bowing of the LC may be related to mechanical or vascular damage to the ONH [65]. The lamina cribrosa curvature index (LCCI) is used to quantify the curvature of this bowing. A reference line of the curvature (see yellow line in Figure 2.12) is set between the two extreme points of the anterior surface of the LC. The LCCI is therefore calculated as:

$$LCCI = \frac{LCCD}{W} \times 100$$

where lamina cribrosa curvature depth (LCCD) is the maximum depth between the reference line and the anterior surface of the LC and W is the length of the curvature reference line [27].

In Wu *et al.* [30] the baseline values of LCCI for patients with IOP $>$ 30 mmHg, 21 mmHg $<$ IOP $<$ 30 mmHg and normal are 76.2 ± 28.0 μm, 58.0 ± 33.8 μm and 35.8 ± 29.2 μm, respectively.

Lamina cribrosa thickness

The LCT is defined as the distance between the anterior and the posterior surface of the LC (see light blue arrow in Figure 2.12), that in OCT images correspond to the borders of the highly reflective region bellow the cup of the ONH [34].

In Wu *et al.* [30] the baseline values of LCT for patients with IOP > 30 mmHg, 21 mmHg $<$ IOP < 30 mmHg and normal are 221.4 ± 69.8 μm, 225.8 ± 46.6 μm and 243.7 ± 47.3 μm, respectively.

Chapter 3

Methods

The following chapter describes the methods employed during this project. First, the available dataset is described. Second, the process to obtain the segmentation ground truth is explained. Third, the neural network used for the automatic segmentation is presented and explained. Lastly, the processes for the automatic calculation of the biomarkers and the model's validation are described.

A flowchart with the summarized tasks comprised in this project can be found in Figure 3.1.

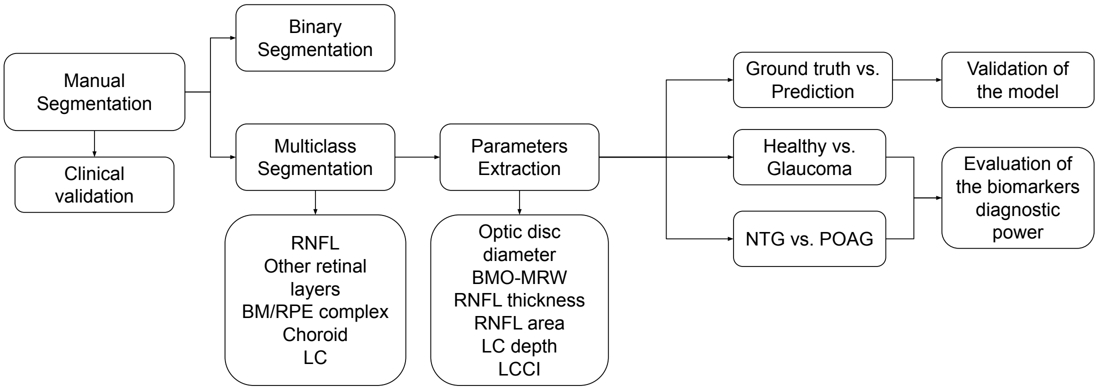


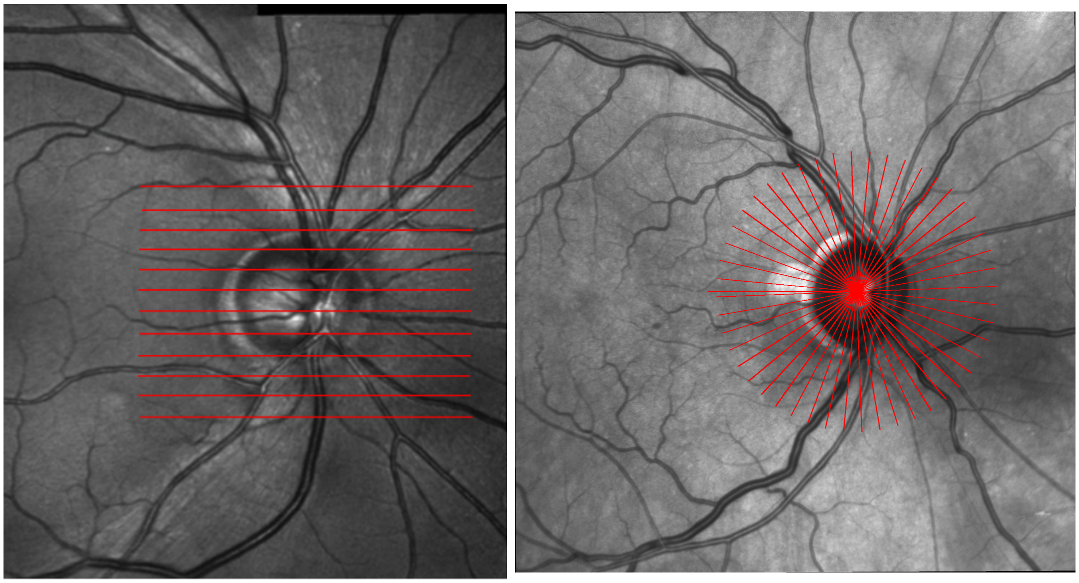
Figure 3.1: Flowchart with the project tasks.

3.1 Datasets

Two different datasets of ONH centered OCT volumes were available for this project:

- **Dataset A:** 23 volumes from 13 healthy subjects were provided by the Cardiovascular R&D Center in Hospital S. João, Porto, Portugal. These volumes were acquired with Spectralis OCT 2 (Heidelberg Engineering, Germany) with a raster scanning protocol (Figure 3.2 (a)). Figure 3.3 (a) shows an example of a B-scan from dataset A.

- **Dataset B:** 20 volumes from 20 healthy subjects and 93 volumes from 46 glaucoma patients acquired with a Spectralis OCT 2 (Heidelberg Engineering, Germany) by the Research Ophthalmology Group in Hospital UZ Leuven, Leuven, Belgium, using a radial scanning protocol (Figure 3.2 (b)). A B-scan from one of these volumes can be seen in Figure 3.3 (b).



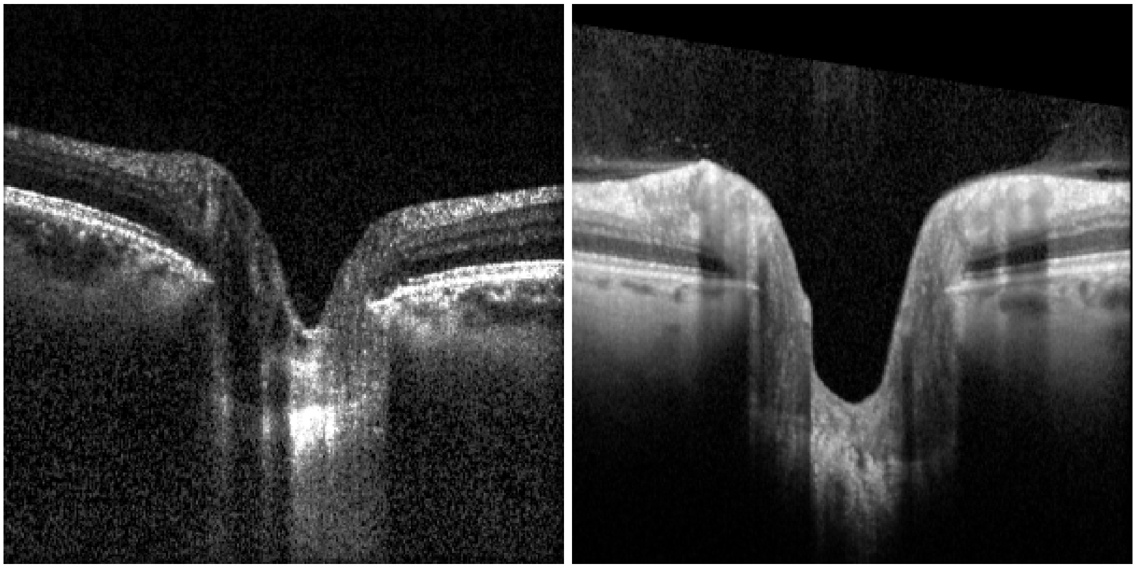
(a) (b)

Figure 3.2: Illustration of a (a) raster and (b) radial acquisition. The red lines represent the B-scans.

The notorious differences observed when comparing the two B-scans on Figure 3.3, from two different protocols, highlight one of the main issues with OCT - not even when using the same device is possible to avoid having vastly different images. These differences between the images of the two different datasets can be explained by the acquisition techniques used. Acquisition protocols differ mostly on the number of scans used for each cross-section, and in the distance between cross-sections. The larger the number of B-scans used in the averaging of a cross-sectional image, the better the signal-to-noise ratio. This is the case for the radial scans of Dataset B.

Dataset A was acquired in the high speed mode. The spacing was 11μm, 3.9μm, 30μm in the lateral, axial and transversal directions respectively. Each volume had 144 scans with a 496 × 384 pixels size.

Dataset B was acquired in the OCT high resolution mode. The spacing was 5.7μm, 3.9μm in the lateral and axial respectively. Each volume had 24 scans with a 496 × 768 pixels size. In both cases, the scan depth into the tissue was 1.9mm.



(a) (b)

Figure 3.3: Example of a B-scan from (a) a healthy subject from dataset A and (b) a healthy subject from dataset B.

3.2 Manual segmentation

Given the lack of publicly available datasets with manual segmentations, 300 B-scans were manually segmented from the described datasets. It was not possible to segmente all available B-scans due to time constraints. The number of segmented images from each dataset are depicted in Figure 3.4.

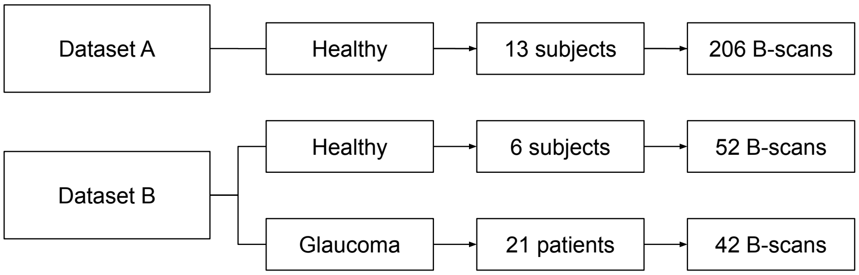


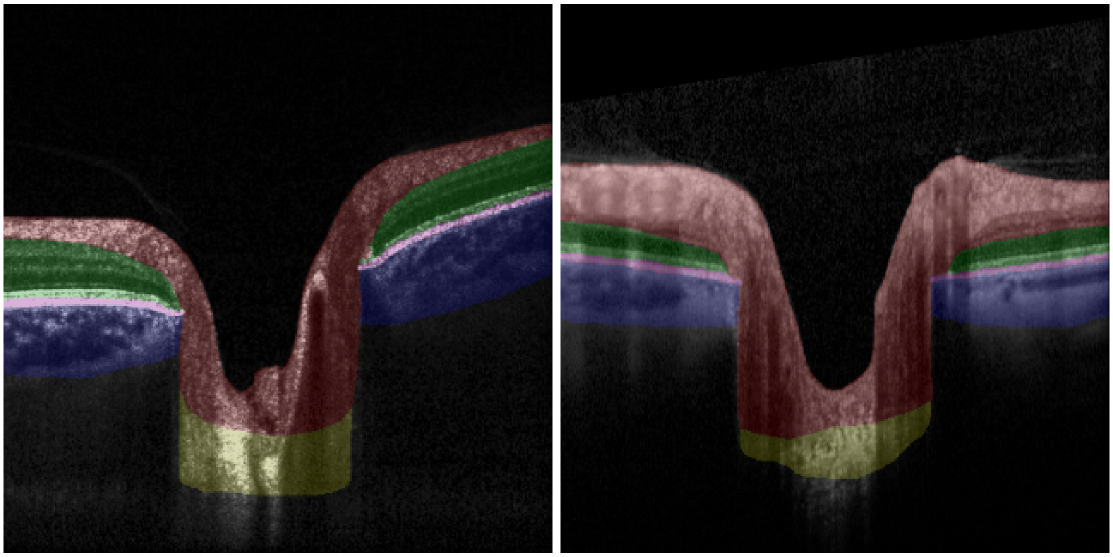
Figure 3.4: Diagram with the number of B-scans segmented per dataset including healthy and glaucoma subjects.

The following layers of tissues were segmented: (1) RNFL; (2) BM/RPE complex; (3) other retinal layers; (4) choroid and (5) LC , as seen in Figure 3.5. The other retinal layers include the ganglion cell layer, the inner and outer plexiform layers, the inner and outer nuclear layers and the external limiting membrane. All B-scans were segmented using ITK-SNAP (version 3.8.0).

In dataset A, 10 B-scans were segmented in each volume. The 10 B-scans belonged to the optic nerve canal and were separated by 4 B-scans each.

In dataset B, 3 random B-scans were segmented in each of the selected volumes. While all the B-scans from dataset B, being radial scans, cross the center of the ONH, in dataset A, with all B-scans being line-rasters, only one in each OCT volume crosses the center of the ONH and the others are positioned in a parallel to each side of the center scan. Therefore, the choice of a lower number of B-scans per volume than in dataset A is due to the lower variability between images of the same volume.

All manual segmentations were subsequently corrected by a retinal specialist from Hospital S. João, Porto, using the same software, ITK-SNAP. The inter-grader variability was assessed by computing the Dice coefficient, sensitivity, specificity and overall accuracy excluding the background between the preliminary and corrected segmentations. However, segmentations of Dataset B were not clinically validated in time to be used to train and test the segmentation model. So, in summary, the trained models used corrected images from Dataset A and preliminary segmentations from Dataset B.



(a) (b)

Figure 3.5: Example of a manual segmentation from (a) dataset A and (b) dataset B. In red, the RNFL. In pink, the BM/RPE complex. In green, the other retinal layers. In blue, the choroid. In yellow, the LC.

The BM and RPE were segmented together, given the difficulty in separating these two hyper reflective layers in OCT images. However, the RPE, as one of the retinal layers, is independent from the BM. For this reason, the termination points of the BM and the RPE not always coincide (peripapillary atrophy [66]). In these cases, the complete segmentation of the BM, even when no longer attached to the RPE, prevailed since an accurate segmentation of its termination is crucial for the correct computation of several biomarkers in the ONH.

Finally, the inter-grader variability effect on the biomarkers computation is also addressed. For this purpose, the biomarkers are computed on all manual segmentations from both graders, and compared between them in the same image.

3.3 Automatic segmentation

As seen in the Section 2.4 (State of the art), deep learning methods achieve accuracies over 90% in the segmentation of both neural and connective tissue of the ONH in OCT. The classification problem at cause is one of semantic segmentation. Semantic segmentation is a form of pixel segmentation, in which every pixel of the image is assigned a label. In this case, we intend to automatically label each pixel of a given ONH centered OCT B-scan in either background or one of the tissues manually identified as the ground truth. The segmentation problem can be tackled in two manners: either with separated binary models for each tissue, or by training a multiclass network

The U-net architecture [53] was chosen for this application due to its efficiency in the domain of the medical imaging segmentation. In the binary segmentation, one U-net was trained per tissue, while for the multiclass segmentation, only a semantic segmentation/multiclass model was trained.

In order to avoid bias in the model evaluation, the data in the train, validation and test sets were separated at patient-level instead of B-scan level. This way, it was avoided the testing on data very similar to the one used for training.

This section starts with a brief introduction to CNNs and some of its basic concepts and later uses these concepts to display and explain the architecture used in this project.

3.3.1 Convolutional neural networks

A CNN is a type of artificial neural network often used in image processing [67]. Since the chosen architecture is a type of CNN, their basic structures and principles are going to be briefly presented and explained.

Convolutional layer

The CNNs use convolutional layers for at least one of its layers. The first convolutional layers are responsible for capturing low-level features, such as edges and color. As more layers are added, the network becomes able to capture high-level features as well.

A convolutional layer uses a three dimensional kernel with predefined dimensions $N \times M \times C$ where N and M correspond to the kernel's height and width, respectively, and C is the number of channels of the kernel, which should be equal to the number of channels of the input volume of the layer. Another important parameter when defining a convolution is the stride. The stride is the length of each step taken by the kernel during the convolution operation. To perform the convolution, the kernel slides across the input image (first to the right and then back to the beginning of the next line until the whole image is covered) with a step defined by the stride. The operation starts by a matrix multiplication between the elements of the kernel and the elements of the image it overlaps with. Finally, the sum of these values is the output (feature map).

It is important to note that, unless the kernel size is 1×1, the output will always be smaller than the input because of the boundary conditions. This can be a problem when multiple convolutions are applied successively. For that reason, and in order to prevent information loss, padding is usually applied. Zero padding is the most commonly used and consists in the addition of zeros on the borders of the image. By doing this, it is possible to either retain or increase the original dimensions of the image after the convolution.

Pooling layer

The pooling layer reduces the dimensions of the feature maps and is normally used after a convolutional layer, which is useful to extract dominant features. As in the convolutional layer, a kernel slides across the input image. There are two main types of pooling: max pooling and average pooling. In max pooling, the returned value is the maximum in the overlapped area. Therefore, it is also able to remove feature noise [68]. In the average pooling, the returned value is the average of all values in the overlapped area.

Activation function

The main purpose of an activation function is to introduce nonlinearities to the model [69]. This is important for the vast majority of applications since most of the phenomena studied are non-linear. One of the first activation functions was the sigmoid, defined as follows [70]:

$$\sigma(x) = \frac{1}{1+e^{-x}} \tag{3.1}$$

It is specially used in models that need to predict the probability as an output, since it exists between 0 and 1.

However, the sigmoid presents some issues, namely vanishing gradient issues [71]. Therefore, a more commonly used function is the rectified linear unit (ReLU) function, defined as [72]:

$$r(x) = max(0, x) \tag{3.2}$$

The ReLU has a reduced likelihood of the gradient to vanish, since it is always either 0 or 1.

As a final example, the softmax [73], often used as the activation function in the output layer of neural network models that predict a multinomial probability distribution. For an input vector, the softmax function gives a normalized vector of the same size where each element is a probability proportional to the exponentials of the input values.

3.3.2 Network architecture

The U-Net [53] is a CNN for fast and precise segmentation of images, originally created and used for biomedical imaging applications. This section provides a description of the base network used during the whole project, starting with an explanation of how the U-net works.

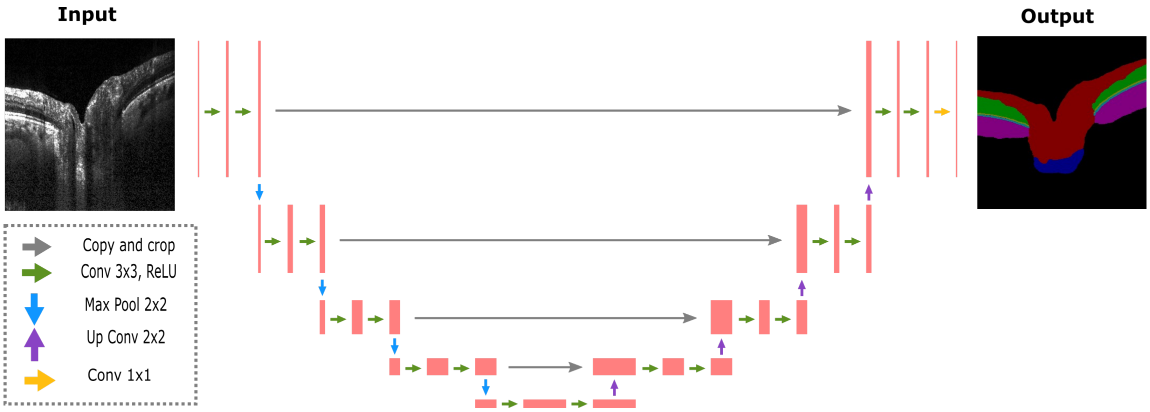


Figure 3.6: U-net architecture

In Figure 3.6 we see can see the U-net architecture. Each pink box corresponds to a multi channel feature map. The architecture's name comes from its U shape. It consists of a contracting path, on the left side, and an expansion path, on the right side.

Contracting path

The contracting path consists on the repeated application of 3 × 3 convolutions followed by a non-linear activation function (ReLU). During this work, padding was used to retain the same size of the input (320 × 320) after the convolutions. Following two 3 × 3 sequential convolutions, a 2 × 2 max pooling is applied. The max pooling operation reduces the size of the feature map. It acts on each channel separately and propagates the maximum activation from each 2 × 2 window to the next feature map. Therefore, after each max pooling operation, the number of feature channels is increased by a factor of 2, and the spatial dimensions cut in half.

The sequence of convolutions and max pooling operations results in an increase on feature channels and a decrease in the spatial dimensions. This means that, while we get closer to the content information on the image, we also move further from the spatial location of that information.

Expansion path

The expansion path creates a high resolution map. The up-convolutions use a learned kernel to map each feature vector to the 2×2 pixel output again, halving the number of feature channels. In order to yield a precise localization, the high resolution features from the contracting path are combined with the upsampled output. This is done through the skip connections (grey arrows in Figure 3.6), which concatenate the feature maps to help give localization information. Each concatenation is followed by two 3×3 convolutions, which are then followed by a ReLU.

The final layer uses a 1×1 convolution and a sigmoid activation function to map the channels into the desired number of classes. A 1×1 convolution [74] involves convolving over the entire image, pixel by pixel, with filters of size 1x1 and stride of 1. The 1×1 convolution is applied the same number of times as the desired number of classes. Afterwards, the image size will be $H \times L \times N$, where H is the height, L the length and N, the number of channels. That is, the sigmoid function is computed for each pixel of the image, assigning it a value between 0 and 1. The closer the value is to one, the higher the confidence level on its classification. In order to have this confidence level interpreted in a probabilistic way, a softmax function converts the sigmoid output into a probability, so that the sum of all the channels is always 1. The channel with higher value is the one that the network considers to be the correct layer.

Other considerations

Training networks can be challenging due to the changes in the distribution of the inputs to the layers when the weights are updated after each batch. These changes are referred to as internal covariate shift [75]. This problem was addressed by normalizing the layer inputs using batch normalization. The batch normalization stabilizes the learning process and reduces the number of training epochs required to train the network.

The Adam optimization algorithm [76], an extension of the stochastic gradient descent, was chosen as the procedure that iteratively updates the network weights. While the classic stochastic gradient descent uses a single learning rate for all weight updates during the course of the whole training, Adam optimizer leverages the power of adaptive learning rates methods to find individual learning rates for each parameter by updating this rate during the learning process.

Further specific details on the binary and multiclass segmentation training are provided in the following sections.

The network was implemented on Python using the Pytorch library [77] and CUDA [78]. The Navigator Cluster from the Laboratório de Computação Avançada (LCA) was used to run all necessary trainings and testings.

3.4 Binary segmentation

In order to get an insight into the layer segmentation complexity and understand which layers would be more complex for training a model, a binary segmentation approach was studied. In a binary segmentation problem, the output of the 1×1 convolutional layer of the U-net has only two channels: one for the background and one for each of the segmented labels individually.

As this was done in the initial stages of the project, not all manual segmentations were ready yet. Therefore, only a subset of B-scans from Dataset A was used: 65 (from 8 patients), 18 (from 2 patients), and 20 (from 2 patients), for training, testing and validation, respectively.

Each of the segmented tissues was isolated and used to train and test the network. Additionally, all labels were joined as a whole and used to train the network to find the segmented area of the images without having to differentiate between tissues.

At this stage of the project, the BM and the RPE were being segmented individually, instead of together as a complex. They were joined together later following the advice of the clinical advisors and that is how they are segmented in the rest of the project.

Different batch sizes (from 4 to 32), number of epochs (from 20 to 100) and learning rates (0.01 and 0.001) were experimented to assess which resulted in the best segmentation of the tissue in question. Only the best performances will be presented in the results.

3.5 Multiclass segmentation

The U-net used for the multiclass segmentation differs from the one used for the binary segmentation in the output layer. For the multiclass segmentation, the output layer has six channels: one for the background and one for each of the five labelled tissues.

From Dataset A, 151 images were used for training, 25 for validation and 30 for testing.

From Dataset B, 68 images were used for training, 11 for validation and 12 for testing while making sure that the model was tested on images from both healthy subjects and glaucoma patients.

Dataset A+B results from the combination of each of the previous corresponding sets together.

A total of 32 segmentation models were trained and tested for parameter tuning. The flowchart on Figure 3.7 illustrates the grid search with the combination of parameters used. The same combinations were used on images only from Dataset A, only from Dataset B and in both datasets together. For each dataset, all possible combinations between two learning rates (0.001 and 0.0001), two arrays with a different number of features filters ([32, 64, 128, 256, 512] and [64, 128, 256, 512, 1024]), and three loss functions (weighted Cross Entropy loss, Dice loss and Jaccard loss) were experimented. The parameters and ranges were chosen based on an initial optimization and on the range of parameters observed in the literature [10, 57, 58].

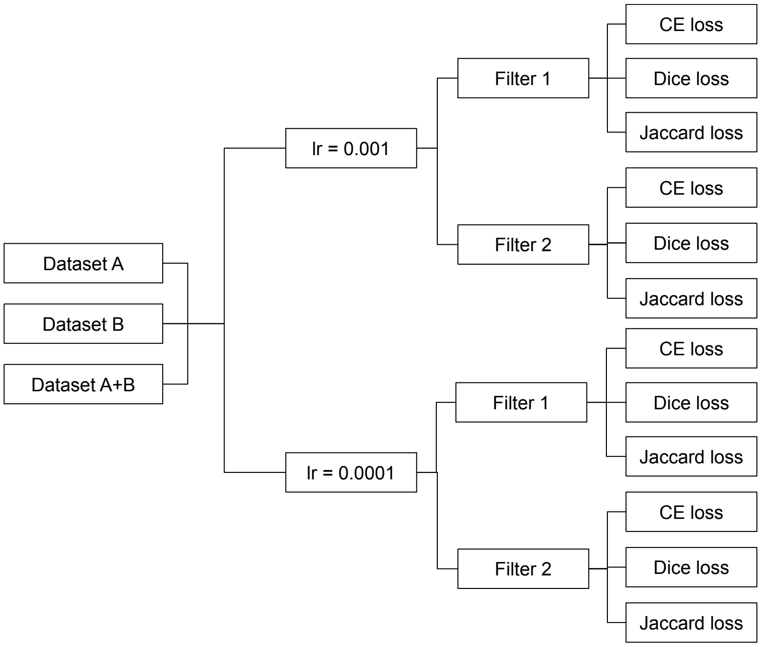


Figure 3.7: Flowchart with all the combinations of parameters used to train the models for each of the three datasets. lr = learning rate. CE = cross entropy.

A batch size of 16 was used for training and validation. The number of epochs was set to 250 but an early stopping function was added, which topped the training when the model stopped improving, avoiding overfitting of the training dataset. The early stopping function threshold was set on 50, meaning that the training stopped when the validation loss of the model did not get lower during 50 consecutive epochs.

In the end, the best performing models were identified based on the following criteria:

1. Overall test accuracy higher than 90%.

2. LC segmentation testing metrics above 60%.

3. BM/RPE complex segmentation testing metrics above 60%.

Five models showing the best combination of these three criteria were selected. The LC and BM/RPE complex segmentation metrics were chosen as discriminatory criteria given their importance in the computation of the biomarkers that depend on their correct segmentation. These five models results were analysed more in depth in order to find the best performing one to use for the biomarkers analysis.

A 5-fold cross validation was used only on the best model because of computational reasons, since training a model with cross validation consumes much more time (5-fold cross validation with the best model parameters took 11 hours and 23 minutes while without cross validation it took approximately 2 hours). Cross-validation is a resampling procedure used to evaluate models on limited data, allowing a better estimation of the model's performance. During cross-validation, the data is divided in N sets, or folds, where one is reserved for validation and the remaining N-1, for training. That ensures that the final performance is less dependant on the data division. In this case, the data was divided in 5 folds, patient-wise, and used to evaluate if the distribution of the patients in the training, validation and test sets had significant differences in the model's performance.

3.5.1 Data augmentation

Data augmentation was performed in order to account for the differences between images, both acquired with the same or a different protocol, and to overcome the sparsity of training data. The data augmentation transformations consisted of rotation, horizontal flipping, brightness and smoothing adjustments. These transformations will be described in detail in this section.

Custom data augmentation functions were created and chained together in the main training script. Each function was applied randomly with a 50% probability. The functions were applied only to the image (brightness and smoothing) or to both the image and the mask (rotation and horizontal flip). A mask is the segmentation image with the same size of the original, in which each pixel is labeled either as background or as one of the segmented tissues.

The data augmentation was implemented using the Python versions of the libraries torchvision [77] and OpenCV [79].

Horizontal flipping

In order to mimic right and left eye perspectives, a random horizontal flip transform was added to the data augmentation. This transform horizontally flips the given image randomly with a given probability.

Rotation

In order to account for the different perspectives the OCT images can be acquired and possible positions of the B-scan/cross-sectional images, a rotation transform was added to the data augmentation, which rotates the image with an angle between -8 and 8 degrees. This choice of degrees was based on similar works from the literature [58].

Brightness

For brightness and contrast adjustments, a look up table (LUT) function was used. The LUT is used to apply a gamma correction, following the equation:

$$output = input^{\frac{1}{\gamma}} \tag{3.3}$$

where $\gamma < 1$ will shift the image towards the darker end of the spectrum, $\gamma > 1$ will make the image appear lighter and $\gamma = 1$ will not affect the image.

A gamma value between 0.7 and 1.3 was chosen for each application of the function. First, a LUT is created that maps the input pixel values to the output gamma corrected values. Then the LUT is applied to the original image to find the correct mappings for each pixel value. Figure 3.8 shows two image examples to which the minimum and maximum gamma correction was applied, respectively. In Figures 3.8 (b) and (e), the BM and the LC, two important and challenging regions to correctly identify, are highlighted. While an highlighting of these regions can also be observed in Figures 3.8 (c) and (f), noise is also amplified. However, since noise is a very common artifact in OCT data, accounting for it in data augmentation is important, as it can contribute to a better generalization.

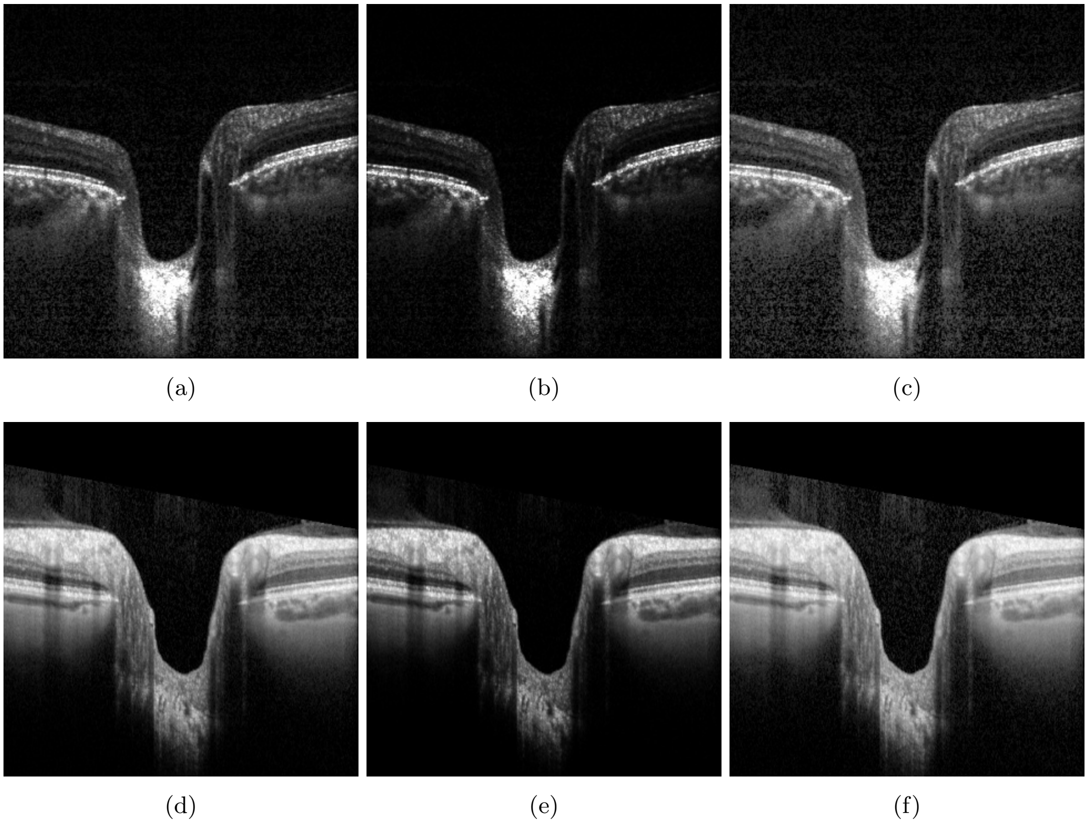


Figure 3.8: Top row: example of a B-scan from a volume from dataset A (a) before applying data augmentation and (b) after applying a $\gamma = 0.7$ and (c) $\gamma = 1.3$ correction with LUT. Bottom row: example of a B-scan from a volume from dataset B (c) before applying data augmentation and (d) after applying a $\gamma = 0.7$ and (e) $\gamma = 1.3$ correction with LUT.

Smoothing

Smoothness adjustments were added to the data augmentation using a bilateral filter. When the images are not well registered, the final averaged image may be slightly blurred. One way to resemble such effect is with smoothing. Blurring of an image can be achieved by convolving the image with a low-pass filter kernel. The bilateral filter was chosen because it is highly effective at noise removal while preserving edges. It uses a Gaussian filter in the space domain and another Gaussian filter component in the intensity domain, which ensures that only pixels with intensities similar to that of the central pixel are included to compute the blurred intensity value.

A 3×3 kernel and 75 sigma were used to obtain the images on Figure 3.9 (b) and (c). By looking at Figure 3.9, it is possible to notice a reduction on the noise after applying the filter. Moreover, the texture is more similar in both protocols after smoothing.

3.5.2 Loss function

A loss function compares the outputs of the model during the training with the desired outputs, and returns a quantification on how close these two values are. Thus, the loss is a representation of the error estimate of the model, and the larger the loss value, the farther the model is from its desired output. Three different loss functions were studied: multiclass cross entropy, Jaccard Index loss, and Dice loss.

The problem in study is one with unbalanced classes. This means that there is a disproportionate ratio of observations in each class due to the different sizes of each tissue labeled. This is particularly relevant in the background, which is the largest class (more than twice the other classes combined).

There are different options to deal with unbalanced classes in segmentation problems. One of them is by using weights, which was done in the cross entropy loss. Another approach is to use metrics that target the pixels in the foreground classes, such as Jaccard index and Dice coefficient. These losses are often used with unbalanced classes because their goal is the maximization of metrics that enforce positive classes (tissue, in our case).

Cross entropy

Cross entropy is a measure of the difference between two probability distributions. Its loss measures the performance of a classification model whose output is a probability value between 0 and 1. Cross entropy loss is a logarithmic loss. So, when the predicted probability diverges from the actual label, the loss value will increase. The formula is the following:

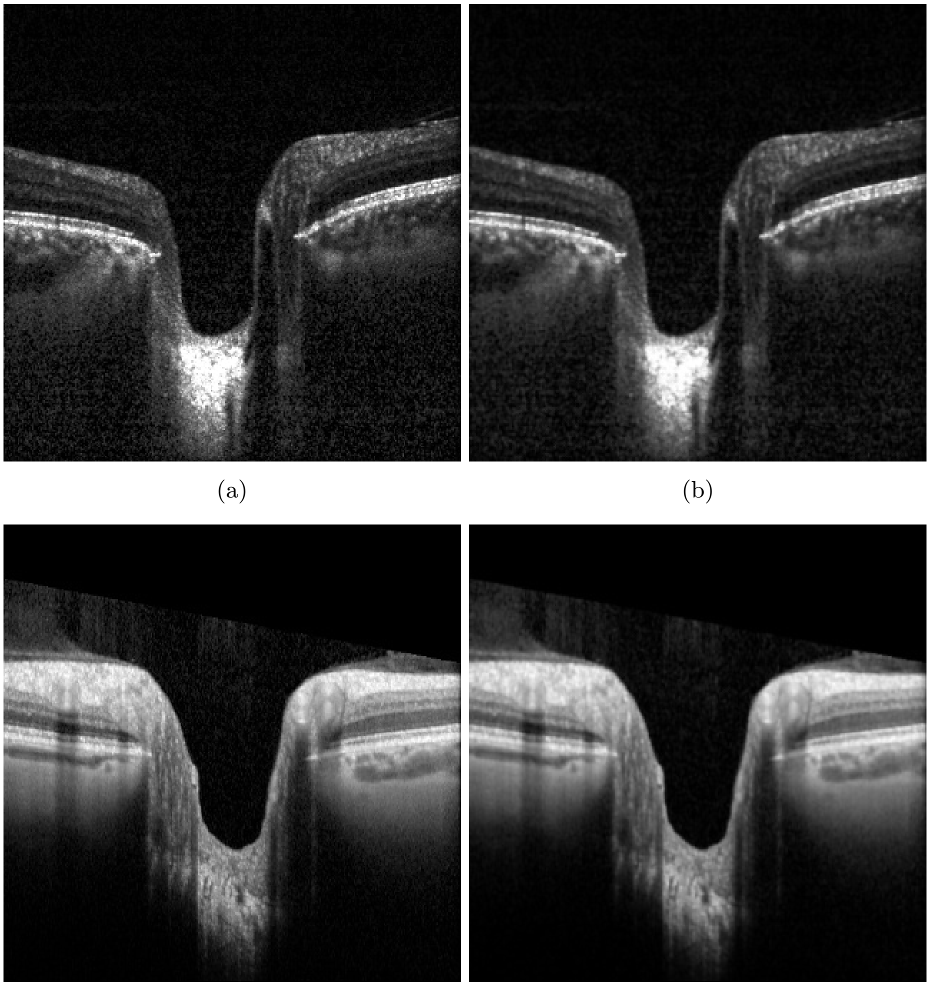


Figure 3.9: Top row: example of a B-scan from a volume from dataset A (a) before and (b) after applying a bilateral filter. Bottom row: example of a B-scan from a volume from dataset B (c) before and (d) after applying a bilateral filter.

$$loss(x, class) = weight[class] \left(-x[class] + log \left(\sum_j exp(x[j]) \right) \right) \tag{3.4}$$

$$\text{Cross entropy loss} = \frac{\sum_{i=1}^{N} loss(i, class[i])}{\sum_{i=1}^{N} weight(i, class[i])} \tag{3.5}$$

where $x[class]$ is the real probability distribution, $x[j]$ is the predicted probability distribution, N is the number of samples and i is the index of the sample.

One of the parameters of the function is the weights for each class, that allows to tackle the class imbalance. A higher weight, 1, was used for the BM and the LC labels, while 0.95 was used for the rest of the labels (including the background). These weights were decided empirically, as the ones that better fitted the data after an initial optimization. The BM and the LC were chosen to have the highest weights since they have the smallest pixel densities compared with the remaining tissues and are specially important for the computation of the biomarkers.

Jaccard index loss

The Jaccard index calculates a ratio, namely the intersection of the manual and predicted labels over their union. The intersection is the area where the manual and predicted labels overlap and the union is the area covered by at least one of the manual or predicted labels. So, the differences between intersection and union are where errors (either over or undersegmentation) are happening, and should be minimized. As the intersection increases so does the Jaccard index, signaling a more accurate overlap of the manual and predicted areas. Since the loss is a decreasing function, the Jaccard index loss formula is given by:

$$\text{Jaccard index loss} = 1 - \frac{\sum_{l=1}^{6} w_l \sum_n t_{nl} p_{nl}}{\sum_{l=1}^{6} w_l \sum_n (t_{nl} + p_{nl}) - t_{nl} p_{nl}} \tag{3.6}$$

where w_l is a computed weight used to provide invariance to different label set properties, t_n are the manual segmentation pixels and p_n are the prediction segmentation pixels.

Dice loss

The Dice coefficient is a measure of similarity. Similarly to the Jaccard index, it computes a ratio between the intersection and union of two sets: pixels belonging to the manual label and pixels belonging to the predicted label. Specifically, it computes twice the intersection over the sum of the manual and predicted areas. The goal is to simultaneously maximize the intersection and minimize the predicted area, since the manual area is constant. The Dice coefficient increases when the segmentation improves, so its loss function is given by [80]:

$$\text{Dice loss} = 1 - 2\frac{\sum_{l=1}^{6} w_l \sum_n t_{nl} p_{nl}}{\sum_{l=1}^{6} w_l \sum_n t_{nl} + p_{nl}} \tag{3.7}$$

where w_l is a computed weight used to provide invariance to different label set properties, t_n are the manual segmentation pixels and p_n are the prediction segmentation pixels.

3.5.3 Evaluation

In order to quantitatively assess the segmentation performance, the accuracy, Dice coefficient, specificity and sensitivity were computed for the testset results. All the metrics take values between 0 and 1, where the higher the value, the better. Accuracy was computed for the overall segmentation of all tissues excluding the background. The Dice coefficient, specificity and sensitivity were computed for each of the following tissues individually: the RNFL; the BM/RPE complex; all other retinal layers; the choroid and the LC.

The accuracy was computed using the confusion matrix. The accuracy measures how often a classifier is correct by dividing the sum of true negatives and true positives by the total of predictions. In a confusion matrix this translates to the sum of values on the diagnonal divided by the sum of all values on the matrix.

The Dice coefficient was used to assess the spatial overlap between the manual and automatic segmentations. For each tissue, the Dice coefficient was computed as:

$$\text{Dice coefficient} = \frac{2|T \cap P|}{|T| + |P|} \tag{3.8}$$

where T stands for the manual segmentation and P for the prediction of the network.

Specificity was used to assess the true negative rate of the predictive model. For each tissue, the specificity was computed as:

$$\text{Specificity} = \frac{|\bar{T} \cap \bar{P}|}{|T|} \times 100\% \tag{3.9}$$

where T stands for the manual segmentation and P for the prediction.

Sensitivity was used to assess the true positive rate and was defined, for each tissue, as:

$$\text{Sensitivity} = \frac{|T \cap P|}{|T|} \times 100\% \tag{3.10}$$

where T stands for the manual segmentation and P for the prediction.

3.6 Optic nerve head biomarkers

This sections describes the analysis for which the biomarkers were used, the methodology for automatic biomarkers extraction from the ONH and the statistical methods used for this analysis. Each biomarker was calculated in pixels and converted to micrometers according to the axial and lateral resolution of the respective OCT acquisition pattern. All the automatic segmentations used for the biomarkers extraction were obtained by the model identified as the best performing one.

The computed biomarkers were used for three purposes:

- **Examine the effect of the automatic segmentation in the biomarkers:** the biomarkers were computed on both the prediction and manual segmentations of the test set used for the chosen model, in order to evaluate the differences in the same image.
- **Comparison between healthy and glaucoma subjects:** the biomarkers were computed on automatic segmentations of all the available B-scans (even those that were not manually segmented and used to train the segmentation models). Then, these measurements were used to evaluate the existence of significant statistical differences between two groups - healthy subjects and glaucoma patients.
- **Comparison between POAG and NTG patients:** from the data used on the previous analysis, the portion that belonged to glaucoma patients only was isolated. Then, it was guaranteed that only B-scans from one volume per patient were present by selecting only the first volume. Moreover, it was assured that only eyes from NTG and POAG glaucoma patients with information regarding visual field and IOP were used. A total of 28 patients were used. The measurements were used to evaluate the existence of significant statistical differences between two groups - POAG and NTG. Information about the VF MD and the

IOP (commonly used in glaucoma evaluation, as described in Section 2.3) were also used to this analysis. Both parameters are known to be related with glaucoma detection/progression and, hence, are used as reference on the statistical analysis to see how informative a given biomarker truly is.

The last two evaluations used all the B-scans from each volume in order to introduce variability, despite the high correlation between B-scans from the same patient. One reason for this inclusion is that glaucoma progression is sectorial, meaning that, within the same volume, B-scans might be perceived differently depending on the acquisition angle, showing different signs of the disease (or even with patients showing no signs of disease at all in certain sectors). By including the complete 3D volume in the analysis, it was possible to overcome a smaller effective dataset, and to further explore the sensibility of the automatic segmentation model to variations on the ONH strucures.

3.6.1 Optic disc diameter

The optic disc diameter is computed as the horizontal distance between the two green lines on Figure 3.10, which are marked on the BM (segmented with pink in Figure 3.10) termination points. In order to find the optic disc diameter, all the *gaps* in the BM were detected and saved. A *gap* is any group of consecutive columns, or just an isolated column, where the BM is not labeled. This was necessary because the BM segmentation is neither always a continuous line until the BMO, nor is always visible in the B-scan borders. While going through all the columns of the B-scan, the indexes that did not contain any BM label were added to a list. Consecutive indexes were grouped as a *gap* and added to the dictionary. The biggest *gap* in the dictionary was defined as the optic disc diameter.

Being always calculated as an horizontal distance, the number of pixels in this biggest *gap* was multiplied by the lateral resolution to obtain the corresponding value in micrometers. Having the column coordinates of the two BM terminations, it was possible to find the row coordinate. Going through each of the two columns, the last row containing the BM labels was saved as the row coordinate of the BM termination. These two coordinates were saved to be used as reference in other biomarkers.

3.6.2 Retinal nerve fiber layer thickness

The RNFL thickness is calculated in a 3.4 mm diameter around the center of the optic disc [81,82]. The previously detected BM termination columns (described in 3.6.1) were used to find the middle point of the optic disc (red line on Figure 3.10). The distance from each column of the image to the middle point of the optic disc was computed and saved. This distance was transformed to millimeters after multiplying the number of pixels by the lateral resolution. The closest value to 1.7 mm in each side of the optic disc corresponds to the image column (illustrated with the yellow lines on Figure 3.10) where the RNFL (segmented in red on Figure 3.10) should be measured. Then, the RNFL thickness was computed as the mean number of pixels that contain the RNFL label in those two columns. The conversion was made by multiplying this value by the axial resolution, since it is a vertical distance.

3.6.3 Retinal nerve fiber layer area

The RNFL area is computed as the sum of pixels containing the RNFL label between an interval of image columns. This interval is calculated from the BM termination column to the column situated 1.7 millimeters from the center of the optic disc. This interval where the area of the red RNFL label is calculated can be seen in Figure 3.10, between the green and yellow vertical lines. Since it is an area, the unit conversion is made by multiplying the value in pixels by the axial and lateral resolutions.

The RNFL area is very dependent on the eye anatomy and, therefore, very patient-specific. For that reason, it is a biomarker more suitable for longitudinal studies and follow-up of the same patient. However, since in this case, an interval was introduced that guaranteed the RNFL area was always calculated inside previously defined inner and outer limits, the limitation was surpassed and the biomarker was used in comparisons between patients. The interval guaranteed that the biomarker was measured on the same anatomical region in all patients and excluded the RNFL accumulated on the optic disc, which is a source of more variability.

This approach evaluate the RNFL is not standard and was not found in the literature. It was computed as an alternative to the already used RNFL thickness based on clinical advice.

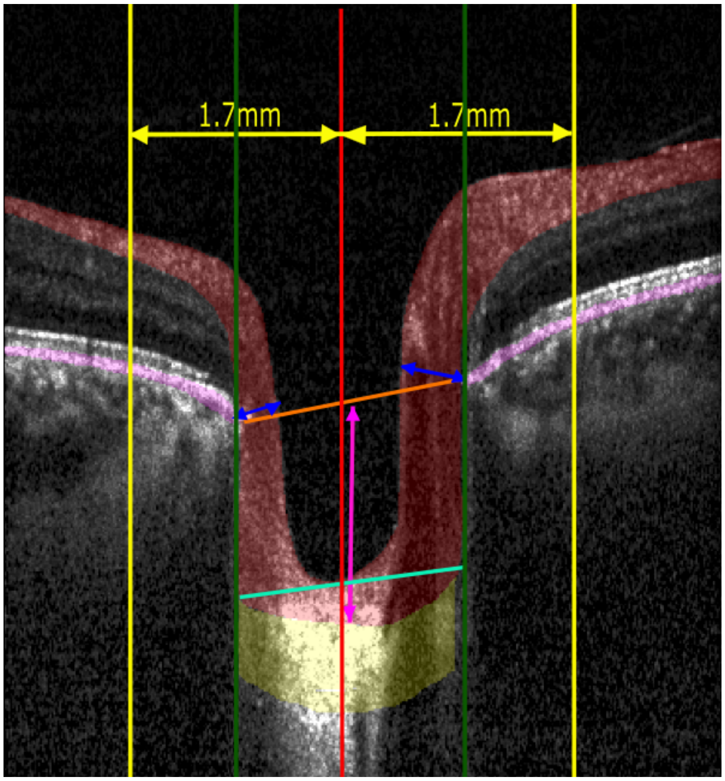


Figure 3.10: Reference lines used to calculate the biomarkers. In red, the line that marks the center of the optic disc. In yellow, the lines that mark a 3.4mm diameter around the optic disc center. In green, the BM reference lines for the termination points. In dark blue, the BMO-MRW. In orange, the BM reference line. In pink, the LC depth. In light blue, the LC insertion points reference line. The RNFL, RPE/BM complex and the LC segmentations are highlighted in red, pink and yellow, respectively.

3.6.4 Bruch's membrane opening - minimum rim width

The first step to compute this biomarker was to locate and save on a list all the x, y coordinates of the top layer of the RNFL label (in red on Figure 3.10). The Euclidean distance from each of the two BM termination points (the end points of the orange line in Figure 3.10 saved in the optic disc diameter computation 3.6.1), to each one of the RNFL top layer coordinates was computed and saved. The result was two lists of distances, one for each side of the ONH opening. The shortest distance on each of the two lists was the BMO-MRW (dark blue arrows in Figure 3.10). So, in the end, there are two values, one for each side of the ONH. Since the BMO-MRW distance is diagonal, the following formula was used for the conversion of the number of pixels:

$$\text{BMO-MRW}_{micrometers} = \sqrt{(dx \times lateral\ resolution)^2 + (dy \times axial\ resolution)^2} \qquad (3.11)$$

where dx and dy are the horizontal and vertical components of the distance, respectively.

3.6.5 Lamina cribrosa depth

The computation of the LC depth (pink line on Figure 3.10) uses as starting point the reference line connecting the two BM termination points (computed as described in 3.6.1). The columns which contain any LC label (in yellow on Figure 3.10) were selected and saved. In each of these columns, the distance from the BMO reference line (orange line in Figure 3.10) to the first row that contains the LC label is computed. The sum of all these distances is in the end divided by the number of columns containing the LC label to find its mean depth. As this distance is always vertical, the number of pixels is multiplied by the axial resolution for conversion.

3.6.6 Lamina cribrosa curvature index

For the LC curvature, the reference line connecting the LC insertion points was defined (light blue line on Figure 3.10. The LC insertion points are the first points labeled as LC on the two extreme LC columns. Next, the distances from the insertion points reference line to the first row with the LC were computed. Finally, the sum of these distances was divided by the length of the LC insertion points reference line, and multiplied by 100 to find the curvature index as described in section 2.4.2.

3.6.7 Statistical analysis

Statistical analysis was performed to assess the significance of the findings in the different experiments. First, to validate the segmentation, the 25, 50 and 75 percentile of the measurements from the manual and automatic segmentations to examine how close they were from each other and to the values reported in the literature. The correlation coefficient between pairs was computed to examine how much the measurements from the predictions resembled the ones from the ground truth, and to quantify the inter-grader variability effect on the biomarkers.

Additionally, the data from the biomarkers calculated from the manual and automatic segmentations was examined on a Bland-Altman plot and in a scatter plot with a linear regression line. The Bland-Altman plot [83] describes the agreement between two quantitative measurements by constructing limits of agreement. To that end, it uses the mean and the standard deviation of the differences between the two measurements. It is common to compute 95% limits of agreement for each comparison (average difference ± 1.96 standard deviation of the difference), which indicates how far apart measurements by the two methods were. Whether a 95% limit of agreement is significant in a clinical context or not, requires further analysis from experts. This approach has been considered more adequate than the correlation coefficient when assessing the comparability between methods [84].

For the two classification analyses, healthy vs. glaucoma and NTG vs. POAG, the data was visualized in box plots since the observations were no longer paired and the size of each group differed.

First, the Shapiro Wilk test [85] was used to assess whether the data was normally distributed, using a significance level 0.05. Depending on the result of the Shapiro Witk test, either the t-test or the Mann Whitney test [86] were used to assess if there was a significant difference between two normally or not normally distributed groups, respectively (significance level 0.05). Finally, additional unpaired t-tests were used to examine the correlation coefficient between groups, the Pearson correlation and the Spearman correlation [87]. The Spearman correlation differs from the Pearson in that it does not assume that both datasets are normally distributed. The groups were considered correlated for p-values > 0.05. Both formulations of the correlation coefficient vary between -1 and 1. Negative correlations imply that when x increases, y decreases; positive correlations imply that both groups vary in the same direction.

Chapter 4

Results

This chapter presents the results for the manual and automatic segmentation of the ONH in OCT data, organized in four sections. The first section shows the inter-grader variability in manual segmentations. The second section describes the results obtained for segmenting each layer individually, using the binary segmentation model. The third section presents the results from the multiclass segmentation models, which are used to make the selection of the best performing model, that will be applied to obtain the automatic segmentations used in the biomarkers analysis. Finally, the fourth section details the analysis of the biomarkers, which comprises a validation of the segmentation model by comparing the biomarkers computed in the manual and automatic segmentations of the same image, and an assessment of the diagnostic power of the biomarkers by comparing their values between data from healthy and glaucoma subjects.

4.1 Manual segmentation

Table 4.1 shows the Dice coefficient, sensitivity, specificity and accuracy results for the inter-grader variability assessment. The lowest Dice coefficient and sensitivity are highlighted in bold, showing how the LC is the structure with more disagreement between graders. However, the Dice coeffcient and sensitivity standard deviation for the LC is considerably large, hinting that there were cases where the agreement was almost as high as on the other tissues (0.95-1) and other cases with metrics between 0.55 and 0.6 where graders had more doubts. After the LC, the choroid segmentation is the structure that required more correction, with a 0.94 Dice coefficient and 0.92 sensitivity. The same effect observed on the LC standard deviations can be observed for the choroid, although on a lower scale. Specificities are always higher than 0.99 which means that in this particular analysis, one of the graders was always under-segmenting, or one was always over-segmenting.

The influence that the inter-grader variability has on the biomarkers computation is depicted in Figure 4.1 with individual scatter plots for each of the six biomarkers. The correlation coefficient was 0.99 for all the biomarkers except for those related to the LC, that is, LC depth and LCCI, for which the correlations were 0.87 and 0.34, respectively. This is consistent with the results presented in Table 4.1 for the differences in the manual segmentations between both graders, which also show large differences only in the LC.

Table 4.1: Inter-grader variability assessment. Dice coefficient, sensitivity and specificity for each structure and overall accuracy excluding the background.

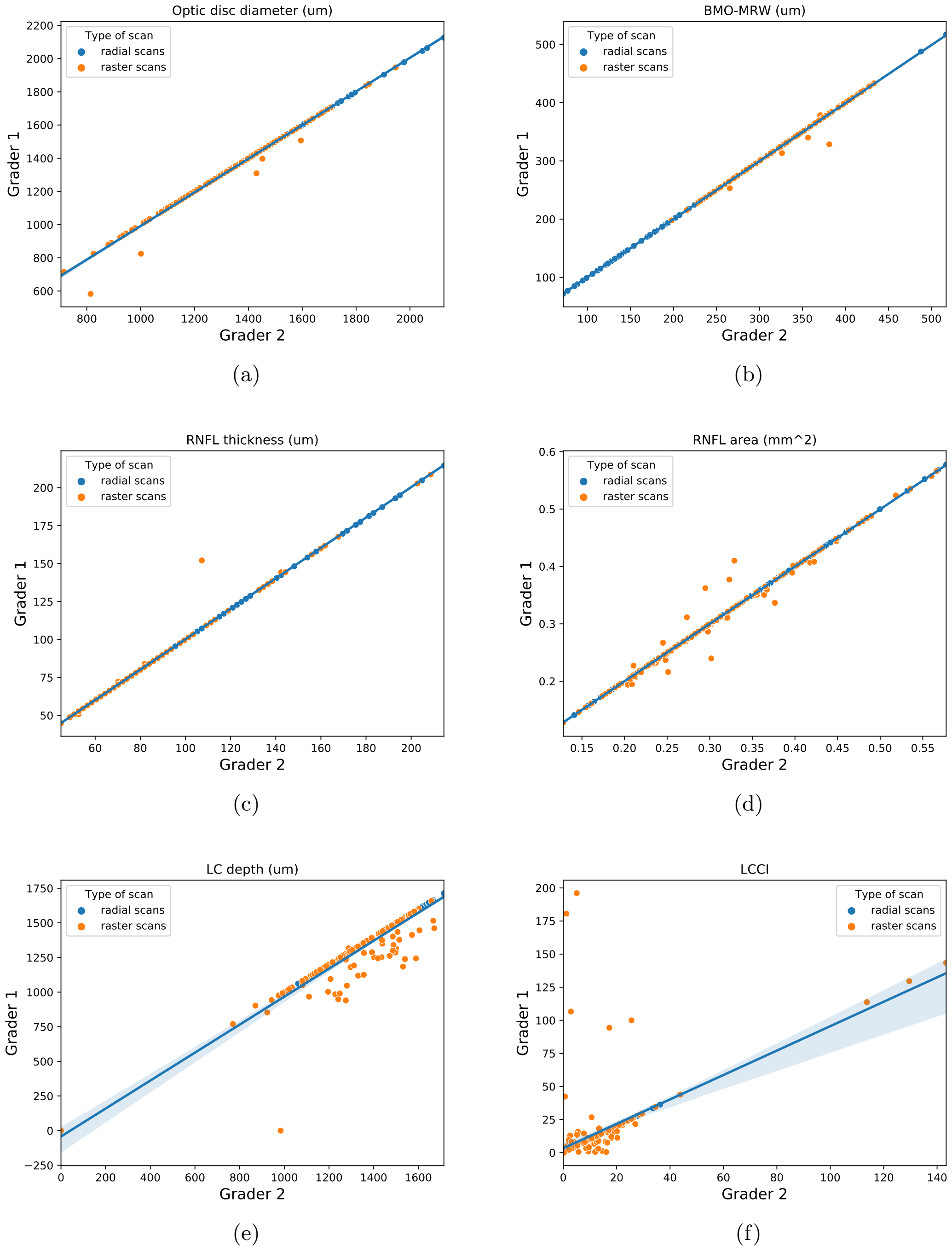
Tissue	Dice Coefficient (Mean ± SD)	Sensitivity (Mean ± SD)	Specificity (Mean ± SD)	Accuracy
RNFL	0.98 ± 0.04	0.97 ± 0.07	1.00 ± 0.00	0.988
Other retinal layers	0.99 ± 0.01	0.99 ± 0.01	1.00 ± 0.00	
BM/RPE complex	0.99 ± 0.00	1.00 ± 0.01	1.00 ± 0.00	
Choroid	0.94 ± 0.11	0.92 ± 0.17	0.99 ± 0.00	
LC	**0.86 ± 0.29**	**0.86 ± 0.30**	0.99 ± 0.01	

4.2 Binary segmentation

Since the binary segmentation model presented a worse performance than the multiclass approach in all the tissues and, particularly, in the LC, this approach was not pursued further, and the results are not discussed in depth. For reference, examples of the predicted images with the binary segmentation model can be seen in the Appendix A (Figures A.1, A.2, A.3, A.4, A.5 and A.6). The only structure missing in the result figures is the LC, as the output for the LC was 100% background for all the test images.

All the presented results showed a train and validation accuracy above 0.85, except for the results with all labels together as an whole, which accuracy was only above 0.65. However, this accuracy was computed including the background. Regarding the train and validation loss, it was between 0.1 and 0.5, except for the LC, for which it was 1 in both cases.

A batch size of 16 and 8 for train and validation, respectively, was used to obtain the results presented for the other retinal layers, the LC, the BM and all the labels together as whole while



(a)

(b)

(c)

(d)

(e)

(f)

Figure 4.1: Inter-grader variability effect in the computation of the biomarkers. The scatter plots compare the measurements obtained in the manual segmentation from each of the graders on (a) optic disc diameter ($r^2 = 0.99$), (b) BMO-MRW ($r^2 = 0.99$), (c) RNFL thickness ($r^2 = 0.99$), (d) RNFL area ($r^2 = 0.99$), (e) LC depth ($r^2 = 0.87$), and (f) LCCI ($r^2 = 0.34$).

only a size 8 and 4 for train and validation, respectively, was used for the RNFL and choroid results. All the presented results were obtained with a 0.001 learning rate and binary cross entropy loss function.

4.3 Multiclass segmentation

The results for the selection of the five best models can be found in Table 4.2. For each model, the dataset, the parameters used to train the model, and the training and validation loss, accuracy, and Dice coefficient, are presented. Additionally, the accuracy was computed for the test set, excluding the background.

Table 4.2: Results from the five best performing models. The variable parameters used in each are presented. The final loss, accuracy and Dice coefficient (DC) values obtained in training and validation are presented. The testing accuracy corresponds to the overall segmentation excluding the background.

ID	Dataset	Learning rate	Filter	Loss function	Training			Validation			Testing accuracy
					Loss	Accuracy	DC	Loss	Accuracy	DC	
1	A	0.0001	2	Dice loss	0.34	0.88	0.92	0.28	0.87	0.95	0.95
2	A	0.0001	2	Jaccard loss	0.45	0.88	0.93	0.39	0.91	0.95	0.95
3	B	0.0001	2	Dice loss	0.40	0.82	0.91	0.37	0.84	0.92	0.94
4	A + B	0.001	1	CE loss	0.30	0.87	0.92	0.24	0.91	0.95	0.95
5	A + B	0.0001	2	Dice loss	0.34	0.86	0.91	0.28	0.88	0.94	0.94

Table 4.3 shows the results (Dice coefficient, sensitivity, and specificity) obtained in the test set for each labelled region individually for each of the models in Table 4.2. The best results for each tissue are highlighted in bold. It can be observed that model 2 showed the best Dice coefficient for all structures except the choroid, for which model 1 and 4 performed better. Sensitivities were more varied, with higher values for the RNFL in model 1 and 5, for the other retinal layers in model 2, for the BM/RPE complex in model 5, for the choroid in model 2 and for the LC in model 3. Specificities were always higher than 0.98.

Overall, the five models had very similar training and testing results. In particular, models 1, 3, and 5 can be directly compared, since they share all the same training parameters, with the only difference of being trained in different datasets. It can be observed how the performance in terms of both global test accuracy and metrics computed in specific structures is very similar for the three datasets with the same configuration, and particularly for models 1 and 5, hinting that this parameter configuration is less optimal when training with dataset B alone.

Table 4.3: Mean Dice coefficients, sensitivity and specificity for all structures across all five models from 4.2 when evaluated on the test set. The best result for each structure is highlighted in bold.

Metrics	Tissue	Model ID				
		1	2	3	4	5
Dice Coefficient (Mean ± SD)	RNFL	0.92 ± 0.04	**0.93 ± 0.03**	0.83 ± 0.06	0.91 ± 0.04	0.91 ± 0.05
	Other retinal layers	0.89 ± 0.05	**0.91 ± 0.04**	0.79 ± 0.08	0.88 ± 0.06	0.88 ± 0.06
	BM/RPE complex	0.71 ± 0.15	**0.73 ± 0.15**	0.63 ± 0.08	0.66 ± 0.14	0.70 ± 0.14
	Choroid	**0.87 ± 0.06**	0.84 ± 0.07	0.83 ± 0.06	**0.87 ± 0.05**	0.82 ± 0.08
	LC	0.64 ± 0.25	**0.66 ± 0.23**	**0.66 ± 0.13**	0.59 ± 0.25	0.63 ± 0.22
Sensitivity (Mean ± SD)	RNFL	**0.93 ± 0.03**	0.92 ± 0.04	0.85 ± 0.07	0.91 ± 0.04	**0.93 ± 0.04**
	Other retinal layers	0.93 ± 0.04	**0.94 ± 0.03**	0.74 ± 0.13	0.89 ± 0.09	0.86 ± 0.09
	BM/RPE complex	0.75 ± 0.12	0.80 ± 0.11	0.77 ± 0.12	0.57 ± 0.13	**0.89 ± 0.10**
	Choroid	0.89 ± 0.05	**0.94 ± 0.02**	0.87 ± 0.08	0.90 ± 0.06	0.92 ± 0.04
	LC	0.63 ± 0.26	0.66 ± 0.23	**0.69 ± 0.13**	0.59 ± 0.25	0.67 ± 0.23
Specificity (Mean ± SD)	RNFL	0.98 ± 0.01	0.99 ± 0.01	0.98 ± 0.01	0.99 ± 0.01	0.98 ± 0.01
	Other retinal layers	0.99 ± 0.00	0.99 ± 0.00	0.99 ± 0.00	0.99 ± 0.00	0.99 ± 0.00
	BM/RPE complex	0.99 ± 0.00	0.99 ± 0.00	0.99 ± 0.00	0.99 ± 0.00	0.99 ± 0.00
	Choroid	0.99 ± 0.01	0.98 ± 0.01	0.99 ± 0.01	0.99 ± 0.01	0.98 ± 0.01
	LC	0.99 ± 0.01	0.99 ± 0.01	0.99 ± 0.00	0.99 ± 0.01	0.99 ± 0.01

The first three models have been trained on a single set. In order to analyse their generalization capabilities on the second dataset, models 1, 2, and 3 were tested on data from the dataset that was not used for their training. The results, in Table 4.4, show an evident drop in testing accuracy and, specially, on the metrics computed on individual tissues. The best metrics for each structure are highlighted in bold, showing that the models trained with dataset A (model 1 and 2) perform better when segmenting images from a different acquisition protocol than a model trained with images from dataset B. Specificities were always higher than 0.90. This confirms that a model trained in one dataset cannot generalize well enough, so models trained in both types of data should be considered.

Models 4 and 5, trained in both datasets, showed very similar performances. The testing metrics reported in Table 4.3 did not show large differences, except for the BM/RPE complex segmentation sensitivity, for which model 5 performed significantly better. For this reason, and considering the importance of this region, model 4 was excluded, and model 5 was chosen for the subsequent analysis. Therefore, from this point onward model 5 will be referred to simply as *model*.

Once this selection on the best model was done, and in order to obtain a better estimation of its accuracy, a five-fold cross-validation approach was used. The mean results for all five folds regarding training and validation can be found in Table 4.5. It is possible to verify that the standard deviations for the loss, accuracy and Dice coefficient results for both training and validation are always very small, showing that the choice of patients used for training and validation did not have a strong influence in the performance of the model.

Table 4.4: Testing results for when model 1 and 2 are tested on data from Dataset B and model 3 is tested on data from Data set A. Mean Dice coefficients, sensitivity and specificity are showed for all tissues and the accuracy for the overall segmentation excluding the background.

Metrics	Tissue	Model 1	Model 2	Model 3
		Tested on dataset:		
		B	B	A
Accuracy	All	0.85	0.84	0.78
Dice Coefficient (Mean ± SD)	RNFL	0.63 ± 0.08	0.67 ± 0.08	0.50 ± 0.12
	Other retinal layers	**0.70 ± 0.11**	0.66 ± 0.12	0.32 ± 0.18
	BM/RPE complex	**0.53 ± 0.08**	0.51 ± 0.08	0.36 ± 0.14
	Choroid	0.66 ± 0.14	**0.68 ± 0.10**	0.48 ± 0.15
	LC	0.36 ± 0.18	**0.48 ± 0.13**	0.14 ± 0.18
Sensitivity (Mean ± SD)	RNFL	**0.85 ± 0.07**	0.77 ± 0.11	0.53 ± 0.12
	Other retinal layers	**0.69 ± 0.16**	0.66 ± 0.13	0.22 ± 0.13
	BM/RPE complex	0.48 ± 0.22	**0.63 ± 0.21**	0.28 ± 0.13
	Choroid	0.57 ± 0.18	**0.77 ± 0.11**	0.38 ± 0.15
	LC	**0.62 ± 0.18**	0.61 ± 0.13	0.11 ± 0.14
Specificity (Mean ± SD)	RNFL	0.90 ± 0.05	0.94 ± 0.04	0.91 ± 0.02
	Other retinal layers	0.99 ± 0.01	0.99 ± 0.01	0.99 ± 0.00
	BM/RPE complex	0.99 ± 0.00	0.99 ± 0.00	0.99 ± 0.00
	Choroid	0.99 ± 0.00	0.97 ± 0.01	0.99 ± 0.01
	LC	0.97 ± 0.01	0.98 ± 0.01	0.99 ± 0.00

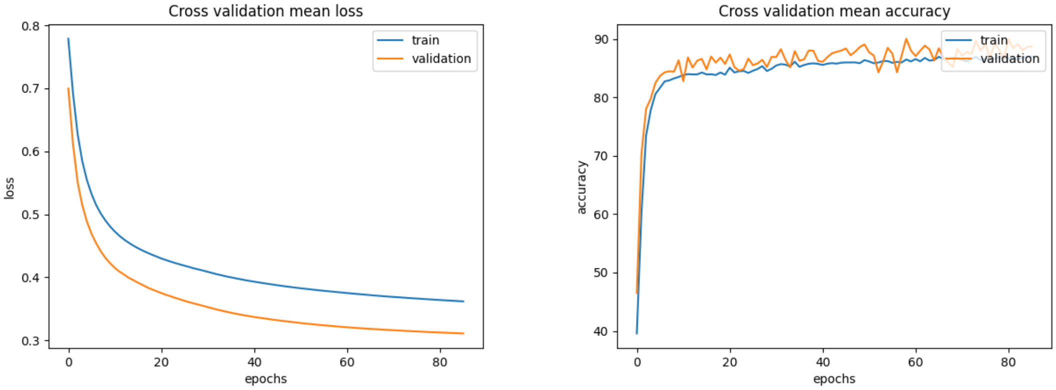
Table 4.5: Mean training and validation results from the 5-fold cross validation using the same training variable parameters used for model 5.

5-fold cross validation mean results					
Training			Validation		
Loss	Accuracy	DC	Loss	Accuracy	DC
0.35 ± 0.03	0.87 ± 0.01	0.91 ± 0.01	0.30 ± 0.03	0.88± 0.02	0.93 ± 0.01

Finally, to confirm that the model was fully trained and converged correctly, the plots for the cross validation mean loss and accuracy are depicted in 4.2, showing how the model stopped improving between epoch 80 and 90.

4.4 Optic nerve head biomarkers

In this section, the main results from the biomarkers' extraction are presented. The experiments involving biomarkers were divided in three subsections. First, the biomarkers are calculated in both the manual segmentations and the predictions given by the automatic segmentation model, and the two results for each image are compared in order to validate how much the automatic segmentation affects the biomarkers measurements. Then, the biomarkers' ability to separate healthy from glaucoma and NTG from POAG in automatically segmented B-scans is statistically analysed.



(a)

(b)

Figure 4.2: Results from the 5-fold cross validation. On (a) the mean training and validation loss and on (b) the mean training and validation overall accuracy.

4.4.1 Analysis of the automatic segmentation bias

The six features previously defined in Section 3.6 (optic disc diameter, BMO-MRW, RNFL area, RNFL thickness, LC depth, and LCCI) were automatically extracted from the segmentations, both ground truth and predictions, obtained in the test set.

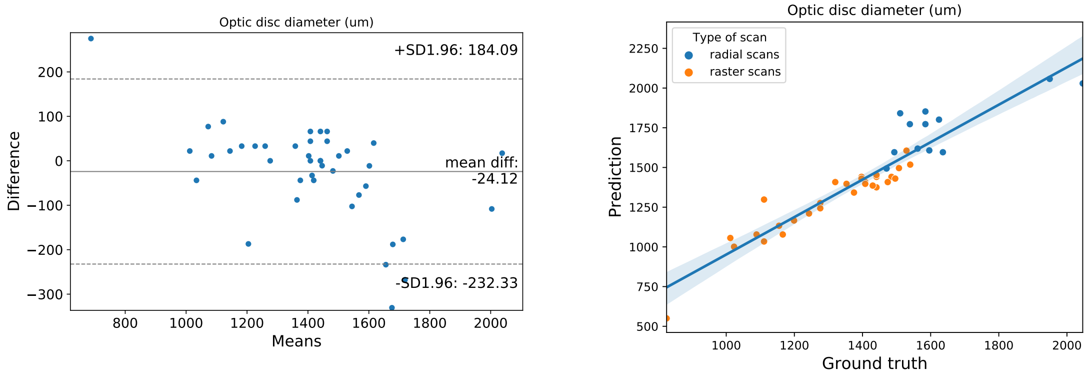
Table 4.6 shows the 25, 50 and 75 percentiles for all measurements on each group and the correlation coefficient of each biomarker between the pairs of segmentations (ground truth, prediction). The percentiles are highly comparable between the two groups in all the biomarkers except the LCCI. The biomarker showing the highest correlation is the BMO-MRW, while the lowest belongs to the LCCI.

Table 4.6: Results of the biomarkers extracted from the segmentations of the ground truth and the predictions given by the best model. The 25th (P25), 50th (P50) and 75th (P75) percentiles of each biomarker are presented. The correlation coefficient (r^2) between the ground truth measurements and its corresponding prediction is also presented.

Biomarker	Ground truth			Prediction			r^2
	P25	P50	P75	P25	P50	P75	
Optic disc diameter (μm)	1276	1441	1524.4	1281.5	1430	1596	0.88
BMO-MRW (μm)	234.2	286.1	346.7	248.3	298.8	358.3	0.94
RNFL area (mm^2)	0.26	0.32	0.42	0.27	0.33	0.411	0.87
RNFL thickness (μm)	84.8	104.3	135.5	79.9	106.3	142.8	0.82
LC depth (μm)	1273.3	1429.4	1575.4	1269.3	1432.7	1541.5	0.61
LCCI	4.9	11.3	15.3	2.5	8.3	18.5	0.0066

The Bland-Altman and scatter plots for the comparison between (ground truth, prediction) pairs for each biomarker is presented next. Since the test set contained both radial and raster scans, they are identified with different colors in the scatter plots.

The results for the optic disc diameter are shown in Figure 4.3. In Figure 4.3 (a), the Bland-Altman plot shows that the worst results (further from the middle line that represents a 0 difference between the biomarker in the prediction and in the ground truth segmentation) happen for larger discs. In Figure 4.3 (b) it is possible to verify that larger discs belong to radial scans.



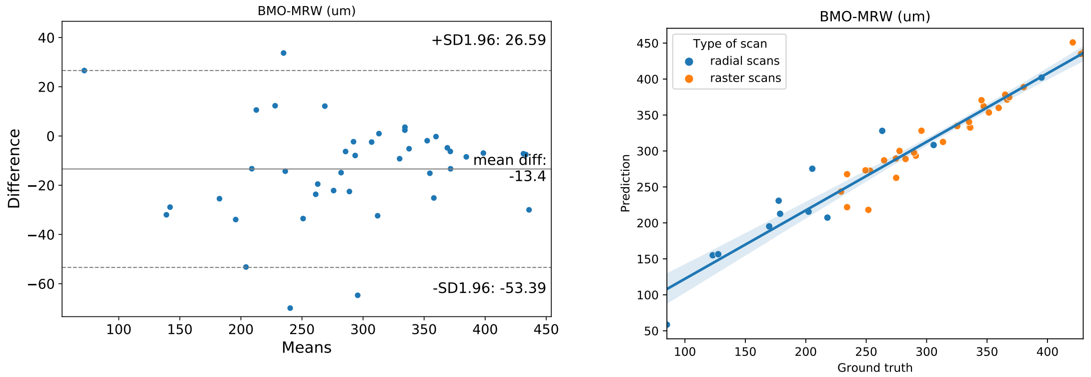
(a)

(b)

Figure 4.3: Comparison between the optic disc diameter obtained from the ground-truth and the model prediction on the test set. (a) Bland-Altman plot; (b) scatter plot with the linear regression line ($r^2 = 0.88$).

The results for the BMO-MRW are shown in Figure 4.4. It is possible to verify that larger BMO-MRW values belong to raster scans. In Figure 4.4 (a) is possible to verify that larger errors happen for shorter BMO-MRW distances but also that overall the points are inside the agreement limits (whether those agreement limits are acceptable in clinical context is a matter of biological factors). In Figure 4.4 (b) it is possible to see that the scatter plot points are all close to the linear regression line, in agreement with the 0.94 correlation coefficient.

The BMO-MRW is a distance calculated at each side of the ONH, the temporal and the nasal. However, in Figure 4.4 the results are shown for the average of both sides. In order to know which side we are looking at in a B-scan, nasal or temporal, we need information about which eye it belongs to and, in the radial scans case, at which angle was it taken. Unfortunately, this information about the angle was not available in the radial dataset, which motivated the decision of combining both sides to provide a global view of the biomarker. Nevertheless, Figure 4.5 shows the results for the



(a)

(b)

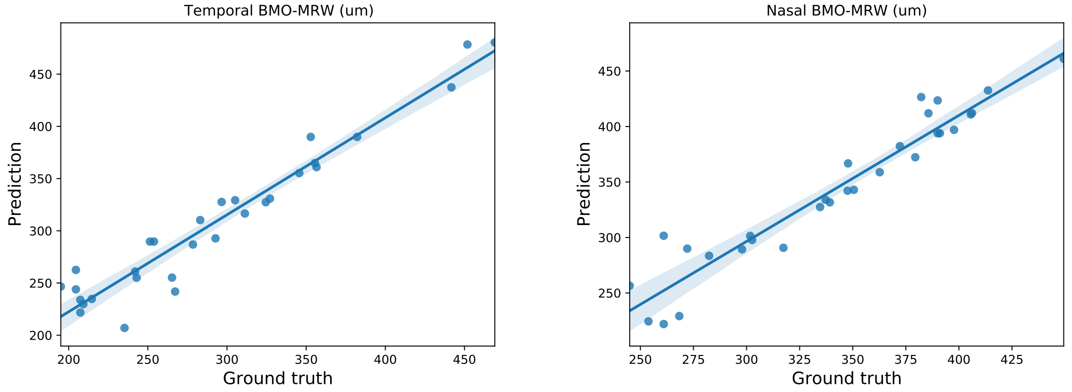
Figure 4.4: Comparison between the BMO-MRW obtained from the ground-truth and the model-5 prediction on the test set. On (a) the Bland-Altman plot and on (b) the scatter plot with the linear regression line ($r^2 = 0.94$).

temporal and nasal BMO-MRW in the raster scans portion of the data. It is possible to verify a high correlation between manual and automatic segmentations at both sides.

The results for the RNFL thickness and area are depicted in Figures 4.6 and 4.7, respectively. The RNFL area shows a higher correlation coefficient ($r^2 = 0.87$ over $r^2 = 0.82$). Both biomarkers show a different behaviour regarding the distribution of values for the different datasets, as smaller RNFL thicknesses can be found in raster scans while this trend is not evident for the RNFL area, where the points are more uniformly spread.

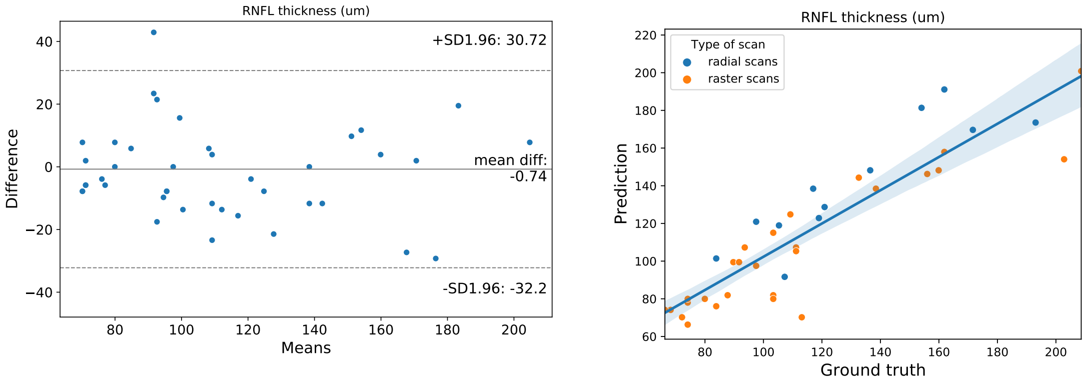
The results for the LC depth are presented in Figure 4.8. Despite showing small differences between measurements in the Bland-Altman and scatter plot, it still presented one of the lowest correlation coefficients ($r^2 = 0.61$). However, when looking at the plots, it was possible to isolate three outliers, that is, three pairs of measurements that showed significantly larger differences when compared to the other values. These cases are shown and analysed in Figures 4.9, 4.10, and 4.11.

The first and second cases (Figures 4.9 and 4.10) belong to a radial scan. The LC segmentation in the first outlier is similar to the ground truth in the central part, but the lateral boundaries are different due to the differences in illumination, that are interpreted as a tissue boundary by the model, but not by the grader. In the second outlier, the segmentation is overall very similar to the ground truth. However, it is also wrongly segmented on the top of the image due to a reflection artifact, misleading the LC depth measurement. Finally, the third case (Figure 4.11) belongs to



(a) (b)

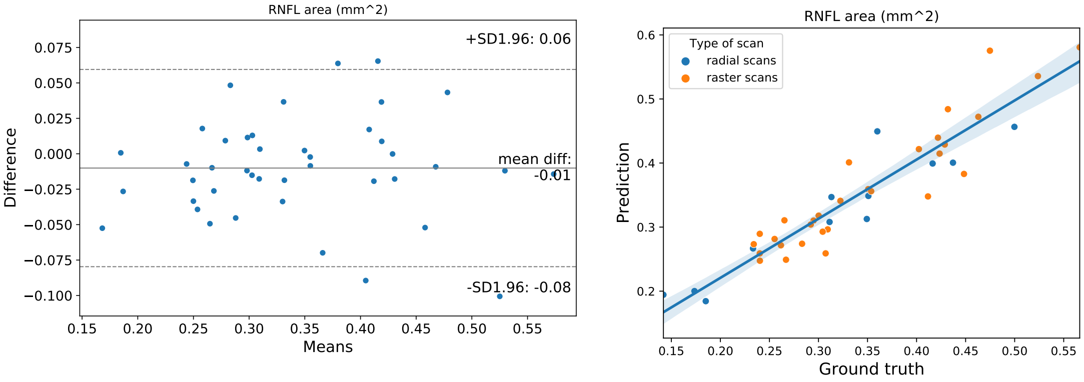
Figure 4.5: Comparison between the BMO-MRW in the ground truth segmentation and corresponding model prediction for the raster scans in the test set. The results are shown in the scatter plots with a linear regression line. (a) Temporal BMO-MRW ($r^2 = 0.93$); (b) nasal BMO-MRW ($r^2 = 0.92$).



(a)

(b)

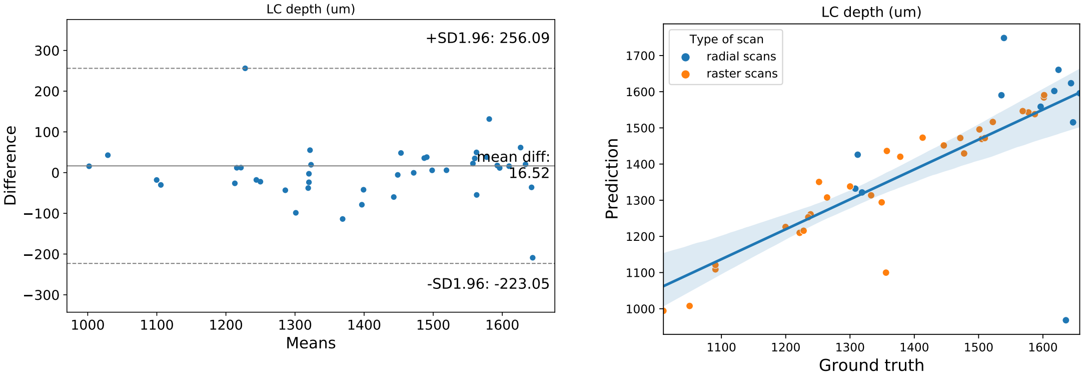
Figure 4.6: Comparison between the RNFL thickness obtained from the ground-truth and the model prediction on the test set. (a) Bland-Altman plot; (b) scatter plot with the linear regression line ($r^2 = 0.82$).



(a)

(b)

Figure 4.7: Comparison between the RNFL area obtained from the ground-truth and the model prediction on the test set. (a) Bland-Altman plot; (b) scatter plot with the linear regression line ($r^2 = 0.87$).



(a)

(b)

Figure 4.8: Comparison between the LC depth obtained from the ground-truth and the model prediction on the test set. (a) Bland-Altman plot; (b) scatter plot with the linear regression line ($r^2 = 0.61$).

a raster scan. Similarly to the second example, there is a region in the top of the image wrongly labelled LC, which also misleads the calculations.

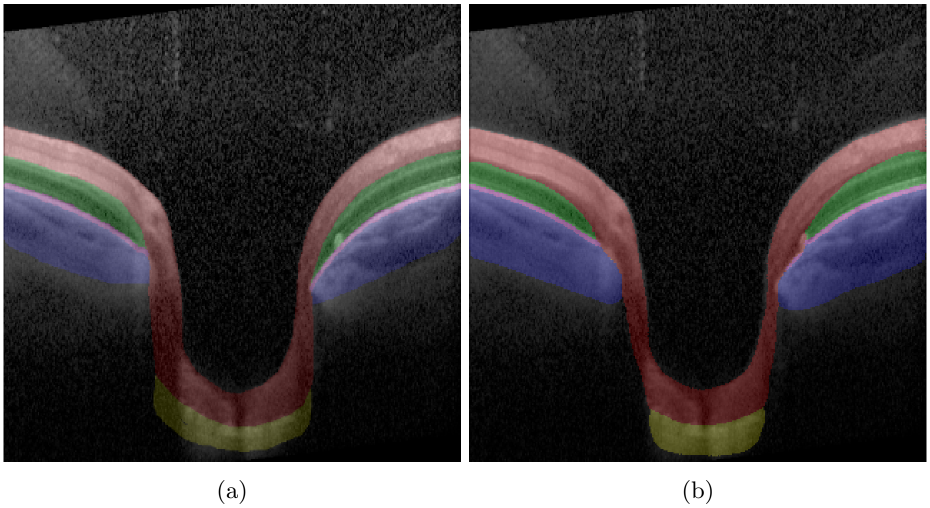
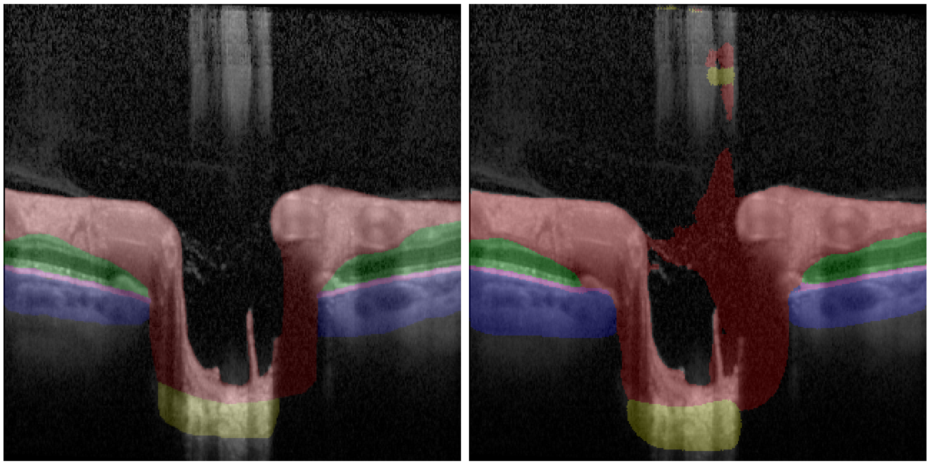


Figure 4.9: Example 1 of incorrect segmentation that affects the LC depth biomarker. (a) Manual segmentation; (b) prediction given by the model.

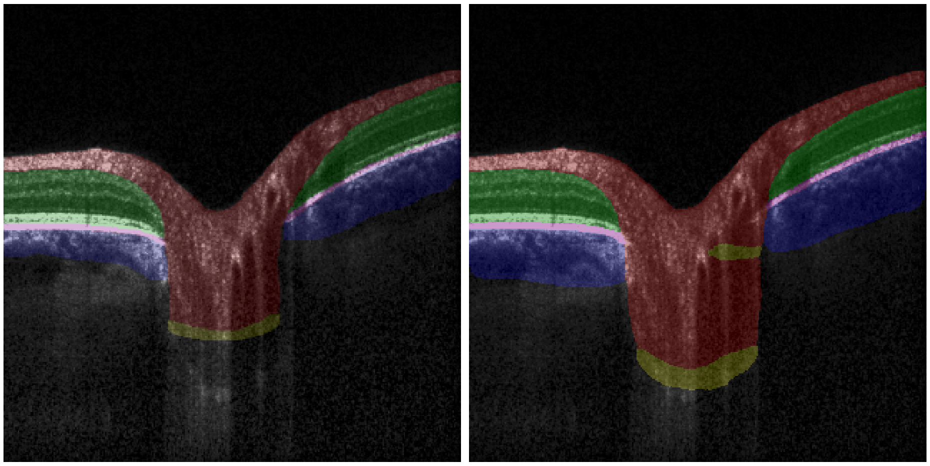
The results obtained after removing these three cases are showed in Figure 4.12. Comparing the Bland-Altman and scatter plots from Figures 4.8 and 4.12, it is possible to see a clear improvement after the exclusion of the three outliers. The Bland Altman shows a decrease on both the mean difference and the standard deviation of the differences between pairs of measurements. The scatter plot shows an increase of the linear regression correlation coefficient from $r^2 = 0.61$ to $r^2 = 0.93$.

The results for the LCCI are shown in Figure 4.13. Based on these, the LCCI biomarker was excluded as clinical feature in the subsequent analyses, as the results show how the current segmentations of the LC are not accurate enough for a correct LCCI measurement. This can be observed in the values for mean difference and standard deviation in the Bland-Altman plot and also in the low correlation value ($r^2 = 0.0066$). This is consistent with the information regarding the LCCI percentiles in Table 4.6, where it is possible to verify differences in the percentiles between the measurements on manual and automatic segmentations.



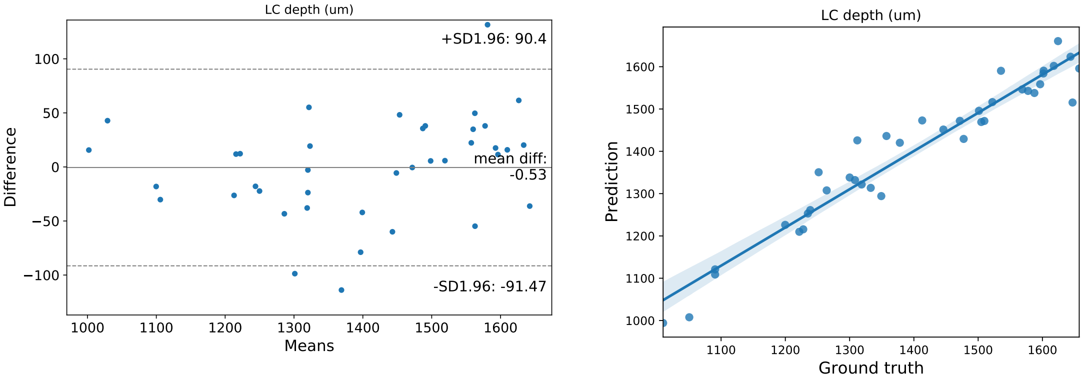
(a) (b)

Figure 4.10: Example 2 of incorrect segmentation that affects the LC depth biomarker. (a) Manual segmentation; (b) prediction given by the model.



(a) (b)

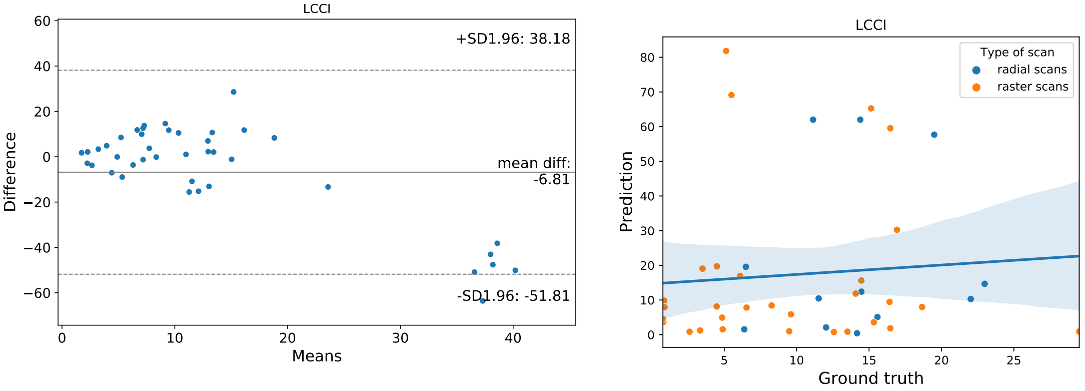
Figure 4.11: Example 3 of incorrect segmentation that affects the LC depth biomarker. (a) Manual segmentation; (b) prediction given by the model.



(a)

(b)

Figure 4.12: Comparison between the LC depth obtained from the ground-truth and the model prediction for the test set after removing the three outliers. (a) Bland-Altman plot; (b) scatter plot with the linear regression line ($r^2 = 0.93$).



(a)

(b)

Figure 4.13: Comparison between the LCCI obtained from the ground-truth and the model prediction on the test set. (a) Bland-Altman plot; (b) scatter plot with the linear regression line ($r^2 = 0.0066$).

4.4.2 Comparison between healthy and glaucoma data

This section presents the results for the biomarkers extracted from the automatic segmentations predicted on all available B-scans. These results were used to evaluate the existence of a significant statistical difference between healthy subject and glaucoma patients based on each biomarker.

The Shapiro Wilk normality test showed that the BMO-MRW measurements in healthy subjects were the only ones following a Gaussian distribution. In consequence, the Mann-Whitney U rank test was used to evaluate if there were significant differences between healthy and glaucoma B-scans for each biomarker. The RNFL thickness was the only parameter that did not show significant differences (p-value$= 0.1$), and was therefore excluded from further analysis. The results for the optic disc diameter, BMO-MRW, RNFL area, and LC depth, are presented in box plots in Figure 4.14.

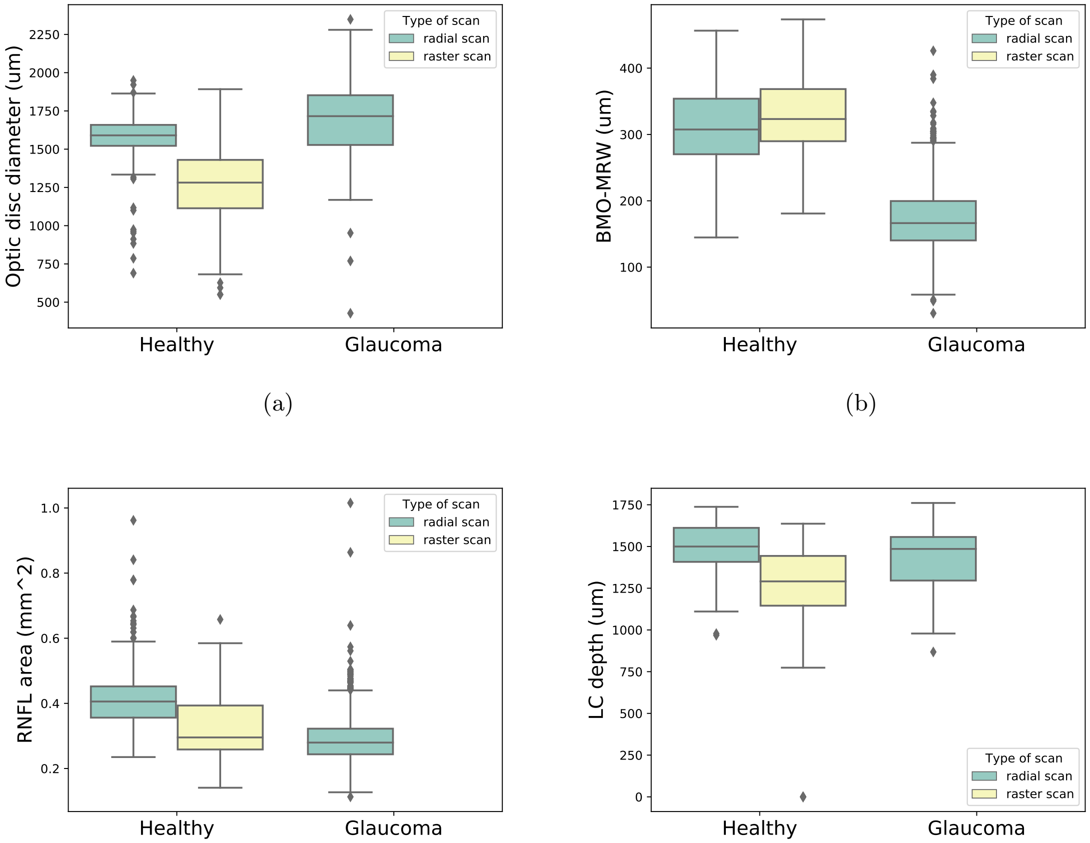
Attending to the differences between radial and raster scans, previously reported in Subsection 4.4.1, the measurements of each dataset were separated in the boxplots. A visual analysis of Figure 4.14 further corroborates these differences on healthy subjects, the only group with raster and radial scans available.

Figure 4.14 (a) shows that the optic disc diameter is larger in glaucoma eyes and, regarding the measurements in healthy eyes, larger in radial scans. The Mann Whitney test corroborates that there is a significant difference between the measurements from the two acquisition protocols (p-value< 0.05).

Figure 4.14 (b) shows that the BMO-MRW is smaller in glaucoma patients. Although visually there does not seem to be a large difference between raster and radial scans, the results from the T-test show a significant difference between them (p-value$= 0.0005$).

Figure 4.14 (c) shows that the RNFL area is smaller in glaucoma patients than in healthy subjects. Moreover, it shows that, between healthy subjects, it is smaller in the raster scans. The Mann Whitney test corroborates that there is a significant difference when comparing the two acquisition protocols (p-value< 0.05).

Finally, Figure 4.14 (d) shows that the LC depth is smaller in raster scans in comparison with radial scans, and slightly smaller in glaucoma patients in comparison with healthy subjects, corroborated by the Mann Whitney test (p-value< 0.05). Despite the difference between healthy and glaucoma not being visually evident, Mann Whitney test showed a significant difference (p-value$= 0.0025$).



(a)

(b)

(c)

(d)

Figure 4.14: Results from the biomarkers measurements on automatic segmentations from healthy and glaucoma data. Box plots show the (a) optic disc diameter, (b) BMO-MRW, (c) RNFL area and (d) LC depth measurements separated in healthy subjects and glaucoma patients. The healthy subjects are divided by dataset (radial and raster scans).

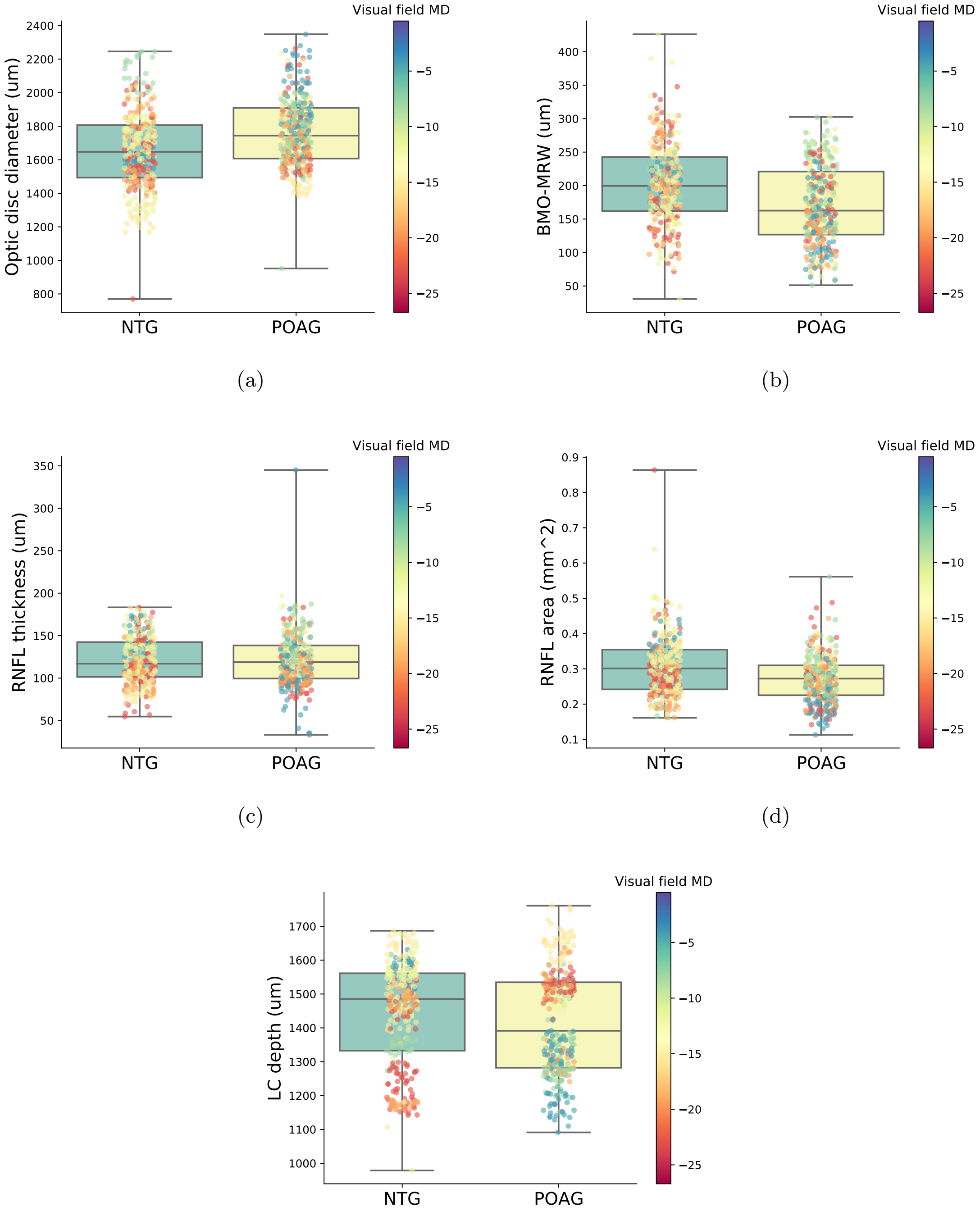
4.4.3 Comparison between NTG and POAG data

Figures 4.15 and 4.16 show the distribution of each computed feature divided according to the type of glaucoma diagnosed, NTG or POAG. Each point corresponds to a B-scan and is colored following the gradient scale according with its VF MD (Figure 4.15) or maximum IOP (Figure 4.16). By looking at the colored points on the box plots it is possible to observe that both VF MD and IOP values are overall lower in NTG. All biomarkers showed statistically significant differences (p-value< 0.05) between groups with the Mann Whitney test, except for the RNFL thickness. By looking at the box plots, it is possible to verify that the optic disc diameter is larger for POAG, and that the BMO-MRW, RNFL thickness, and LC depth are lower for POAG.

Table 4.7 shows the correlation between each biomarker and the VF MD and the IOP, separately. Regarding the VF MD, all biomarkers show a significant correlation, except for the LC depth. The remaining biomarkers decrease as the visual field deteriorates (lower values) except for the optic disc diameter, which increases. The visual field shows a weak correlation with all biomarkers except with the BMO-MRW, to which it has a moderate correlation. Regarding the IOP, all biomarkers show a significant correlation, except for the RNFL thickness. All the remaining biomarkers decrease as IOP increases, except for the optic disc diameter, which also increases. The correlation between the IOP and the biomarkers is always weak, specially for the LC, to which it shows the weakest correlation.

Table 4.7: Spearman correlation coefficient and p-value between each biomarker and the VF MD and the IOP.

	Vidual field MD		Maximum IOP	
	Correlation	p-value	Correlation	p-value
Optic disc diameter	-0.14	1.91×10^{-4}	0.31	1.69×10^{-16}
BMO-MRW	0.53	1.08×10^{-49}	-0.22	1.14×10^{-8}
RNFL thickness	0.19	6.57×10^{-7}	-0.02	0.55
RNFL area	0.36	2.42×10^{-22}	-0.32	6.09×10^{-18}
LC depth	-0.07	0.061	-0.12	0.001



(a)

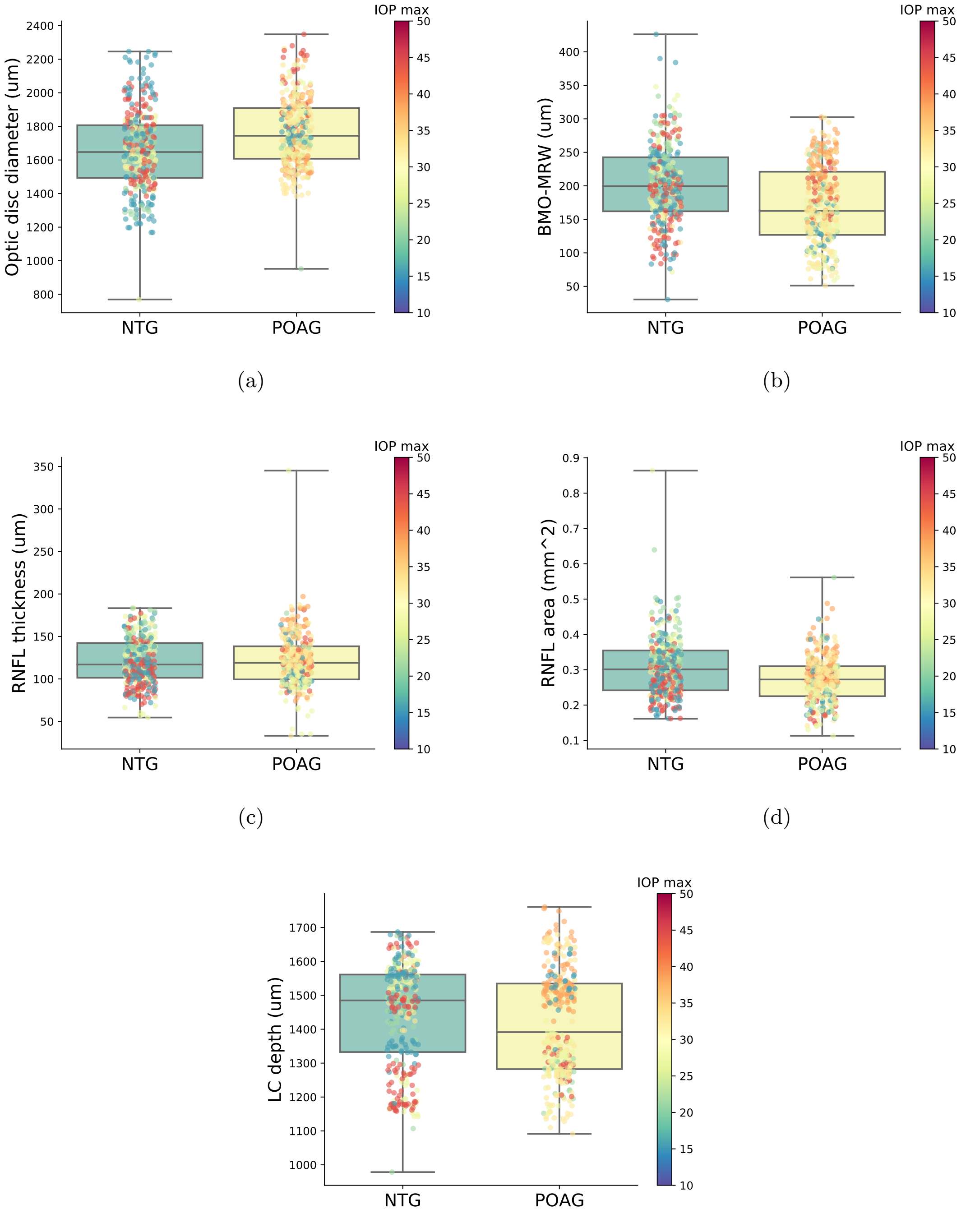
(b)

(c)

(d)

(e)

Figure 4.15: Box plots for the biomarkers measurements on automatic segmentations from NTG and POAG data. The color of each individual point indicates the VF MD for that B-scan.



(a)

(b)

(c)

(d)

(e)

Figure 4.16: Box plots for the biomarkers measurements on automatic segmentations from NTG and POAG data. The color of each individual point indicates the IOP for that B-scan.

Chapter 5

Discussion

In this chapter, the results presented in Chapter 4 are discussed and compared with those found in literature. First, the segmentation model development and obtained results are discussed. Then, the model validation is further analysed taking into consideration the results from the biomarkers extraction. Finally, the biomarkers extracted from automatic segmentations are used to infer differences between healthy and glaucoma subjects and between NTG and POAG patients.

5.1 Segmentation model

The challenges associated with the manual segmentation of the ONH are a common problem in OCT [88]. To tackle this issue, the manual segmentations used as ground truth for training and testing the model were segmented by two different graders. The results of the inter-grader variability assessment showed a good correlation with metrics (Dice coefficient, sensitivity and specificity) above 97% for all tissues except for the choroid and LC, which had metrics above 94% and 86%, respectively. Due to the poor signal-to-noise ratio of the LC and the common artifacts at that region, it was often a difficult challenge, even for specialists, to precise the exact boundaries of the LC.

Regarding the inter grader variability analysis of the biomarkers, the LC related parameters were the only ones showing a correlation coefficient bellow 0.99 (0.87 and 0.34 for the LC depth and LCCI, respectively). These results may be related to the uncertainty observed on the LC segmentation between graders. However, excluding the LCCI, the inter grader manual segmentation variability did not significantly impact the biomarkers computation.

Excluding the LC, the performance of the binary classifier for the remaining labels was overall successful (accuracy > 0.65, including the background), with model predictions that visually agreed with what was expected from their respective manual segmentations. Furthermore, despite not being correctly identified when isolated, results from the binary segmentation of all the labels as a whole, as seen in Figure A.6, show inclusion of the LC area in the segmentation. From this observation, it was possible to learn that contextual information was important in the LC segmentation.

Regarding the multiclass segmentation, the proposed model was able to simultaneously segment the RNFL, the BM/RPE complex, all other retinal layers, the choroid, and the LC. A grid search hyperparameter optimization was used to find the best set of hyperparameters for the segmentation model. However, the ranges of the values used for this optimization were limited due to time and computational constraints. Therefore, it may be possible to obtain even higher accuracy with finer parameter tuning. Furthermore, alternative hyperparameter optimization methods could be explored in the future (e.g. Bayesian [89]). Finally, an in-depth study of each hyperparameter's impact on the model was out of this project's scope. Because of this, conclusions can be provided only regarding the set of parameters that yielded the best results. Still, it was observed that best performances overall were associated with lower learning rates and a larger number of features (filter 2).

Only three studies were reported, all by Devalla *et al.* [10, 57, 58], which present a region based segmentation of ONH centred OCT data using deep learning. The authors divided the data into the following regions: the RNFL, the RPE (which is isolated from the BM), the other retinal layers, the choroid, the sclera, and the LC. The first study [57] used an eight layer CNN, while the two others [10, 58] used more complex architectures based on the U-net. All these works reported results in radial scans acquired using Spectralis [57,58]. Therefore, the results should be comparable with this work and, specifically, with dataset B, which uses the same acquisition pattern (although changes in the protocol will likely to exist).

The Dice coefficient for the proposed model, the eight layer CNN [57] and the DRUNET [58], respectively, were 0.91 ± 0.05, 0.82 ± 0.05 and 0.93 ± 0.04 for the RNFL segmentation; 0.70 ± 0.14, 0.84 ± 0.02 and 0.84 ± 0.04 for the BM/RPE complex and RPE only; 0.88 ± 0.06, 0.86 ± 0.03 and

0.95 ± 0.02 for the other retinal layers; and 0.82 ± 0.08, 0.85 ± 0.02 and 0.91 ± 0.04 for the choroid. Therefore, the proposed model has an almost equal Dice coefficient as the DRUNET for the RNFL, and slightly lower for the remaining tissues. However, it obtained higher Dice coefficients than the eight layer CNN for the RNFL, the other retinal layers, and the choroid.

The sensitivity for the proposed model, the eight layer CNN [57] and the DRUNET [58], respectively, were 0.93 ± 0.04, 0.89 ± 0.04 and 0.92 ± 0.02 for the RNFL segmentation; 0.89 ± 0.10, 0.90 ± 0.03 and 0.88 ± 0.04 for the BM/RPE complex and RPE only; 0.86 ± 0.09, 0.98 ± 0.02 and 0.96 ± 0.03 for the other retinal layers; and 0.92 ± 0.04, 0.91 ± 0.02 and 0.90 ± 0.04 for the choroid. In summary, the proposed model achieves almost equal sensitivities as the models presented in literature for all the tissues, except for the other retinal layers, where it performed slightly lower.

The latest work by Devalla *et al.* [10] focuses on comparing the performance of a model in three different OCT devices. However, values for the Dice coefficient, sensitivity and specificity on individual tissues are reported for Spectralis data, and they are always higher than 0.90, outperforming their previous works and the proposed model in all the quantified tissues (RNFL, RPE, other retinal layers, and choroid). Despite taking the tissue in the segmentation, none of the works in the literature [10,57,58] reported quantitative results for the LC segmentation. Therefore, further comparisons of the testing results for this structure were not possible.

The present study uses data from both raster and radial scans of the same OCT device. Table 4.4 shows that the model performs better when trained on images from dataset A and tested on images from dataset B than the other way around. Several factors were hypothesised to explain these results. The tested parameters might work better for dataset A, which can be due to the fact that dataset A was the first to be accessed and therefore had more time to be fine tuned. Also, the data augmentation used may be mimicking more characteristics of dataset B on dataset A, making it easier for patterns of dataset B to be recognized by a model that only learned from dataset A. Furthermore, dataset A has more variability, since the scans image different parts of the ONH and not only the middle, which might have led to a better generalization of the model. When training on dataset B and testing on A, the model was trying to recognize the anatomy from the borders of the ONH without knowing it. Moreover, dataset B has clearer boundaries overall, which might simplify its segmentation. Finally, dataset A is more than twice as large as dataset B (206 vs 94) - and models trained on more data (which, on top of that, is more varied) are expected to perform better. These results raise the discussion of whether an OCT acquisition protocol should be defined as standard when studying the ONH. That way, there could be more comparability across studies as the same region of the ONH was being imaged.

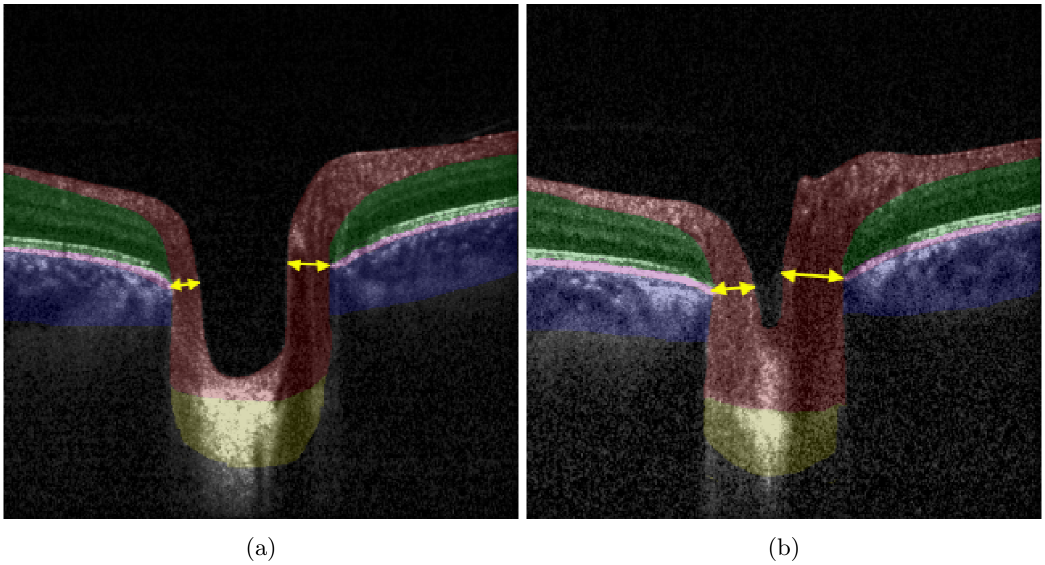
5.2 Optic nerve head biomarkers

5.2.1 Analysis of the automatic segmentation bias

The percentiles values from Table 4.6 are similar between ground truth and automatic segmentations. When comparing the biomarkers measurements with the ones reported in the literature and referred in section 2.4.1, the BMO-MRW, LC depth, and LCCI were the ones showing the largest differences. While the BMO-MRW has been reported between 129 ± 38 μm (IOP $>$ 30mmHg) and 195 ± 99 μm (normal IOP) [30], in this project it was 284 ± 82 μm and 298 ± 81 μm for biomarkers extracted from manual and automatic segmentations, respectively. The results presented from this project include data from both healthy and glaucoma patients, which cover a wide range of IOP. The LC depth has been reported between 404 ± 103 μm (normal IOP) and 655 ± 174 μm (IOP $>$ 30mmHg) [50]. In this project the LC depth was 1408 ± 179 μm and 1392 ± 189 μm for manual and automatic segmentations, respectively. The LCCI has been reported between 35 ± 29 (normal IOP) and 76 ± 28 (IOP $>$ 30mmHg) [50], and in this project it was 11 ± 7 and 17 ± 22 in manual and automatic segmentations, respectively. However, it must be noted that measurements of the same parameter can vary greatly between computation approach, OCT devices used to acquire the data, acquisition protocols and populations targeted, making it difficult for a direct comparison between results from different studies [90, 91].

Regarding the optic disc diameter, Figure 4.3 shows that larger differences between measurements often belong to larger discs. One possible explanation is that some of the images from the dataset might belong to myopes. The optic disc diameter is often overestimated in myopes eyes in OCT [92] which may translate to higher uncertainty in the manual segmentation. Moreover, bigger values of the optic disc measurements belong to radial scans as they all cross the center of the ONH, where the optic disc diameter is expected to be at its maximum. Therefore, when comparing the obtained values with reference values, it is possible to verify that the optic disc measurements are within the expected range of values (1.1-2.0 mm [61]).

The BMO-MRW width can only be correctly measured in scans acquired perpendicularly to the disc border, which is the case for all radial scans. When measuring the biomarker in raster scans, the measured width may be artificially increased. That can be explained by the fact that the raster scans only cross the center of the ONH at one single B-scan while the remaining B-scans are parallel to the central one. As the distance to the center increases, the BMO-MRW is expected to increase as well, since the shape of the ONH is attenuated and merges back to the macular retinal layers. Figure 5.1 (a) and (b) show a segmented B-scan from the middle and border of the ONH, respectively, where it is possible to verify this increase on the distance from the BM termination to the anterior surface of the RNFL. Considering this, the results showing higher BMO-MRW width for raster scans (Figure 4.4) are in agreement with what was expected. Finally, the analysis on both nasal and temporal sides, despite being done only on raster scans, allows further validation of the BMO-MRW and shows that the computation of the biomarker in both sides produces equally robust results in the manually annotated and automatically segmented data. The main motivation behind this separated analysis is that the temporal side, being closer to the macula, is more informative of the disease.



(a)

(b)

Figure 5.1: Raster B-scan from the (a) middle and (b) border of the ONH. The BMO-MRW is illustrated with yellow arrows.

The main difference between the two RNFL-based biomarkers, area and thickness, is that, while the thickness corresponds to the mean value between two single distances at each side of the ONH, the area covers a more extensive region of the RNFL. Therefore, the latter is more robust to errors in the automatic segmentation of the RNFL. This may explain the higher correlation between RNFL area measured in the manual and predicted segmentations when compared to the correlation of the RNFL thickness measurements under the same circumstances. Possible solutions that might help the RNFL thickness robustness is an average of the measurement on all B-scans of each volume and/or previously correcting all B-scan for the same rotation angle.

The LC depth showed a high correlation between measurements on manual segmentations and automatic predictions of the same B-scan after the removal of the three outliers (the correlation coefficient improved from 0.61 to 0.93). These outliers were excluded due to poor image quality, unanimously decided by the clinical team after review. Image quality factors influencing this decision include imaging artifacts and poor eye alignment during acquisition, further discussed next. Artifacts are a known problem within the OCT imaging, and it is a common practice to exclude images bellow a reasonable minimum quality from the analysis since the clinician cannot clearly see the structures and, therefore, cannot provide a ground truth.

A low image quality affects each structure differently in the retina but, due to how OCT light beam penetrates the tissues during acquisition, the least visible structures are usually the deeper ones, due to a higher likelihood of scattering and attenuation. This includes the LC, and is linked with the previously discussed difficulty in the LC manual segmentation (Figure 4.1 (e)), which will consequently affect the determination of the LC depth. Even if a clinician validated all the manual segmentations, the prediction on Figure 4.9 (b) was considered more accurate by the experts than the initial manual segmentation. This is because the manual segmentation may also include speculation based on anatomical knowledge of the region. The prediction, on the other hand, is a data-driven approach, and thus only segments the LC where the image shows the changes in intensity and structure that the model has been trained to detect.

Images with the type of artifacts shown in Figure 4.10, which shows a hyperreflective region in the vitreous, were rare in both datasets, explaining the trained model difficulty to identify that reflection as part of the background. However, that B-scan also showed additional artifacts due to the vitreous humor detachment from the retina that may help explain the mistakes on the RNFL layer automatic segmentation (Figure 4.10, in red).

Finally, Figure 4.11 is another example of how the trained model can interpret the images in a different way than the human specialists. Apart from the LC label wrongly placed at the upper region of the prelaminar tissue, Figure 4.11 (b) correctly places the LC at a lower position than the manual segmentation. Despite having a good image quality, the B-scan from Figure 4.11 was acquired far from the center of the ONH, almost at its border, where the LC is at its lowest visibility due to an higher probability of the light being scattered/attenuated in the tissues above, contributing for the difficulty of its segmentation.

One possible solution for the reported problems could be to train the model in 3D volumes instead of 2D B-scans, which would leverage sequential information [93] from the OCT volume instead of considering only a single B-scan as input. In addition, ensuring layer ordering [93], that is, including knowledge of the LC anatomy and ensure, for example, that the LC cannot be above the RNFL, could also be part of the solution. These would already partly solve the issue on Figures 4.10 and 4.11.

Despite the difficulties associated with the segmentation of the LC, the segmentation results only affected the outcome of the LC depth in 3 out of the 42 B-scans used in the test set. Therefore, even though state of the art results were not available for comparison with the LC segmentation, the achievement of a high correlation between the LC depth measurements on manual and automatic segmentations shows that the segmentation model prediction is accurate enough for computing relevant LC clinical parameters.

Regarding the LC curvature, it is known to be highly sensitive to local variations [94]. These local variations were common in the analysed segmentations, both manual and automatic, due to the inherent noise of OCT data, artifacts, low image quality, and uncertainty associated with the manual segmentation. The effect of these local variations is translated to the LCCI results, where differences between the pairs of measurements are too large (as can be observed in the Bland-Altman plot, Figure 4.13 (a)) and the correlation too weak ($r^2 = 0.0066$) to be acceptable for a clinical context. Additionally, the discrepancy between the obtained values and the theoretically expected [30], stand by the fact that the LC anterior surface curvature delineation is not accurate enough. One possible solution to increase the viability of the LCCI could be to smooth the segmentation of the LC before the biomarker determination. This could be achieved using a low order polynomial fitting on the anterior surface of the LC. Following this idea, Rahman *et al.* [95] recently presented a deep learning boundary based segmentation model for the computation of LC related parameters. They first implemented a semantic segmentation model. Then, they applied a polynomial regression curve that estimates the curve of the anterior LC boundary, based on the acquired BMO and LC information from the model. They achieved a 0.96 and 0.7 correlation coefficient between manual and predicted segmentations based LC depth and LCCI measurements, respectively. Comparing with the correlations presented in this project for the same parameters, which are similar for the LC depth and much lower for the LCCI, the LC depth seems to be a more robust measurement, less sensitive to local variations, since it can be accurately measured even with a more imprecise delineation of the anterior surface.

5.2.2 Comparison between healthy and glaucoma subjects

The discrepancy between the results computed from raster and radial scans for the biomarkers measurements on automatic segmentations from healthy and glaucoma is evident across Figure 4.14. That may be explained by the fact that biomarkers are being measured at different locations of the ONH. Moreover, since the comparisons are not being made between the same patients, the inter-subject anatomical variability may also have an effect on this discrepancy.

The optic disc diameter is normally larger at the center of the ONH. The lowest values for the optic disc diameter found in the healthy raster scans may be due to the fact that some of those B-scans were acquired at ONH border, where the diameter is expected to be small. It is also expected for the optic disc diameter to be larger in glaucoma eyes [61], a trend that can be verified in Figure 4.14 (a).

The BMO-MRW distance is expected to decrease as we approach the ONH center. This way, the slightly higher values for the BMO-MRW found in healthy raster scans when compared with the radial scans may be explained by the fact that some of those B-scans were taken further from the center, where this distance is expected to be larger. However, BMO-MRW results show the smallest difference between radial and raster scans in the healthy group. This may be due to the fact that the tissues involved in the computation of this biomarker are two of the easiest to segment, with very distinct intensity patterns, therefore making a very stable biomarker. The BMO-MRW is also expected to be smaller in glaucoma patients [96] since this disease is characterized by a loss of nerve fibers in the RNFL. This can be verified in Figure 4.14 (b). Additionally, the BMO-MRW biomarker has shown to be the best to differentiate between healthy and glaucoma based on the box plots.

As previously mentioned in Chapter 1, glaucoma is characterized by a RNFL thinning [96]. This was expected to cause a decrease in the RNFL area, which can be verified in Figure 4.14 (c), being more evident between radial scans of both groups. However, despite the fact that RNFL thickness is the current RNFL evaluation biomarker used in the clinical management of glaucoma, it was not further analysed since it did not show significant differences between groups. This might be related to the robustness of the RNFL thickness as computed in this project, discussed in section 5.2.1.

The LC depth is expected to be significantly greater in patients with glaucoma than normal subjects [94]. Despite not being visually clear in the boxplot from Figure 4.14 (d), results confirmed there were statistically signifant differences between the two groups. The differences between acquisition protocols on healthy subjects on Figure 4.14 (d) may be explained by the fact that the LC is at its deepest at the center of the ONH, where all radial scans cross.

5.2.3 Comparison between NTG and POAG

As it is detailed in Chapter 1, glaucoma is not a single disease but a family of diseases. Most of the types of glaucoma, such as POAG, are characterized by a rise in IOP, but this is not the case of NTG, a form of glaucoma where IOP is expected to be lower than 21mmHg. However, the clinical similarities between NTG and POAG, and the controversy around the role of IOP on the pathogenesis of NTG, have prompted doubts on the diagnostic and treatment of both NTG and POAG [97, 98]. As a consequence, the role played by several ONH factors on these pathologies is still being studied, and the results discussed here will be compared with what has already been reported in the literature, within the possible limits given the data available.

While drawing conclusions from these results, it is important to consider that POAG and NTG groups were significantly different for the IOP and visual field. This may bias further analysis and comparisons of these two groups in a conventional way. The IOP and VF MD are confounding factors that may be causing the differences verified on the biomarkers between the two glaucoma groups and must be accounted for as much as possible. Such is verified by the significant correlation between IOP, VF MD and most of the biomarkers presented on Table 4.7, which verifies their effect on the measurements. By showing a significant correlation between visual field, or IOP, and a ONH biomarker, it implies that the changes on the biomarker measurments may be related to those clinical factors.

The optic disc diameter has been reported to be either similar or larger in patients with NTG compared to POAG [61]. Such results could not be verified in this work (Figure 4.16 (a) shows larger optic discs in POAG with higher IOP values), which point that optic disc diameter increases with higher IOP values, characteristic of POAG.

The BMO-MRW has been reported to decrease with increasing IOP, which is characteristic of POAG [30]. Such can be verified in the box plots in Figure 4.16 (b) and the Spearman correlation coefficients for the BMO-MRW biomarker, which imply a decrease on this distance on POAG patients.

Parameters extracted from the RNFL are reported to be higher for NTG than POAG by the literature [99]. Such observation was not verified in this project for the RNFL thickness, which also does not show a correlation with IOP. However, a significantly statistical difference was observed for the RNFL area, corroborating that this measurement of the RNFL is more robust, and therefore more fit to evaluate the alterations in this structure.

The results in this work also show that the LC depth decreases with IOP increase, and it is lower for POAG patients. However, it has been reported in the literature that the increase of the LC depth slows down when the IOP reaches 25 mmHg [30]. This may explain the spread of some points colored with higher IOP at lower LC depths on Figure 4.16 (e). One possible explanation is due to the elastic nature of the LC, which stops the progression of the membrane at a certain level of pressure [100].

The definition of POAG places it at IOPs higher than 21 mmHg, whether NTG is used to describe the remaining cases. However, high IOP values do not always correspond to POAG. For example, in Figure 4.16 it is possible to spot some red dots, which translate to pressures above 40mmHg, on the NTG side of the boxplots. This brings to light the separation between NTG and POAG, which has been put into question by several studies [97] and may play some influence on the drawn conclusions regarding the IOP and visual field influence on the differences between biomarkers.

In summary, the results from this analysis are comparable with what has been reported. However, further research into the NTG vs. POAG separation is still needed, and more clinical factors should be included in this analysis. One possible line of research could be including these confounding factors into a multiple linear regression. Linear mixed models are also something to consider, but they would benefit from a bigger patient sample.

Chapter 6

Conclusion

6.1 General conclusions

The main objectives stated on Chapter 1 were overall successfully achieved. A deep learning approach that is able to capture both local and contextual features to simultaneously segment connective and neural tissues from OCT B-scans of the ONH was presented and evaluated. Furthermore, it was possible to successfully extract five clinically relevant parameters that not only showed high correlation with the ground truth, but also a good agreement with what was expected from the literature based on the current knowledge of the structural and functional characteristics of glaucoma.

The work developed during the project tackled some difficult tasks, such as the differences between acquisition protocols, despite both used datasets being acquired with the same device. The definition of the radial acquisition protocol as standard for evaluations of the ONH could benefit future research and help studies to be more comparable between them. Additionally, the manual segmentation of the OCT data to generate the ground truth for training the model was time consuming and difficult, even for experienced clinicians. The results show some discrepancy between graders, specially on the LC, which may have implications when analysing the results.

The choice of the U-net was an important one given its proven success in the biomedical imaging field [101]. Considering the scarcity of studies that developed a region based segmentation of the ONH in OCT data and the fact that all use more complex deep learning architectures based on

U-nets, this project allowed more insight on the segmentation potential and limitations of a simple U-net architecture [102]. The developed segmentation model was able to successfully segment all the defined ONH structures. When compared with the literature, the results are either similar or a bit lower regarding of Dice coefficient, sensitivity and specificity. Moreover, to the best of our knowledge, it was the first work to present a quantitative assessment of an automatic region-based segmentation of the LC.

In conclusion, a simple U-net architecture, despite achieving slightly lower segmentation than more complex ones, provides a segmentation robust enough to compute accurate and reliable biomarkers, which is a pressing need on the daily clinical management of glaucoma.

6.2 Future work

The segmentation model developed in this work could be further improved. Some options to achieve this improvement could be to use transfer learning [103], which applies information learned in a previous problem to help solving the target problem; or to incorporate more information from the morphology of the LC into the network to address the segmentation flaws on this structure, which are highly undesirable given the significance of LC morphology in glaucoma. Further data augmentation and post-processing of the dataset [9] could also be explored for this purpose.

The results presented bring into discussion the data-centric vs. model-centric debate that has recently been gaining ground [104]. While a model-centric approach focuses on fine-tuning the architecture to improve the performance, a data-centric approach aims to achieve the same goal by focusing on systematically changing and/or improving the dataset. The industry, which closely follows academic research, has been mostly focusing on model-centric approaches, since state-of-the-art and cutting-edge advances are readily available to anyone [105,106]. Thanks to this availability, hyperoptimizing models is becoming less and less necessary. However, datasets that have high quality, are consistently defined, cover important cases, and are appropriately sized, are not so readily available. The observations made in this project regarding the manual labelling difficulties and the impact that the image quality had on the prediction mistakes, such as the ones on Figures 4.10 and 4.11, exemplify the benefits that a data-centric approach could have. It could be the bridge to higher segmentation metrics as achieved by more complex architectures [10] while assuring the reliability of the data. Therefore, future work should include a focus on enhancing the data, and on better data augmentation. Methods for a more systematic and consistent labelling of the complex ONH anatomy [107], such as data versioning, smart labelling, and tracking, should also be explored. Additionally, relaxing the dependency on the labelled data through transfer learning [108], multi tasking learning [109], and semi-supervised learning [110] could also be a possible solution. To reduce the dependency on human skill will diminish the impact that inconsistencies on the ground truth cause on the performance, results, and development of the model.

The emergence and democratization of OCT as the clinical gold-standard for in-vivo structural ophthalmic examinations [111] has encouraged the entry of new manufacturers to the market as well. Given that preparing reliable manual segmentations is time-consuming and requires highly skilled experts, it will soon become practically infeasible to perform manual segmentations for all OCT brands, device models, generations, and applications [10]. Therefore, despite the fact that only one device independent segmentation algorithm has been reported for the ONH segmentation [10], such an approach has already been further explored and found to be important on other regions of the eye (e.g. the macula) [112, 113] and should be considered for future work.

The knowledge acquired from the biomarkers differences between healthy and glaucoma, and NTG and POAG, could be integrated in machine learning models that would thereafter classify each patient in either of these groups and could aid diagnostics. Moreover, the patient-specificity of some biomarkers, such as the RNFL area and the LC [34], may cause bias when comparing subjects. However, these biomarkers could also be explored for the clinical follow-up of patients with glaucoma, where the patient-specificity is not an issue. Particularly, the NTG and POAG analysis still lacks information and needs to be further explored, while accounting for other confounding factors. Since the NTG and POAG relationship with IOP is not deterministic and still needs further research, these biomarkers could also be used on follow-ups to get more insight on these diseases characteristics and development.

Finally, the integration of the proposed fully-automatic pipeline into an interface that could be easily used by clinicians in daily practice would enable faster and more reliable analysis of OCT data, providing objective parameters for diagnostics and follow-up, and reaching a wider portion of the population. This interface would show the output of the segmented regions and the biomarkers values. By doing this, it would be possible for the clinician to identify outlier biomarkers measurements and double check manually to guarantee the safety of the process.

Bibliography

[1] Jaimie D Steinmetz, Rupert RA Bourne, Paul Svitil Briant, Seth R Flaxman, Hugh RB Taylor, Jost B Jonas, Amir Aberhe Abdoli, Woldu Aberhe Abrha, Ahmed Abualhasan, Eman Girum Abu-Gharbieh, et al. Causes of blindness and vision impairment in 2020 and trends over 30 years, and prevalence of avoidable blindness in relation to vision 2020: the right to sight: an analysis for the global burden of disease study. *The Lancet Global Health*, 9(2):e144–e160, 2021.

[2] Harry A Quigley. Neuronal death in glaucoma. *Progress in retinal and eye research*, 18(1):39–57, 1999.

[3] Paul GD Spry and Chris A Johnson. Identification of progressive glaucomatous visual field loss. *Survey of ophthalmology*, 47(2):158–173, 2002.

[4] Alfred Sommer, James M Tielsch, Joanne Katz, Harry A Quigley, John D Gottsch, Jonathan Javitt, and Kuldev Singh. Relationship between intraocular pressure and primary open angle glaucoma among white and black americans: the baltimore eye survey. *Archives of ophthalmology*, 109(8):1090–1095, 1991.

[5] Nicholas Y.Q. Tan, Victor Koh, Michaël J.A. Girard, and Ching Yu Cheng. Imaging of the lamina cribrosa and its role in glaucoma: a review. *Clinical and Experimental Ophthalmology*, 46(2):177–188, 2018.

[6] Guihua Xu, Robert N. Weinreb, and Christopher K.S. Leung. Optic nerve head deformation in glaucoma: The temporal relationship between optic nerve head surface depression and retinal nerve fiber layer thinning. *Ophthalmology*, 121(12):2362–2370, 2014.

[7] Hae Young Lopilly Park, So Hee Jeon, and Chan Kee Park. Enhanced depth imaging detects lamina cribrosa thickness differences in normal tension glaucoma and primary open-angle glaucoma. *Ophthalmology*, 119(1):10–20, 2012.

[8] Hana L. Takusagawa, Ambika Hoguet, Anna K. Junk, Kouros Nouri-Mahdavi, Sunita Radhakrishnan, and Teresa C. Chen. Swept-Source OCT for Evaluating the Lamina Cribrosa: A Report by the American Academy of Ophthalmology. *Ophthalmology*, 126(9):1315–1323, 2019.

[9] Jean Martial Mari, Nicholas G Strouthidis, Sung Chul Park, and Michaël JA Girard. Enhancement of lamina cribrosa visibility in optical coherence tomography images using adaptive compensation. *Investigative ophthalmology & visual science*, 54(3):2238–2247, 2013.

[10] Sripad Krishna Devalla, Tan Hung Pham, Satish Kumar Panda, Liang Zhang, Giridhar Subramanian, Anirudh Swaminathan, Chin Zhi Yun, Mohan Rajan, Sujatha Mohan, Ramaswami Krishnadas, Vijayalakshmi Senthil, John Mark S. de Leon, Tin A. Tun, Ching Yu Cheng, Leopold Schmetterer, Shamira Perera, Tin Aung, Alexandre H. Thiéry, and Michaël J.A. Girard. Towards label-free 3d segmentation of optical coherence tomography images of the optic nerve head using deep learning. *arXiv*, 11(11):6356–6378, 2020.

[11] Rita Marques, Danilo Andrade De Jesus, João Barbosa Breda, Jan Van Eijgen, Ingeborg Stalmans, Theo van Walsum, Stefan Klein, Pedro G Vaz, and Luisa Sánchez Brea. Automatic segmentation of the optic nerve head region in optical coherence tomography: A methodological review. *arXiv preprint arXiv:2109.02322*, 2021.

[12] Colin E. Willoughby, Diego Ponzin, Stefano Ferrari, Aires Lobo, Klara Landau, and Yadollah Omidi. Anatomy and physiology of the human eye: Effects of mucopolysaccharidoses disease on structure and function - a review. *Clinical and Experimental Ophthalmology*, 38(SUPPL. 1):2–11, 2010.

[13] Barry D. Kels, Andrzej Grzybowski, and Jane M. Grant-Kels. Human ocular anatomy. *Clinics in Dermatology*, 33(2):140–146, 2015.

[14] Jost B Jonas, Eduard Berenshtein, and Leonard Holbach. Anatomic relationship between lamina cribrosa, intraocular space, and cerebrospinal fluid space. *Investigative ophthalmology & visual science*, 44(12):5189–5195, 2003.

[15] Anat London, Inbal Benhar, and Michal Schwartz. The retina as a window to the brain. *Nature Reviews Neurology*, 9(1):44–53, 2012.

[16] Sophie Lemmens, Toon Van Craenendonck, Jan Van Eijgen, Lies De Groef, Rose Bruffaerts, Danilo Andrade de Jesus, Wouter Charle, Murali Jayapala, Gordana Sunaric-Mégevand, Arnout Standaert, et al. Combination of snapshot hyperspectral retinal imaging and optical coherence tomography to identify alzheimer's disease patients. *Alzheimer's research & therapy*, 12(1):1–13, 2020.

[17] Heidelberg Engineering. Know your layers. `https://business-lounge.heidelbergengineering.com/be/en/news/news/know-your-retinal-layers-33401465`. Accessed: 09.07.2021.

[18] Akram Belghith, Christopher Bowd, Robert N Weinreb, and Linda M Zangwill. A hierarchical framework for estimating neuroretinal rim area using 3d spectral domain optical coherence tomography (sd-oct) optic nerve head (onh) images of healthy and glaucoma eyes. In *2014 36th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, pages 3869–3872. IEEE, 2014.

[19] Shoji Kishi. Impact of swept source optical coherence tomography on ophthalmology. *Taiwan journal of ophthalmology*, 6(2):58–68, 2016.

[20] A Yasin Alibhai, Chris Or, and Andre J Witkin. Swept source optical coherence tomography: a review. *Current Ophthalmology Reports*, 6(1):7–16, 2018.

[21] ZOFIA Michalewska, JANUSZ Michalewski, and JERZY Nawrocki. Swept-source oct. *Retina Today*, pages 50–56, 2013.

[22] Sunil Kumar Yadav and Ella Maria Kadas. Optic nerve head three-dimensional shape analysis. *Journal of Biomedical Optics*, 23(10):1, 2018.

[23] Harry Quigley and A. T. Broman. The number of people with glaucoma worldwide in 2010 and 2020. *British Journal of Ophthalmology*, 90(3):262–267, 2006.

[24] Danilo A Jesus, João Barbosa Breda, Karel Van Keer, Amândio Rocha Sousa, Luis Abegão, and Pinto Ingeborg. Quantitative automated circumpapillary microvascular density measurements : a new angioOCT-based methodology. *Eye*, pages 320–326, 2019.

[25] Sergio Claudio Saccà, Maurizio Rolando, Antonio Marletta, Angelo Macrì, Piera Cerqueti, and Giuseppe Ciurlo. Fluctuations of intraocular pressure during the day in open-angle glaucoma, normal-tension glaucoma and normal subjects. *Ophthalmologica*, 212(2):115–119, 1998.

[26] Amerens Bekkers, Noor Borren, Vera Ederveen, Ella Fokkinga, Danilo Andrade De Jesus, Luisa Sánchez Brea, Stefan Klein, Theo van Walsum, João Barbosa-Breda, and Ingeborg Stalmans. Microvascular damage assessed by optical coherence tomography angiography for glaucoma diagnosis: a systematic review of the most discriminative regions. *Acta ophthalmologica*, 98(6):537–558, 2020.

[27] Seung Hyen Lee, Tae Woo Kim, Eun Ji Lee, Michaël J.A. Girard, and Jean Martial Mari. Diagnostic power of lamina cribrosa depth and curvature in glaucoma. *Investigative Ophthalmology and Visual Science*, 58(2):755–762, 2017.

[28] Danilo Andrade De Jesus, Luisa Sánchez Brea, João Barbosa Breda, Ella Fokkinga, Vera Ederveen, Noor Borren, Amerens Bekkers, Michael Pircher, Ingeborg Stalmans, Stefan Klein, et al. Octa multilayer and multisector peripapillary microvascular modeling for diagnosing and staging of glaucoma. *Translational Vision Science & Technology*, 9(2):58–58, 2020.

[29] Danilo A Jesus, Joao Barbosa Breda, Karel Van Keer, Amândio Rocha Sousa, Luis Abegao Pinto, and Ingeborg Stalmans. Quantitative automated circumpapillary microvascular density measurements: a new angiooct-based methodology. *Eye*, 33(2):320–326, 2019.

[30] Jian Wu, Yifan Du, Jiaying Li, Xiaowei Fan, Caixia Lin, and Ningli Wang. The influence of different intraocular pressure on lamina cribrosa parameters in glaucoma and the relation clinical implication. *Scientific Reports*, pages 1–13, 2021.

[31] Luís Abegão Pinto, Koen Willekens, Karel Van Keer, Abraham Shibesh, Geert Molenberghs, Evelien Vandewalle, and Ingeborg Stalmans. Ocular blood flow in glaucoma–the leuven eye study. *Acta ophthalmologica*, 94(6):592–598, 2016.

[32] J Crawford Downs, Michael D Roberts, and Claude F Burgoyne. The mechanical environment of the optic nerve head in glaucoma. *Optometry and vision science: official publication of the American Academy of Optometry*, 85(6):425, 2008.

[33] Dae Seung Lee, Eun Ji Lee, Tae-Woo Kim, Young Ho Park, Jungeun Kim, Joon Woo Lee, and SangYun Kim. Influence of translaminar pressure dynamics on the position of the anterior lamina cribrosa surface. *Investigative ophthalmology & visual science*, 56(5):2833–2841, 2015.

[34] Alice Paulo, Pedro G Vaz, Danilo Andrade De Jesus, Luisa Sánchez Brea, Jan Van Eijgen, João Cardoso, Theo van Walsum, Stefan Klein, Ingeborg Stalmans, and João Barbosa Breda. Optical coherence tomography imaging of the lamina cribrosa: Structural biomarkers in nonglaucomatous diseases. *Journal of Ophthalmology*, 2021, 2021.

[35] Anthony J Bellezza, Christopher J Rintalan, Hilary W Thompson, J Crawford Downs, Richard T Hart, and Claude F Burgoyne. Deformation of the lamina cribrosa and anterior scleral canal wall in early experimental glaucoma. *Investigative ophthalmology & visual science*, 44(2):623–637, 2003.

[36] Delia Cabrera DeBuc, Magdalena Gaca-Wysocka, Andrzej Grzybowski, and Piotr Kanclerz. Identification of Retinal Biomarkers in Alzheimer's Disease Using Optical Coherence Tomography: Recent Insights, Challenges, and Opportunities. *Journal of Clinical Medicine*, 8(7):996, 2019.

[37] Muhsin Eraslan, Eren Cerman, Sevcan Yildiz Balci, Hande Celiker, Ozlem Sahin, Ahmet Temel, Devran Suer, and Nese Tuncer Elmaci. The choroid and lamina cribrosa is affected in patients with Parkinson's disease: Enhanced depth imaging optical coherence tomography study. *Acta Ophthalmologica*, 94(1):e68–e75, 2016.

[38] Akram Belghith, Christopher Bowd, Felipe A Medeiros, Robert N Weinreb, and Linda M Zangwill. Automated segmentation of anterior lamina cribrosa surface: How the lamina cribrosa responds to intraocular pressure change in glaucoma eyes? In *2015 IEEE 12th International Symposium on Biomedical Imaging (ISBI)*, pages 222–225. IEEE, 2015.

[39] Mary Comer, Charles A. Bouman, Marc De Graef, and Jeff P. Simmons. Bayesian methods for image segmentation. *Jom*, 63(7):55–57, 2011.

[40] Antoni Buades, Bartomeu Coll, and Jean-Michel Morel. A review of image denoising algorithms, with a new one. *Multiscale modeling & simulation*, 4(2):490–530, 2005.

[41] Jason Chang and John W. Fisher. Efficient MCMC sampling with implicit shape representations. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 2081–2088, 2011.

[42] Zaixing Mao, Atsuya Miki, Song Mei, Ying Dong, Kazuichi Maruyama, Ryo Kawasaki, Shinichi Usui, Kenji Matsushita, Kohji Nishida, and Kinpui Chan. Deep learning based noise reduction method for automatic 3D segmentation of the anterior of lamina cribrosa in optical coherence tomography volumetric scans. *Biomedical Optics Express*, 10(11):5832, 2019.

[43] Michaël J.A. Girard, Nicholas G. Strouthidis, C. Ross Ethier, and Jean Martial Mari. Shadow removal and contrast enhancement in optical coherence tomography images of the human optic nerve head. *Investigative Ophthalmology and Visual Science*, 52(10):7738–7748, 2011.

[44] Kim L. Boyer, Artemas Herzog, and Cynthia Roberts. Automatic recovery of the optic nervehead geometry in optical coherence tomography. *IEEE Transactions on Medical Imaging*, 25(5):553–570, 2006.

[45] Zhihong Hu, Michael D. Abràmoff, Young H. Kwon, Kyungmoo Lee, and Mona K. Garvin. Automated segmentation of neural canal opening and optic cup in 3D spectral optical coherence tomography volumes of the optic nerve head. *Investigative Ophthalmology and Visual Science*, 51(11):5708–5717, 2010.

[46] Bhavna J Antony, Mohammed S Miri, Michael D Abràmoff, Young H Kwon, and Mona K Garvin. Automated 3d segmentation of multiple surfaces with a shared hole: segmentation of the neural canal opening in sd-oct volumes. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 739–746. Springer, 2014.

[47] Mona Kathryn Garvin, Michael David Abràmoff, Xiaodong Wu, Senior Member, Stephen R Russell, Trudy L Burns, and Milan Sonka. Automated 3-D Intraretinal Layer Segmentation of Macular Spectral-Domain Optical Coherence Tomography Images. *IEEE Trans Med Imaging*, 28(9):1436–1447, 2009.

[48] Mohammad Saleh Miri, Michael D Abràmoff, Young H Kwon, Milan Sonka, and Mona K Garvin. A machine-learning graph-based approach for 3d segmentation of bruch's membrane opening from glaucomatous sd-oct volumes. *Medical image analysis*, 39:206–217, 2017.

[49] Kai Yu, Fei Shi, Enting Gao, Weifang Zhu, Haoyu Chen, and Xinjian Chen. Shared-hole graph search with adaptive constraints for 3D optic nerve head optical coherence tomography image segmentation. *Biomedical Optics Express*, 9(3):962, 2018.

[50] Menglin Wu, Theodore Leng, Luis de Sisternes, Daniel L. Rubin, and Qiang Chen. Automated segmentation of optic disc in SD-OCT images and cup-to-disc ratios quantification by patch searching-based neural canal opening detection. *Optics Express*, 23(24):31216, 2015.

[51] Hyunseok Seo, Masoud Badiei Khuzani, Varun Vasudevan, Charles Huang, Hongyi Ren, Ruoxiu Xiao, Xiao Jia, and Lei Xing. Machine learning techniques for biomedical image segmentation: An overview of technical aspects and introduction to state-of-art applications. *Medical physics*, 47(5):e148–e167, 2020.

[52] Zailiang Chen, Peng Peng, Hailan Shen, Hao Wei, Pingbo Ouyang, and Xuanchu Duan. Region-segmentation strategy for Bruch's membrane opening detection in spectral domain optical coherence tomography images. *Biomedical Optics Express*, 10(2):526, 2019.

[53] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015.

[54] Morgan Heisler, Mahadev Bhalla, Julian Lo, Zaid Mammo, Sieun Lee, Myeong Jin Ju, Mirza Faisal Beg, and Marinko V. Sarunic. Semi-supervised deep learning based 3D analysis of the peripapillary region. *Biomedical Optics Express*, 11(7):3843, 2020.

[55] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1125–1134, 2017.

[56] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *arXiv preprint arXiv:1506.01497*, 2015.

[57] Sripad Krishna Devalla, Khai Sing Chin, Jean-Martial Mari, Tin A Tun, Nicholas G Strouthidis, Tin Aung, Alexandre H Thiéry, and Michaël JA Girard. A deep learning approach to digitally stain optical coherence tomography images of the optic nerve head. *Investigative ophthalmology & visual science*, 59(1):63–74, 2018.

[58] Sripad Krishna Devalla, Prajwal K Renukanand, Bharathwaj K Sreedhar, Giridhar Subramanian, Liang Zhang, Shamira Perera, Jean-Martial Mari, Khai Sing Chin, Tin A Tun, Nicholas G Strouthidis, et al. Drunet: a dilated-residual u-net deep learning network to segment optic nerve head tissues in optical coherence tomography images. *Biomedical optics express*, 9(7):3244–3265, 2018.

[59] Sripad Krishna Devalla, Giridhar Subramanian, Tan Hung Pham, Xiaofei Wang, Shamira Perera, Tin A Tun, Tin Aung, Leopold Schmetterer, Alexandre H Thiéry, and Michaël JA Girard. A deep learning approach to denoise optical coherence tomography images of the optic nerve head. *Scientific reports*, 9(1):1–13, 2019.

[60] Biomarkers Definitions Working Group, Arthur J Atkinson Jr, Wayne A Colburn, Victor G DeGruttola, David L DeMets, Gregory J Downing, Daniel F Hoth, John A Oates, Carl C Peck, Robert T Schooley, et al. Biomarkers and surrogate endpoints: preferred definitions and conceptual framework. *Clinical pharmacology & therapeutics*, 69(3):89–95, 2001.

[61] Esther M Hoffmann, Linda M Zangwill, Jonathan G Crowston, and Robert N Weinreb. Optic disk size and glaucoma. *Survey of ophthalmology*, 52(1):32–49, 2007.

[62] Tarek Alasil, Kaidi Wang, Pearse A. Keane, Hang Lee, Neda Baniasadi, Johannes F. De Boer, and Teresa C. Chen. Analysis of normal retinal nerve fiber layer thickness by age, sex, and race using spectral domain optical coherence tomography. *Journal of Glaucoma*, 22(7):532–541, 2013.

[63] Christopher Bowd, Robert N. Weinreb, Julia M. Williams, and Linda M. Zangwill. The retinal nerve fiber layer thickness in ocular hypertensive, normal, and glaucomatous eyes with optical coherence tomography. *Archives of Ophthalmology*, 118(1):22–26, 2000.

[64] Jonas M.D. Gmeiner, Wolfgang A. Schrems, Christian Y. Mardin, Robert Laemmer, Friedrich E. Kruse, and Laura M. Schrems-Hoesl. Comparison of bruch's membrane opening minimum rim width and peripapillary retinal nerve fiber layer thickness in early glaucoma assessment. *Investigative Ophthalmology and Visual Science*, 57(9Special Issue):OCT575–OCT584, 2016.

[65] Yong Woo Kim, Jin Wook Jeoung, Dai Woo Kim, Michael JA Girard, Jean Martial Mari, Ki Ho Park, and Dong Myung Kim. Clinical assessment of lamina cribrosa curvature in eyes with primary open-angle glaucoma. *PloS one*, 11(3):e0150260, 2016.

[66] Varsha Manjunath, Heeral Shah, James G Fujimoto, and Jay S Duker. Analysis of peripapillary atrophy using spectral domain optical coherence tomography. *Ophthalmology*, 118(3):531–536, 2011.

[67] Saad Albawi, Tareq Abed Mohammed, and Saad Al-Zawi. Understanding of a convolutional neural network. In *2017 International Conference on Engineering and Technology (ICET)*, pages 1–6. Ieee, 2017.

[68] Shuai Sun, Bin Hu, Zhou Yu, and Xiaona Song. A stochastic max pooling strategy for convolutional neural network trained by noisy samples. *International Journal of Computers Communications & Control*, 15(1), 2020.

[69] Vinod Nair and Geoffrey E Hinton. Rectified linear units improve restricted boltzmann machines. In *Icml*, 2010.

[70] Sridhar Narayan. The generalized sigmoid activation function: Competitive supervised learning. *Information sciences*, 99(1-2):69–82, 1997.

[71] Matías Roodschild, Jorge Gotay Sardiñas, and Adrián Will. A new approach for the vanishing gradient problem on sigmoid activation. *Progress in Artificial Intelligence*, 9(4):351–360, 2020.

[72] Xavier Glorot, Antoine Bordes, and Yoshua Bengio. Deep sparse rectifier neural networks. In *Proceedings of the fourteenth international conference on artificial intelligence and statistics*, pages 315–323. JMLR Workshop and Conference Proceedings, 2011.

[73] John S Bridle. Probabilistic interpretation of feedforward classification network outputs, with relationships to statistical pattern recognition. In *Neurocomputing*, pages 227–236. Springer, 1990.

[74] Min Lin, Qiang Chen, and Shuicheng Yan. Network in network. *arXiv preprint arXiv:1312.4400*, 2013.

[75] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International conference on machine learning*, pages 448–456. PMLR, 2015.

[76] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[77] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc., 2019.

[78] NVIDIA, Péter Vingelmann, and Frank H.P. Fitzek. Cuda, release: 10.2.89, 2020.

[79] G. Bradski. The OpenCV Library. *Dr. Dobb's Journal of Software Tools*, 2000.

[80] Carole H Sudre, Wenqi Li, Tom Vercauteren, Sebastien Ourselin, and M Jorge Cardoso. Generalised dice overlap as a deep learning loss function for highly unbalanced segmentations. In *Deep learning in medical image analysis and multimodal learning for clinical decision support*, pages 240–248. Springer, 2017.

[81] Giacomo Savini, Piero Barboni, Michele Carbonelli, and Maurizio Zanini. The effect of scan diameter on retinal nerve fiber layer thickness measurement using stratus optic coherence tomography. *Archives of ophthalmology*, 125(7):901–905, 2007.

[82] Joel S Schuman, Tamar Pedut-Kloizman, Ellen Hertzmark, Michael R Hee, Jason R Wilkins, Jeffery G Coker, Carmen A Puliafito, James G Fujimoto, and Eric A Swanson. Reproducibility of nerve fiber layer thickness measurements using optical coherence tomography. *Ophthalmology*, 103(11):1889–1898, 1996.

[83] J Martin Bland and Douglas G Altman. Measuring agreement in method comparison studies. *Statistical methods in medical research*, 8(2):135–160, 1999.

[84] Davide Giavarina. Understanding bland altman analysis. *Biochemia medica*, 25(2):141–151, 2015.

[85] Samuel Sanford Shapiro and Martin B Wilk. An analysis of variance test for normality (complete samples). *Biometrika*, 52(3/4):591–611, 1965.

[86] Henry B Mann and Donald R Whitney. On a test of whether one of two random variables is stochastically larger than the other. *The annals of mathematical statistics*, pages 50–60, 1947.

[87] Stephen Kokoska and Daniel Zwillinger. *CRC standard probability and statistics tables and formulae*. Crc Press, 2000.

[88] Maribel Hidalgo-Aguirre, Julian Gitelman, Mark R Lesk, and Santiago Costantino. Automatic segmentation of the optic nerve head for deformation measurements in video rate optical coherence tomography. *Journal of biomedical optics*, 20(11):116008, 2015.

[89] Jonas Močkus. On bayesian methods for seeking the extremum. In *Optimization techniques IFIP technical conference*, pages 400–404. Springer, 1975.

[90] Marta Para-Prieto, Raul Martin, Sara Crespo, Laura Mena-Garcia, Andres Valisena, Lisandro Cordero, Gloria Gonzalez Fernandez, Juan F Arenillas, Nieves Tellez, and Jose Carlos Pastor. Oct variability prevents their use as robust biomarkers in multiple sclerosis. *Clinical Ophthalmology (Auckland, NZ)*, 15:2025, 2021.

[91] Alberto Domínguez-Vicent, Rune Brautaset, and Abinaya Priya Venkataraman. Repeatability of quantitative measurements of retinal layers with sd-oct and agreement between vertical and horizontal scan protocols in healthy eyes. *PLoS One*, 14(8):e0221466, 2019.

[92] Christopher Kai-shun Leung, Arthur Chak Kwan Cheng, Kelvin Kam Lung Chong, King Sai Leung, Shaheeda Mohamed, Charles Sing Lok Lau, Carol Yim Lui Cheung, Geoffrey Chin-hung Chu, Ricky Yiu Kwong Lai, Calvin Chi Pui Pang, et al. Optic disc measurements in myopia with optical coherence tomography and confocal scanning laser ophthalmoscopy. *Investigative ophthalmology & visual science*, 48(7):3178–3183, 2007.

[93] Dmitrij Sitenko, Bastian Boll, and Christoph Schnörr. Assignment flow for order-constrained oct segmentation. In *Pattern Recognition: 42nd DAGM German Conference, DAGM GCPR 2020, Tübingen, Germany, September 28–October 1, 2020, Proceedings 42*, pages 58–71. Springer, 2021.

[94] Sri Gowtham Thakku, Yih-Chung Tham, Mani Baskaran, Jean-Martial Mari, Nicholas G Strouthidis, Tin Aung, Ching-Yu Cheng, and Michael JA Girard. A global shape index to characterize anterior lamina cribrosa morphology and its determinants in healthy indian eyes. *Investigative ophthalmology & visual science*, 56(6):3604–3614, 2015.

[95] Md Habibur Rahman, Hyeon Woo Jeong, Na Rae Kim, and Dae Yu Kim. Automatic quantification of anterior lamina cribrosa structures in optical coherence tomography using a two-stage cnn framework. *Sensors*, 21(16):5383, 2021.

[96] Frédéric Pollet-Villard, Christophe Chiquet, Jean-Paul Romanet, Christian Noel, and Florent Aptel. Structure-function relationships with spectral-domain optical coherence tomography retinal nerve fiber layer and optic nerve head measurements. *Investigative ophthalmology & visual science*, 55(5):2953–2962, 2014.

[97] Stephanie Mroczkowska, Alexandra Benavente-Perez, Anil Negi, Velota Sung, Sunni R Patel, and Doina Gherghel. Primary open-angle glaucoma vs normal-tension glaucoma: the vascular perspective. *JAMA ophthalmology*, 131(1):36–43, 2013.

[98] B Turgut and FA Turgut. Differences between the characteristics of normal tension glaucoma and high tension glaucoma. *Adv Ophthalmol Vis Syst*, 7(7):00250, 2017.

[99] Penpe Gul Firat, Selim Doganay, Ersan Ersin Demirel, and Cemil Colak. Comparison of ganglion cell and retinal nerve fiber layer thickness in primary open-angle glaucoma and normal tension glaucoma with spectral-domain oct. *Graefe's Archive for Clinical and Experimental Ophthalmology*, 251(3):831–838, 2013.

[100] M Rosario Hernandez. Ultrastructural immunocytochemical analysis of elastin in the human lamina cribrosa. changes in elastic fibers in primary open-angle glaucoma. *Investigative ophthalmology & visual science*, 33(10):2891–2903, 1992.

[101] Fabian Isensee, Philipp Kickingereder, Wolfgang Wick, Martin Bendszus, and Klaus H Maier-Hein. No new-net. In *International MICCAI Brainlesion Workshop*, pages 234–244. Springer, 2018.

[102] Luisa Sanchez Brea, Danilo Andrade De Jesus, Stefan Klein, and Theo van Walsum. Deep learning-based retinal vessel segmentation with cross-modal evaluation. In *Medical Imaging with Deep Learning*, pages 709–720. PMLR, 2020.

[103] Ehsan Hosseini Asl, Mohammed Ghazal, Ali Mahmoud, Ali Aslantas, Ahmed Shalaby, Manual Casanova, Gregory Barnes, Georgy Gimel'farb, Robert Keynton, and Ayman El-Baz. Alzheimer's disease diagnostics by a 3d deeply supervised adaptable convolutional network. *Front. Biosci. Landmark*, 23(3):584–596, 2018.

[104] Nithya Sambasivan, Shivani Kapania, Hannah Highfill, Diana Akrong, Praveen Paritosh, and Lora M Aroyo. "everyone wants to do the model work, not the data work": Data cascades in high-stakes ai. In *proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, pages 1–15, 2021.

[105] Peter Henderson, Riashat Islam, Philip Bachman, Joelle Pineau, Doina Precup, and David Meger. Deep reinforcement learning that matters. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32, 2018.

[106] Zachary C Lipton and Jacob Steinhardt. Research for practice: troubling trends in machine-learning scholarship. *Communications of the ACM*, 62(6):45–53, 2019.

[107] Efstathios D Gennatas, Jerome H Friedman, Lyle H Ungar, Romain Pirracchio, Eric Eaton, Lara G Reichmann, Yannet Interian, José Marcio Luna, Charles B Simone, Andrew Auerbach, et al. Expert-augmented machine learning. *Proceedings of the National Academy of Sciences*, 117(9):4571–4577, 2020.

[108] Fuzhen Zhuang, Zhiyuan Qi, Keyu Duan, Dongbo Xi, Yongchun Zhu, Hengshu Zhu, Hui Xiong, and Qing He. A comprehensive survey on transfer learning. *Proceedings of the IEEE*, 109(1):43–76, 2020.

[109] Michael Crawshaw. Multi-task learning with deep neural networks: A survey. *arXiv preprint arXiv:2009.09796*, 2020.

[110] Jesper E Van Engelen and Holger H Hoos. A survey on semi-supervised learning. *Machine Learning*, 109(2):373–440, 2020.

[111] James Fujimoto and Eric Swanson. The development, commercialization, and impact of optical coherence tomography. *Investigative ophthalmology & visual science*, 57(9):OCT1–OCT13, 2016.

[112] Alexander Ehnes, Yaroslava Wenner, Christoph Friedburg, Markus N Preising, Wadim Bowl, Walter Sekundo, Erdmuthe Meyer Zu Bexten, Knut Stieger, and Birgit Lorenz. Optical coherence tomography (oct) device independent intraretinal layer segmentation. *Translational vision science & technology*, 3(1):1–1, 2014.

[113] Luis de Sisternes, Jerry Hong, Theodore Leng, and Daniel L Rubin. A machine learning approach for device-independent automated segmentation of retinal cysts in spectral domain optical coherence tomography images. *Proceeding Optima Challenge-MICCAI*, 2015.

[114] Connor Shorten and Taghi M Khoshgoftaar. A survey on image data augmentation for deep learning. *Journal of Big Data*, 6(1):1–48, 2019.

Appendix A

Binary segmentation

This appendix comprises examples from the binary segmentation results on each of the isolated tissues individually and on all together as a whole. Each figure shows the original image, the manual segmentation mask and the automatic segmentation mask.

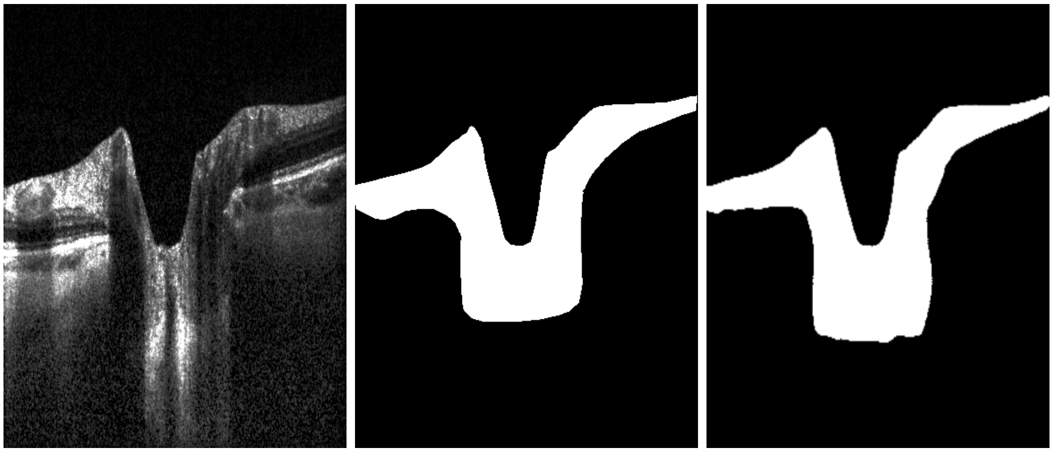


Figure A.1: Example from the results of the binary classification of the RNFL. (a) Original image. (b) Manual segmentation. (c) Automatic segmentation.

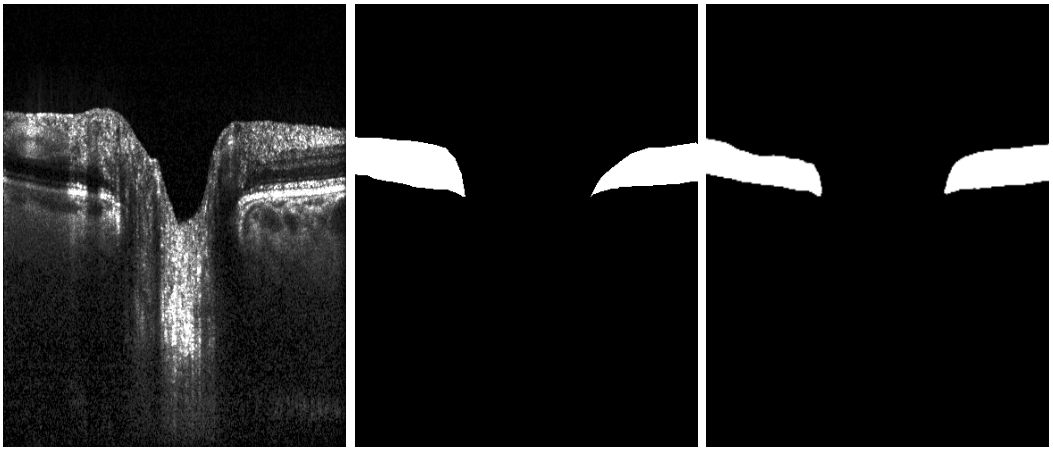


Figure A.2: Example from the results of the binary classification of the other retinal layers. (a) Original image. (b) Manual segmentation. (c) Automatic segmentation.

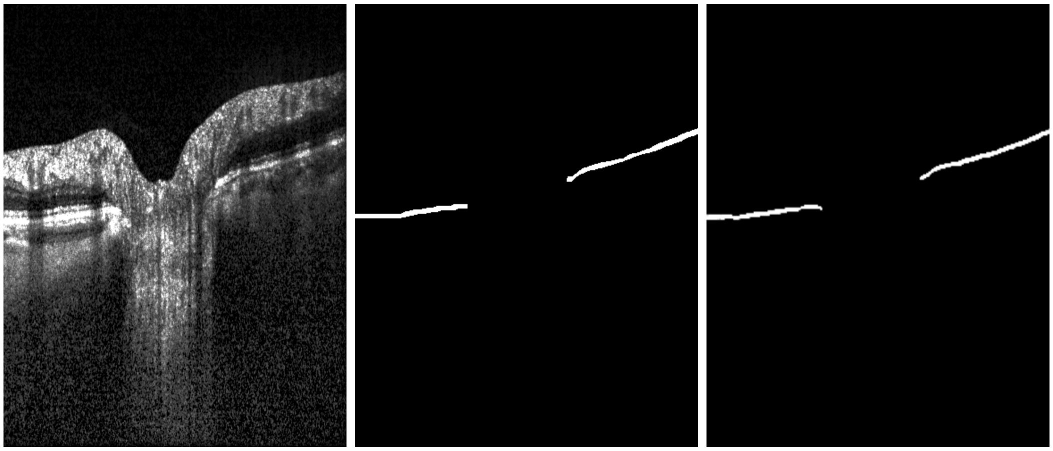


Figure A.3: Example from the results of the binary classification of the RPE. (a) Original image. (b) Manual segmentation. (c) Automatic segmentation.

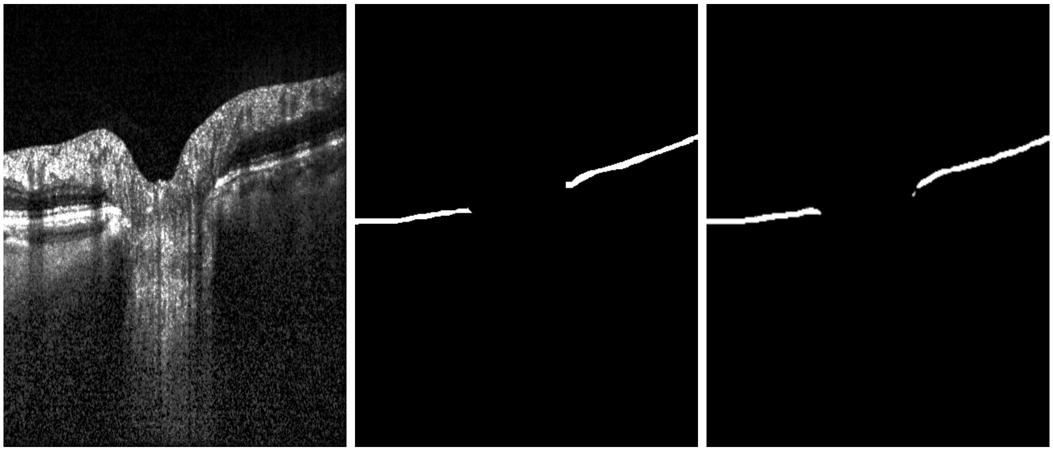


Figure A.4: Example from the results of the binary classification of the BM. (a) Original image. (b) Manual segmentation. (c) Automatic segmentation.

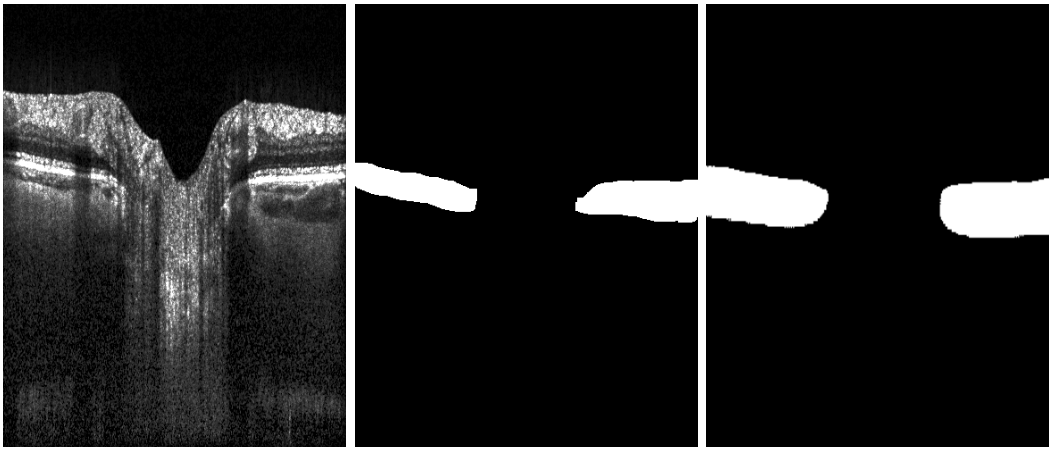


Figure A.5: Example from the results of the binary classification of the choroid. (a) Original image. (b) Manual segmentation. (c) Automatic segmentation.

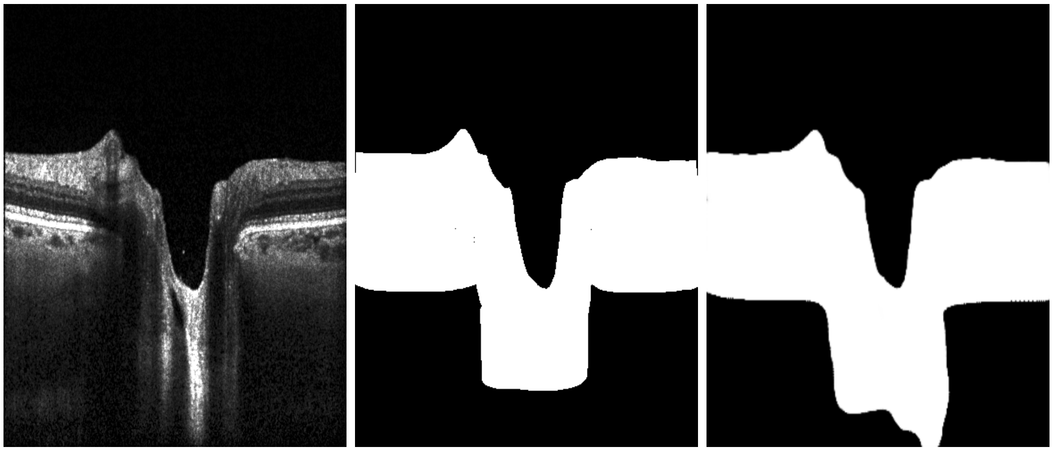


Figure A.6: Example from the results of the binary classification of all the labels as a whole. (a) Original image. (b) Manual segmentation. (c) Automatic segmentation.

Addendum

During the defense of the master thesis entitled "Automatic Segmentation of the Optic Nerve Head in Optical Coherence Tomography Data" of Rita Maria Vieira Carvalho Marques, for the fulfilment of the requirements of the Master's Degree in Biomedical Engineering, that took place on 9th November 2021, at the Departamento de Física da Universidade de Coimbra, before the jury composed of Professors António Miguel Morgado, Pedro Serranho, Rui Bernardes, and Doctor Danilo Jesus (supervisor), a fundamental methodological question raised on the similarity and dimension of the training set after data augmentation that may compromise the proposed neural network to carry semantic segmentation. Despite the disclosed facts that horizontal image flip was applied and rotations between -8 and +8 degrees, the large majority of the images in the training and test sets had the same optic disc location. Consequently, it was not assured the sufficient variability on the optic disc location to assume the neural network is taking enough image information towards the proposed segmentation rather than location.

The jury proposed the candidate to carry additional tests, namely to test the model with images with displacement of the optic disc location from the center of the image. If this results in bad segmentation, then sufficient images with displacements of the optic disc off the center of the image should be added to the training set, increasing its dimension and variability. Results should be then assessed for the new obtained network. Because many networks were tested throughout the work, the jury proposed these tests to be applied only to the network that achieved the best performance. Finally, a new section should be added to the presented work with these findings, explaining the methodology, the number of images added to the training set, results achieved after that, and the comparison of these with previous ones.

Additional tests

There were doubts regarding whether or not the proposed model could generalize to non-centred ONH B-scans because “the large majority of the images in the training and test sets had the same optic disc location”. In order to answer the raised question and as suggested by the Jury, the best model (model 5, see Tables 4.2 and 4.3) was re-tested on images with the optic disc location displaced from the image centre. For that, four new test sets, based on the test set A+B, were created by shifting the ONH 50 and 100 pixels to the right and to the left. These values were chosen in order to have large displacements while ensuring that the optic disc remained in the image, as shown in Figure A.7.

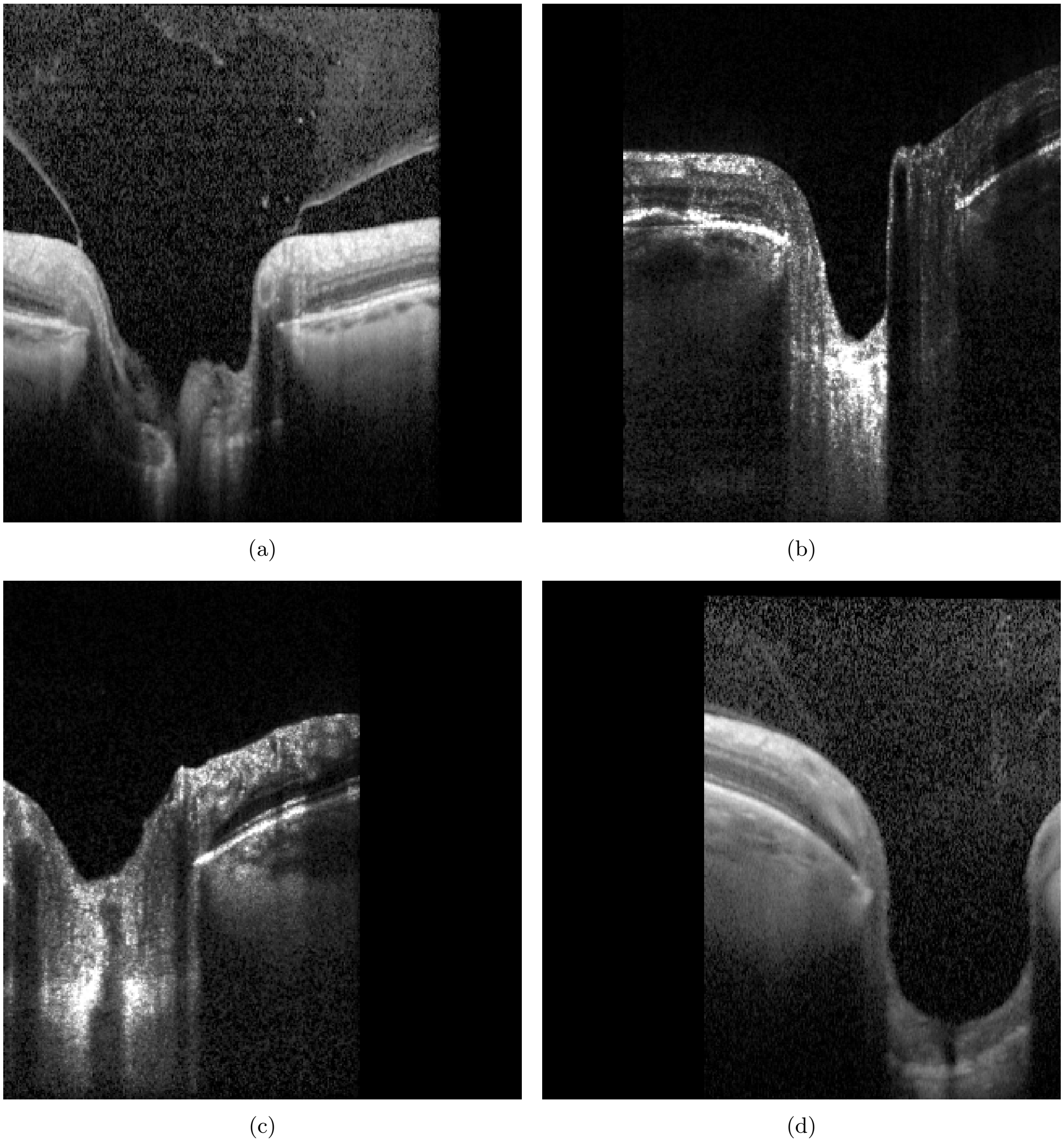
The results obtained for model 5 applied to the four displaced test sets can be compared to the original results in Table A.1.

Table A.1: Testing results for four tests of model 5 from Table 4.2: on the original test set and with all images from the test set shifted 50 and 100 pixels to the right and to the left. Mean Dice coefficient, sensitivity and specificity are showed for all tissues.

Metrics	Tissue	Best model original results	Shift to the left		Shift to the right	
			50 pixels	100 pixels	50 pixels	100 pixels
Dice coefficient (Mean ± SD)	RNFL	0.91 ± 0.05	0.88 ± 0.04	0.85 ± 0.07	0.89 ± 0.05	0.83 ± 0.05
	Other retinal layers	0.88 ± 0.06	0.83 ± 0.08	0.78 ± 0.09	0.83 ± 0.09	0.80 ± 0.08
	BM/RPE complex	0.70 ± 0.14	0.68 ± 0.14	0.66 ± 0.13	0.69 ± 0.15	0.69 ± 0.18
	Choroid	0.82 ± 0.08	0.82 ± 0.07	0.79 ± 0.08	0.82 ± 0.06	0.79 ± 0.08
	LC	0.63 ± 0.22	0.63 ± 0.24	0.61 ± 0.24	0.63 ± 0.23	0.62 ± 0.23
Sensitivity (Mean ± SD)	RNFL	0.93 ± 0.04	0.92 ± 0.05	0.88 ± 0.09	0.92 ± 0.05	0.85 ± 0.08
	Other retinal layers	0.86 ± 0.09	0.82 ± 0.10	0.77 ± 0.11	0.81 ± 0.12	0.78 ± 0.13
	BM/RPE complex	0.89 ± 0.10	0.80 ± 0.12	0.77 ± 0.11	0.81 ± 0.14	0.80 ± 0.18
	Choroid	0.92 ± 0.04	0.85 ± 0.07	0.83 ± 0.08	0.81 ± 0.09	0.76 ± 0.13
	LC	0.67 ± 0.23	0.64 ± 0.25	0.60 ± 0.25	0.64 ± 0.25	0.61 ± 0.24
Specificity (Mean ± SD)	RNFL	0.98 ± 0.01	0.98 ± 0.01	0.98 ± 0.01	0.98 ± 0.01	0.98 ± 0.01
	Other retinal layers	0.99 ± 0.00	0.99 ± 0.00	0.99 ± 0.00	0.99 ± 0.00	0.99 ± 0.00
	BM/RPE complex	0.99 ± 0.00	0.99 ± 0.00	0.99 ± 0.00	0.99 ± 0.00	0.99 ± 0.00
	Choroid	0.98 ± 0.01	0.99 ± 0.01	0.99 ± 0.01	0.99 ± 0.01	0.99 ± 0.00
	LC	0.99 ± 0.01	0.99 ± 0.01	0.99 ± 0.01	0.99 ± 0.01	0.99 ± 0.00

The results show that the proposed model is able to segment all structures even if the ONH B-scan is not centred. The metrics obtained for the displaced test sets are very similar, although slightly lower, than those obtained for the original A+B test set. The highest difference for the Dice coefficient was obtained for the RNFL and the “other retinal layers”. Sensitivities were lower for the “other retinal layers”, the BM/RPE complex, and the choroid, namely the choroid on the test set shifted 100 pixels to the right. Specificities were almost equal across all test sets. Examples of the results obtained for each of the four test sets are shown in Figure A.8.

By visual inspection of the segmentations predicted on the displaced images, it is possible to



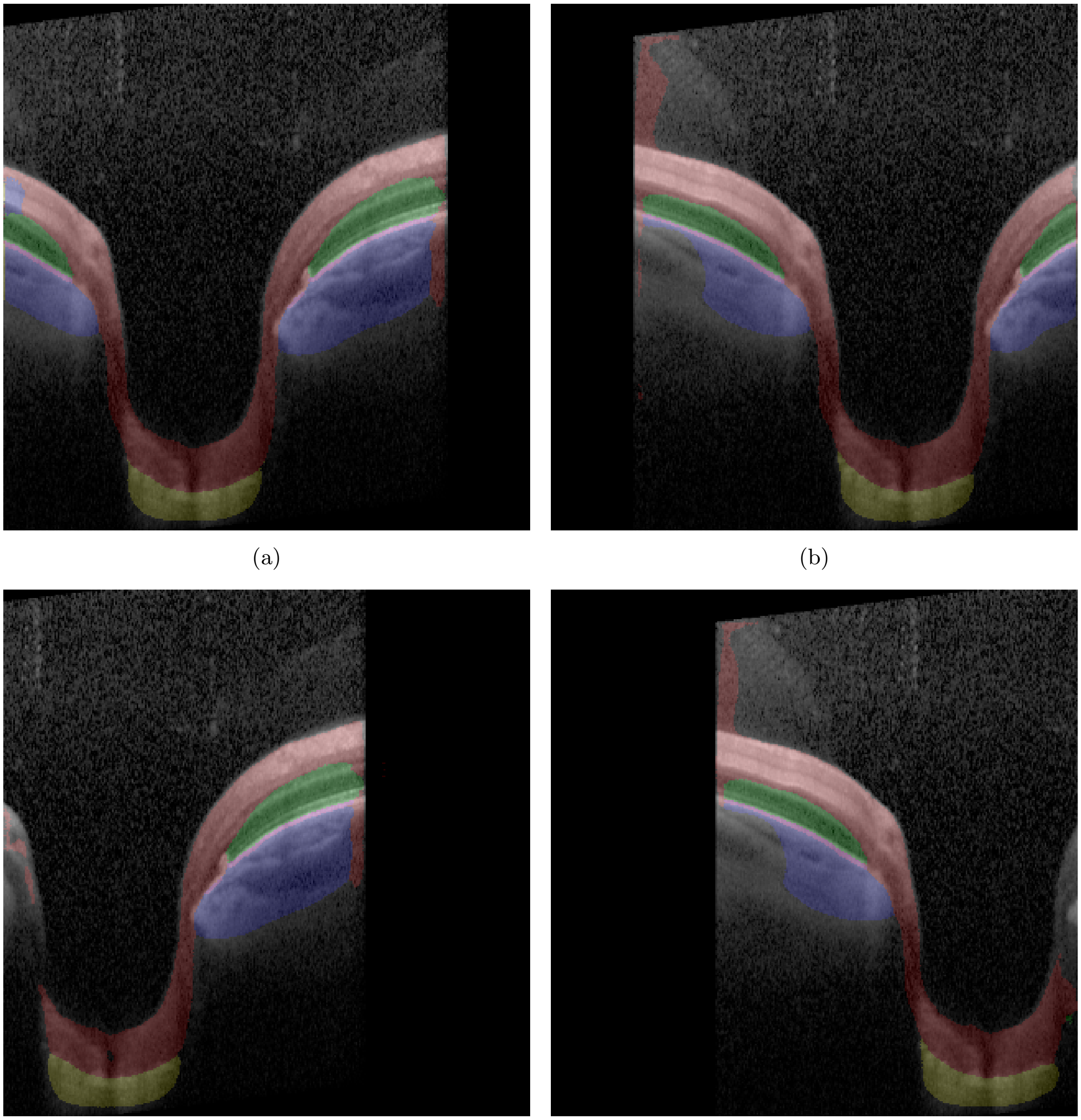
(a)

(b)

(c)

(d)

Figure A.7: Examples of B-scans from the test set shifted (a) 50 pixels to the left, (b) 50 pixels to the right, (c) 100 pixels to the left, and (d) 100 pixels to the right.



(a)

(b)

(c)

(d)

Figure A.8: Example of the model prediction for a B-scan shifted (a) 50 pixels to the left, (b) 50 pixels to the right, (c) 100 pixels to the left, and (d) 100 pixels to the right.

verify that most errors occur at the RNFL, choroid, the BM/RPE complex and the "other retinal layers", which is in agreement with the lower metrics obtained for these tissues in comparison to the ones of the original test set. However, as shown in Figure A.8, these segmentation errors occur at the image borders due to the border effect introduced when shifting the ONH B-scans. A discontinuity, not representative of real world data, is artificially created on the images leading to a loss of information, and therefore, errors on the tissue borders. The LC, on the other hand, does not suffer any border effect and the obtained results are similar when comparing the four displaced test sets to the centred ONH test set. Even so, the experiment performed for this Addendum shows that the proposed model is robust to severe lateral displacements of the ONH. This means that model is not biased towards the low variability of the ONH location in the training and testing sets.

Nevertheless, hypothesising that the model was biased, and it could not generalize to non-centred ONH images, adding more displacements of the optic disc to the training set in order to increase its dimension and variability is not the solution. In order to show that, a new model was trained using augmentation with large ONH displacements. This new model used the same parameters as the best model presented in Chapter 4, with additional data augmentation. The data augmentation consisted of rotations until 30 degrees, and vertical and horizontal translations up to 100 pixels. After the training, the new model was tested on all test sets, the A+B test set and the four variations as mentioned above. The results are shown in Table A.2.

Table A.2: Testing results for the best model and for the new model on the original test set, and on the test sets shifted 50 and 100 pixels to the right and to the left. Mean Dice coefficient, sensitivity and specificity are showed for all tissues. ORL stands for "other retinal layers".

Metrics	Tissue	Best model original results	New model results				
			Original dataset	50 pixels left	100 pixels left	50 pixels right	100 pixels right
Dice coefficient (Mean ± SD)	RNFL	0.91 ± 0.05	0.86 ± 0.10	0.60 ± 0.16	0.58 ± 0.14	0.62 ± 0.16	0.59 ± 0.16
	ORL	0.88 ± 0.06	0.80 ± 0.16	0.07 ± 0.11	0.08 ± 0.09	0.16 ± 0.16	0.17 ± 0.13
	BM/RPE	0.70 ± 0.14	0.63 ± 0.20	0.13 ± 0.12	0.16 ± 0.14	0.19 ± 0.15	0.21 ± 0.15
	Choroid	0.82 ± 0.08	0.80 ± 0.09	0.17 ± 0.14	0.12 ± 0.10	0.26 ± 0.18	0.29 ± 0.23
	LC	0.63 ± 0.22	0.54 ± 0.28	0.46 ± 0.30	0.42 ± 0.34	0.44 ± 0.26	0.37 ± 0.32
Sensitivity (Mean ± SD)	RNFL	0.93 ± 0.04	0.82 ± 0.14	0.47 ± 0.16	0.44 ± 0.14	0.49 ± 0.18	0.45 ± 0.16
	ORL	0.86 ± 0.09	0.77 ± 0.22	0.04 ± 0.07	0.04 ± 0.05	0.10 ± 0.11	0.10 ± 0.09
	BM/RPE	0.89 ± 0.10	0.67 ± 0.25	0.08 ± 0.09	0.10 ± 0.10	0.12 ± 0.11	0.14 ± 0.10
	Choroid	0.92 ± 0.04	0.84 ± 0.13	0.10 ± 0.10	0.07 ± 0.06	0.18 ± 0.13	0.20 ± 0.18
	LC	0.67 ± 0.23	0.53 ± 0.31	0.43 ± 0.30	0.40 ± 0.35	0.37 ± 0.23	0.34 ± 0.33
Specificity (Mean ± SD)	RNFL	0.98 ± 0.01	0.99 ± 0.01	0.99 ± 0.00	0.99 ± 0.00	0.99 ± 0.00	0.99 ± 0.00
	ORL	0.99 ± 0.00	0.99 ± 0.00	1.00 ± 0.00	1.00 ± 0.00	0.99 ± 0.00	1.00 ± 0.00
	BM/RPE	0.99 ± 0.00	0.99 ± 0.00	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00
	Choroid	0.98 ± 0.01	0.98 ± 0.01	0.99 ± 0.00	0.99 ± 0.00	0.99 ± 0.00	0.99 ± 0.00
	LC	0.99 ± 0.01	0.99 ± 0.01	0.99 ± 0.00	0.99 ± 0.00	0.99 ± 0.00	0.99 ± 0.00

The results show that there is a clear decrease in performance introduced by the increase in dimension and variability in the augmentation. Although combining augmentations can result in massively inflated dataset sizes, it is not guaranted to be advantageous, mainly when training models with very limited data. Hence, it is important to define a reasonable augmentation search space for deriving an optimal subset of augmented data to train the models, as performed in this project. As shown by the AI community over the last years, data augmentation cannot overcome all biases present in a small dataset [114].