Glaucoma in Fundus Image

José Luís Pacheco Martins

U. PORTO

FEUP **FACULDADE DE ENGENHARIA**
UNIVERSIDADE DO PORTO

Glaucoma in Fundus Image

José Luís Pacheco Martins

Mestrado Integrado em Engenharia Informática e Computação

Approved in oral examination by the committee:

Chair: Doctor Henrique Daniel de Avelar Lopes Cardoso
External Examiner: Doctor Catarina de Brito Carvalho
Supervisor: Doctor Jaime dos Santos Cardoso

July 24, 2019

Abstract

Glaucoma, an eye condition which leads to permanent blindness, is typically asymptomatic and therefore difficult to be diagnosed in time if a doctor is not seen regularly. However, if diagnosed in time, Glaucoma can effectively be slowed down by using adequate treatment, therefore an early diagnosis is of utmost importance.

Nonetheless, the conventional approaches to diagnose Glaucoma adopt expensive and bulky equipment that requires qualified experts, making it difficult, costly and time-consuming to diagnose large amounts of people.

Considering the aforementioned, this work proposes a solution to this problem, by developing a computer-aided diagnosis system that is capable of diagnosing Glaucoma, in an autonomous way, using fundus images. These images contain the necessary morphological features to determine the presence of Glaucoma and can be obtained with considerably more affordable equipment, being therefore very relevant for the task at hand.

Several computer-aided systems for Glaucoma assessment with fundus images have already been proposed, however, they lack in several aspects. This work is a step forward in the field, by improving existing techniques and addressing still unexplored fields.

The developed solution uses Deep Learning techniques powered by a collection of public datasets and an acquired private dataset to construct an interpretable Glaucoma assessment pipeline that runs offline in mobile and embedded devices.

This pipeline is then integrated on the EyeFundusScope android application, developed by Fraunhofer AICOS, which is part of a larger system that already provides solutions for fundus image acquisition and Diabetic Retinopathy detection.

The result demonstrates the potential that these systems can have in the contribution to an early Glaucoma diagnosis and enhances the EyeFundusScope system, turning it into a more complete screening tool.

Resumo

O Glaucoma é uma condição ocular que conduz à cegueira permanente, geralmente é assintomático e, portanto, difícil de ser diagnosticado a tempo se um médico não for visitado regularmente. No entanto, se diagnosticado a tempo, o Glaucoma pode ser efetivamente abrandado através de tratamento adequado, desta forma, um diagnóstico precoce é de extrema importância.

No entanto, as abordagens convencionais para diagnosticar o Glaucoma recorrem a equipamentos caros e volumosos que exigem especialistas qualificados, tornando difícil, dispendioso e demorado o diagnóstico de uma grande quantidade de pessoas.

Este trabalho propõe uma solução para este problema, desenvolvendo um sistema de diagnóstico assistido por computador capaz de diagnosticar o Glaucoma, de uma forma autónoma, usando imagens do fundus. Estas imagens contêm as características morfológicas necessárias para determinar a presença de Glaucoma e podem ser obtidas por equipamentos consideravelmente mais acessíveis, sendo por isso muito relevantes para a tarefa em questão.

Vários sistemas auxiliados por computador para avaliação do Glaucoma com imagens do fundus foram já propostos, no entanto, estes carecem em diversos aspetos. Este trabalho é um passo em frente nesta área, melhorando as técnicas existentes e abordando tópicos ainda não explorados.

A solução desenvolvida usa técnicas de *Deep Learning*, contruídas com um conjunto de *datasets* públicos e um *dataset* privado para construir uma *pipeline* interpretável de avaliação do Glaucoma que pode ser executada *offline* em dispositivos móveis e embebidos.

Esta *pipeline* é posteriormente integrada na aplicação android EyeFundusScope, desenvolvida pela Fraunhofer AICOS, que faz parte de um sistema maior que já fornece soluções para aquisição de imagens do fundus e detecção de Retinopatia Diabética.

O resultado do trabalho desenvolvido demonstra o potencial que estes sistemas podem ter na contribuição para um diagnóstico precoce do Glaucoma e melhora o sistema EyeFundusScope, transformando-o numa ferramenta de triagem mais completa.

Acknowledgements

First I would like to show my gratitude to Fraunhofer AICOS, for providing me a challenging and empowering environment for the development of this work and especially to my Supervisor at the institution, Filipe Soares, for all the guidance and advice given throughout this journey at Fraunhofer. I also thank my second supervisor, Jaime Cardoso, for the help provided and for constantly proposing new ideas and solutions.

To all my friends, and especially to my academic and Erasmus buddy, David Azevedo, thank you for all the support, and for all the moments of joy during this period.

Finally, one very special thank to my family and to my girlfriend for always being there for me, nothing would have been possible without you.

José Luis Pacheco Martins

"Nothing in life is to be feared, it is only to be understood.
Now is the time to understand more, so that we may fear less."

Marie Curie

Contents

CONTENTS

List of Figures

List of Tables

LIST OF TABLES

Abbreviations

ACG   Angle-closure Glaucoma
ASPP   Atrous Spatial Pyramid Pooling
CAD   Computer Aided Diagnosis
CDR   Cup-to-disc Ratio
CNN   Convolutional Neural Network
DDLS   Disk damage likelihood scale
DL   Deep Learning
EFS   EyeFundusScope
GFI   Glaucoma in Fundus Image
IOP   Intraocular Pressure
IoU   Intersection over union
ML   Machine Learning
NRR   Neuro-Retinal Rim
OC   Optic Cup
OD   Optic Disc
ONH   Optic Nerve Head
POAG   Primary open-angle Glaucoma
PPA   Peripapillary Atrophy
RDAR   Rim to disc area ratio
RNFL   Retinal Nerve Fiber Layer
ROC   Receiver Operating Characteristic
ROI   Region of Interest
SCAS   Semana das Ciências Aplicadas à Saúde
SPP   Spatial Pyramid Pooling
VCD   Vertical cup diameter
VCDR   Vertical Cup-to-disc Ratio
VDD   Vertical disc diameter

Chapter 1

Introduction

In this chapter the context and motivation of this work are briefly explained in section 1.1, followed by the work main goals in section 1.2. The chapter provides the reader with an overview of the problems addressed in this work as well as the vital methodologies that are used. Finally, the chapter closes with an explanation of the remaining structure of the document in section 1.3.

1.1 Context and Motivation

Glaucoma is a group of chronic eye diseases and a leading cause of irreversible blindness worldwide [FCX⁺18a], being therefore considered a great public health concern.

Glaucoma occurs due to an inappropriate intraocular pressure, related with an inability of the eye to properly manage the balance between the fluid produced in the eye and the fluid that leaves the eye through its drainage canal [HKT⁺18]. The undesired pressure inside the eye ends up damaging the optic nerve head, which starts to progressively create blind spots in the peripheral vision, that gradually convey to the center [Shr16]. As mentioned, Glaucoma is a group of diseases, these diseases differ between each other in the way this abnormal level of pressure is achieved inside the eye.

Currently, primary open angle Glaucoma (POAG) is the most common type of Glaucoma, accounting for 74% of all Glaucoma cases [QB06]. In POAG the level of intraocular pressure increases due to the clogging of the drainage canal in the eye [Shr16].

Additionally, Glaucoma is known as a 'silent thief of sight', as it typically does not have any early symptoms, turning its diagnose into a challenging task if a doctor is not seen regularly.

The traditional methods to diagnose Glaucoma consist in a comprehensive eye examination, including slit lamp examination, tonometry to measure the intraocular pressure, gonioscopy to examine the angle of the anterior chamber, perimetry to detect functional defects in the visual field and others [TLP11]. All the mentioned methods require bulky and expensive equipment as well as highly qualified professionals to operate them. Additionally, due to the high costs incurred

by these methods, some of them are only performed if Glaucoma is already suspected. This turns Glaucoma diagnosis into a rather challenging task, prone to human errors, which can lead to a substantial number of undiagnosed cases.

This number of undiagnosed cases is a big concern in the healthcare community since, despite being irreversible, Glaucoma can be effectively slowed down with appropriate treatment, therefore reducing the risk of total blindness. Hence, achieving an early diagnosis of Glaucoma is of utmost importance. This results in the demand for new diagnose techniques that aid or replace the existing ones, turning Glaucoma assessment into a more efficient process.

To respond to this demand, fundus images have been used; these images contain the necessary morphological features to identify Glaucoma and other diseases. Furthermore, fundus images can be obtained with considerably more affordable and portable cameras, turning them into a very relevant research topic nowadays.

In the past decade, a considerable amount of research has been done on how to use this images to build new CAD systems for Glaucoma detection and new emerging techniques inside the machine learning field have been widely used. The term 'machine learning' generally refers to any pattern-classification method that requires a training scheme, and inside the machine learning field, deep learning is the newest and fastest growing topic; it can be considered as a brute-force approach to determine the most appropriate image-processing filters and tools that can quantify various biomarkers of a disease [GORT18]. These techniques need large amounts of data and computational power for a successful application, especially considering deep learning techniques. This has been an issue in the past, but nowadays these requirements are starting to be fulfilled at a growing pace, making these approaches very suitable for Glaucoma diagnosis.

A variety of CAD systems have already been developed and surveyed [Shr16, HKT+18, TJ18, SBS+18] using fundus images and machine learning techniques. These systems provide a quick diagnose of Glaucoma taking a fundus image as an input and are considerably less costly when compared to the typical techniques. However, there isn't still a CAD system that provides high usability, efficiency, affordability and interpretability that could be used in practice for Glaucoma detection.

One system that aims at solving these issues, is the EyeFundusScope[1], developed by Fraunhofer AICOS. The system consists of an Android application and a custom optical prototype that can be attached to an Android smartphone, and acquire fundus images. This system was already explored to be used for Diabetic Retinopathy diagnosis, although, at the present moment, it isn't capable of diagnosing other diseases, like Glaucoma.

1.2 Goals

The main goal of this work consists in the development of a CAD system that provides high usability, efficiency, affordability and interpretability for Glaucoma detection with fundus images.

---

[1]https://www.fraunhofer.pt/content/dam/portugal/en/documents/Flyers/EyeFundusScope%20-%20FhP-AICOS%20Project%20Brochure.pdf

To accomplish these properties, the system should be able to run efficiently in an android device, and then be integrated into the Fraunhofer EyeFundusScope system mentioned in section 1.1, resulting in an end-to-end mobile-based Glaucoma diagnosis system.

The underlying goals of this work are then to enable massive Glaucoma screenings, increasing the chance of detecting Glaucoma at an earlier stage and hence decrease both individual and economic burdens caused by Glaucoma.

1.3 Document Structure

Besides the present introductory chapter, this document contains 9 additional chapters. In chapter 2, the relevant background is uncovered, providing the reader with the necessary knowledge about the main concepts of the work. Chapter 3 reviews the state-of-the-art, disclosing the main techniques, resources, and related work. Moreover, chapter 4 goes into more detail about the problem that this work addresses, giving a clear overview of all its nuances, and how they will be handled, by defining in detail the proposed solution.

Then, the chapters related to the implementation arise, first, chapter 5 reveals the construction of the datasets used throughout the development of this work and, chapter 6 addresses the main preprocessing techniques applied to the images that constitute the datasets. Furthermore, the implemented segmentation and classification models are addressed in chapters 7 and 8, respectively and, then, the insights of all the previous chapters are combined in chapter 9, where a complete overview of the developed CAD system is provided.

Lastly, in chapter 10, the main conclusions of this work along with the main suggestions for future work are presented.

Introduction

Chapter 2

Background

This chapter uncovers the main concepts of this work, providing the necessary background for understanding its contents.

Firstly, in section 2.1, Glaucoma is explained in detail. Then on section 2.2, fundus images are addressed explaining why these images are relevant for Glaucoma detection, describing their morphology, and identifying features that have been extracted from these images to diagnose Glaucoma.

Moreover, in section 2.3, an overview of machine learning and deep learning techniques is given, describing both classifiers and commonly used metrics. Additionally, on section 2.4, the importance of CAD systems in healthcare is analyzed with some examples of current applications of these systems.

Finally, a summary and some conclusions of the chapter are given in section 2.5.

2.1 Glaucoma

Glaucoma is a cluster of eye diseases that damage the optic nerve head, leading to a progressive and irreversible vision loss. The optic nerve head is responsible for carrying visual information from the retina to the brain, allowing the human to visualize the outside world [TJ18]. When this nerve is damaged, peripheral vision losses start to occur. These losses gradually converge to the center of the vision, ending in permanent blindness in the worst case scenario.

The rise of intraocular pressure (IOP) inside the eye is considered the main cause of Glaucoma; this pressure starts to damage the optic nerve as can be seen in figure 2.1. This is usually related to an inability of the eye to properly manage the balance between the amount of fluid that is produced in the eye and the amount that is drained [HKT+18]. Although it is confirmed that the "raised" intraocular pressure is a causal risk factor for Glaucoma and the only one that can be treated, it is neither sufficient nor necessary for the diagnosis of the disease [TLP11].

Besides the increase in IOP, there are other identified risks for Glaucoma [SBS+18], like:

GLAUCOMA
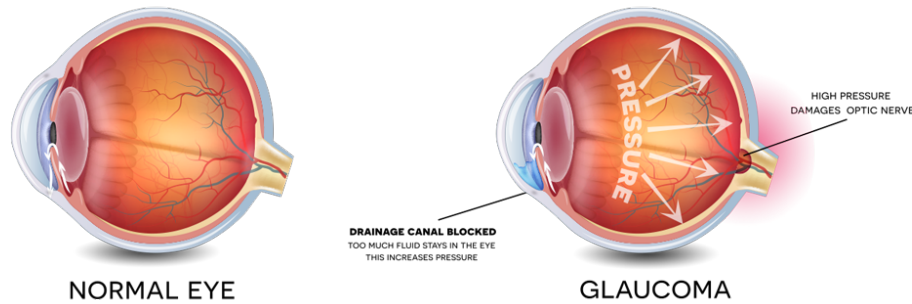
Figure 2.1: Increase in intraocular pressure leading to Glaucoma [Pra18]

- Family history;

- Having a cardiovascular disease;

- Africa, Latino or US ancestry;

- Being over 70 years;

- Blood pressure rate, Steroid use;

- Having myopia.

In the majority of Glaucoma cases, no early symptoms or pain occur, and for that reason, Glaucoma is often called a 'silent thief of sight'. Glaucoma damage is relatively asymptomatic in most cases because of redundancy in the sensory system and the binocular nature of vision, that is, one eye may compensate for early losses in the other [VLGK11]. This turns Glaucoma diagnoses into a very challenging task if a doctor is not seen regularly.

The usual approaches for diagnosing Glaucoma are performed by ophthalmologists and consist of a comprehensive eye examination. This encompasses an external examination of the eye, ocular mobility, examination of the pupil, slit lamp examination, tonometry to measure the intraocular pressure, gonioscopy to examine the angle of the anterior chamber and perimetry to detect functional defects in the visual field [TLP11]. All these methods require trained professionals and expensive equipment, therefore, due to the high costs incurred by these methods, some of them are only performed if Glaucoma is already suspected which contributes to the increasing number of undiagnosed Glaucoma cases.

Furthermore, although Glaucoma cannot be reversed, it can be treated and effectively slowed down; therefore, performing an early diagnosis of the disease is extremely important.

The treatment of Glaucoma diseases depends on the type of Glaucoma [TJ18], although it always aims at managing the intraocular pressure inside the eye. The treatment can be pharmaceutical using eye drops or surgical, consisting in either trabeculectomy, laser surgery or drainage implants [SBS+18].

Currently, Glaucoma is a leading cause of irreversible blindness worldwide, with a continuously growing prevalence over the last years, it is expected that in 2020 there will be around 79.6 million people affected by Glaucoma [QB06]. Therefore, Glaucoma is considered a great public health issue.

2.1.1 Types of Glaucoma

As mentioned, Glaucoma consists of a group of diseases, although the prevalence of Glaucoma can be explained mainly by two of them, the Primary Open-Angle Glaucoma (POAG) and Angle-Closure Glaucoma (ACG). Glaucoma prevalence studies usually only feature these types of Glaucoma.

Below an overview of Glaucoma types can be found [SBS+18, Shr16].

- **Primary open-angle Glaucoma (POAG)** - This is the most common type of Glaucoma, usually, when people only mention Glaucoma they are referring to POAG. It is caused by a slow clogging of the drainage canal, that leads to the rise of IOP, and consequently optic nerve damage. It has no early symptoms and continues to progress for many years until a substantial vision loss has been achieved.

- **Angle-closure Glaucoma (ACG)** - This is a painful and abrupt Glaucoma. In this type, the drainage canals are blocked and a sudden rise in IOP occurs. Symptoms like, vomiting, extreme eye, and head torment and obscured vision can occur. Immediate medical treatment is needed otherwise blindness will occur within two days.

- **Normal tension Glaucoma** - This is sometimes called low-tension Glaucoma. In this type, there is no increase in the IOP, but the normal pressure of the eye is already causing damage to the optic nerve head. It usually occurs due to poor blood quality and can be associated with vascular diseases.

- **Congenital Glaucoma** - This is ordinarily present since the moment of birth. It can be the result of incomplete or abnormal development of the eye's drainage canal or other disorders in the eye or body.

- **Secondary Glaucoma** - This type of Glaucoma occurs due to other eye diseases/issues. It can happen in situations like surgeries or inflammations.

2.2   Fundus Images

Fundus images are photographs of the fundus of the eye, this is the interior surface of the eye opposite to the lens; an example of these images can be seen in figure 2.2. These images turned into a very relevant research topic due to their morphological characteristics, since they have the necessary morphological features to diagnose several important eye conditions. One of them is Glaucoma, the topic of this work, although other leading causes of blindness such as age-related macular degeneration (AMD) and diabetic retinopathy (DR), can also be diagnosed recurring to these images. [HKT$^+$18].

Combining the previously stated with the fact that these images are acquired with fundus cameras, that can be handheld and are considerably more affordable when compared to the existing equipment for diagnosing this diseases, makes the use of these images a very compelling option.

2.2.1   Morphology

Fundus images contain several elements, that can be seen in figure 2.2. Firstly a variety of retinal vessels can be easily spotted, these are usually identified as veins or arteries and are useful for the diagnose of various diseases like diabetic retinopathy [HKT$^+$18]. The Macula is a darkly pigmented area in the center of the retina that acts as a sunblock and subdivides into the Fovea that is the darkest region of the fundus image. It consists of a small and slightly concave area without any vessels [TJ18]. Exudates are not featured in figure 2.2, but they are also relevant features commonly used for detecting diabetic retinopathy and consist in bright scattered sections that are formed after the leakage of the retinal vessels [TJ18].

The most relevant structures present in these images for Glaucoma diagnosis are all within a region named as the region of interest (ROI). This region can be divided into four sections, the superior section in the top of the image, inferior section in the bottom of the image, on the left of the image the temporal section and lastly on the right of the ROI there is the nasal section. In the ROI, 3 important elements can be identified as shown in figure 2.3, these elements are the following.

- **Optic Disc** - This is the central yellowish part of the retina, it has an elliptical shape and is an entry point for outgoing vessels into the retina. This area is also a blind spot since it does not contain any rods or cones [TJ18].

- **Optic Cup** - The Optic cup is the bright central depression in the Optic Disc [TJ18].

- **Neuro-Retinal Rim** - This Rim is the region between the edge of the optic cup and the edge of the optic disc [Shr16].

Besides these 3 elements, the state of the retinal nerve fiber layer can also be inferred from Fundus images and it will be further explained in section 2.2.2.
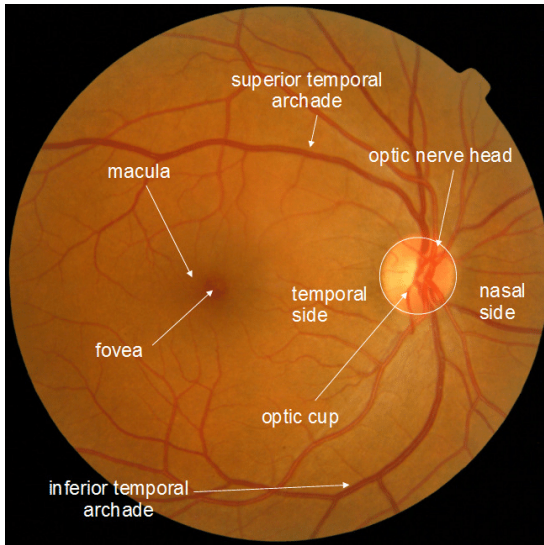
Figure 2.2: Fundus Image [Mor11]
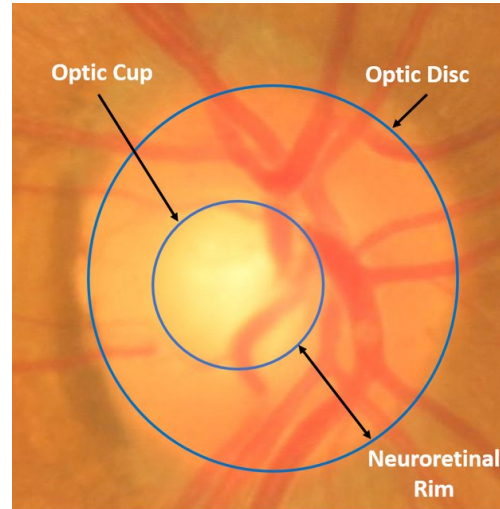
Figure 2.3: Region of interest (ROI) in fundus image [ABWAN+18]

2.2.2 Morphological Features for Glaucoma Detection

Various morphological features can be inferred from these images that are useful for Glaucoma detection. The most commonly used is the cup to disc ratio (CDR) [TLP11]. An overview of these features can be found below.

- **Cup to Disc Ratio (CDR)** - As the name suggests it consists of a ratio between the cup and the disc. It can be calculated based on the horizontal length, vertical length or area of the cup and the disc [TJ18], however, the clinical convention consists in using the vertical lengths. The enlargement of this ratio is known as 'cupping' [HKT+18] and can be classified into mild (CDR up to 0.4), moderate (CDR between 0.5 and 0.7) and severe (CDR above 0.7) Glaucoma classes [SBS+18]. In figure 2.4, an example of a fundus image with a normal CDR (on the left) and an abnormal CDR (on the right) can be seen.
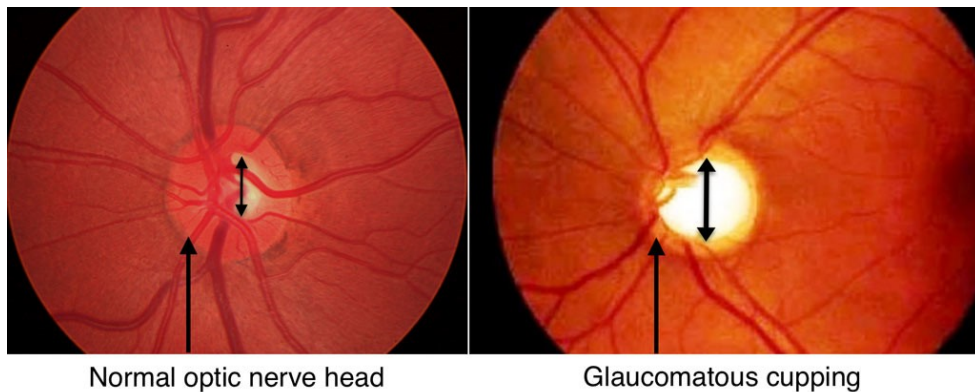
Figure 2.4: Fundus images with and without cupping [Joy17]

- **Disk damage likelihood scale (DDLS)** - This likelihood scale was proposed by [Hen06], and claims to be superior to the CDR ratio when describing Glaucomatous eyes from non-Glaucomatous. The intuition is that the CDR can misclassify large optic nerves as Glaucomatous and small optic nerves as normal. Secondly, focal changes that are usually associated with Glaucoma are not reflected in CDR. To address those issues, this scale uses the radial NRR width and the disc diameter and consists of the following expression:

$$DDLS = \frac{MinNRRWidth}{DiscWidth} \qquad (2.1)$$

  In the expression, *MinNRRWidth* is the radial NRR width at the thinnest point and *DiscWidth* is the diameter of the disc in the same axis. Although some of the CDR issues are solved by the DDLS, this feature also carries some disadvantages, since it is not very capable of characterizing tilted optic nerves and it is not capable of identifying new areas of NRR damage if there is already an NRR area with more damage. The DDLS classifies Glaucoma in 10 stages.

- **NRR and the ISNT rule** - The ISNT rule states that in normal eyes the NRR width in the inferior section (I) of the ROI is greater than in the superior section (S) which in turn is greater than the area in the nasal section (N) which in turn is greater than the area in temporal section (T); this can be translated by the following inequation.

$$I > S > N > T \qquad (2.2)$$

  It has been declared that the compliance of this rule is useful in differentiating normal from Glaucomatous optic nerves and is not affected by race [Sci14].

- **Atrophy of the Retinal Nerve Fiber Layer (RNFL)** - Another Glaucoma flag that can be inferred from fundus images is the atrophy of the RNFL, this translates to thickness variations of the RNFL, more precisely a decrease in RNFL thickness. In fundus images without the red color channel, the RNFL can be observed as a bright striped pattern on the background. Furthermore, this pattern modifies its visual appearance (becoming darker in abnormal cases) in accordance with thickness changes, and this, although not necessarily, can indicate Glaucoma [Jan14]. A comparison between a normal and an abnormal RNFL in a fundus image can be seen in figure 2.5.

- **Peripapillary Atrophy (PPA)** - PPA is one of the risk factors of Glaucoma [Shr16, MHS+11, HKT+18, Jon05], it can be perceived as a bright region in the fundus image that sometimes is confused with the disc by automatic tools [MHS+11]. As stated by [Jon05], PPA can be divided into two zones 2.6, alpha, and beta, both located outside the optic disc with the beta zone being always the closest to the disc. This zones can co-occur or occur in isolation and
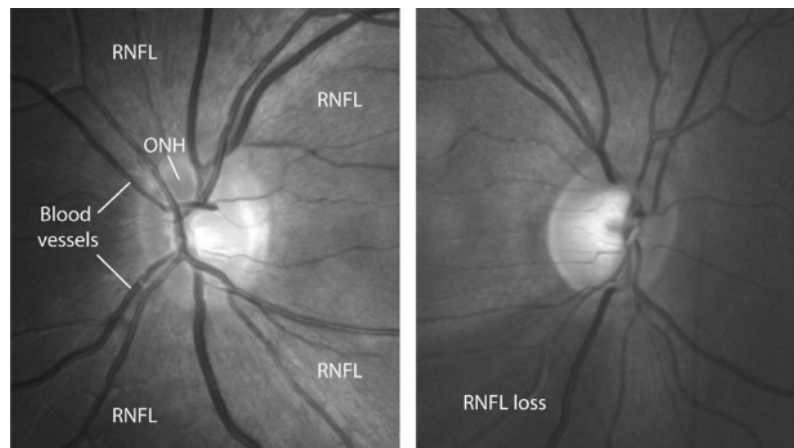
Figure 2.5: Normal fundus image on the left and Glaucomatous fundus image on the right with a visible atrophy of the RNFL translated by the darkening of the image [Jan14]

are considerably bigger in Glaucomatous images. Besides signalizing Glaucoma, PPA is also useful to distinguish between various types of Glaucoma.
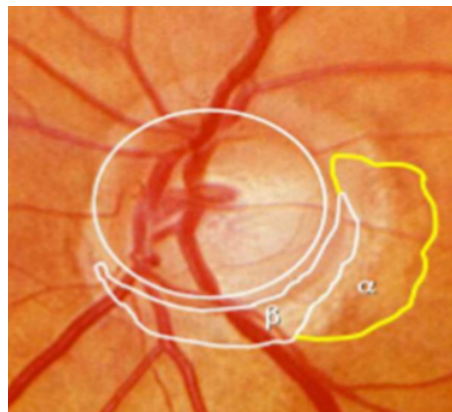
Figure 2.6: Alpha zone and Beta zone in PPA [Shr16]

2.3 Machine Learning

2.3.1 Overview

Machine learning (ML) is one of the most active research fields at the moment, it can be defined as a set of methods to detect patterns in data. These methods can learn by analyzing data and then make new predictions on unseen data. The learning process can be supervised when labeled data is available or unsupervised when the data is unlabeled. Furthermore, when the prediction variable is a continuous value, the training process is called regression yet when it is a discrete variable, it is called classification. [LJC+17]

Background

The most relevant topic for this work is supervised classification, therefore an overview of the most common models used for this task can be found below:

- **Naive Bayes** - These models focus on the probability distribution of the input data, and as the names suggest are based on the application of the Bayes theorem. The name 'Naive' comes from the fact that these models assume that features are conditionally independent of each other in a given class. This is naive because it is almost never true [LJC+17].

- **Support Vector Machines (SVM)** - Support vectors are a distance based classifier that generates a decision hyperplane with the maximum possible margin to the existing samples. The samples on the margin are defined as support vectors [VRN16].

- **k-nearest neighbors (KNN)** - A very simple classifier, as the name suggests a new sample is labeled based on its k-nearest neighbors, with k being a hyperparameter.

- **Random Forests** - Random forests consist in the combination of several random decision trees and are therefore considered an ensemble.

- **Ensembles** - Combination of different models.

Deep Learning (DL) is a branch of machine learning that is rapidly evolving, inspired by the multilayered human cognition system [LJC+17]. Contrary to what might be thought, deep learning techniques are not new, although they only became popular a few years ago. This is due to 3 key elements, the computational power increase, available data increase and algorithm advances in deep learning research.

Deep learning techniques originated from artificial neural networks, and try to learn hierarchical representations of the data with multiple non-linear combinations between layers. They require more data and computational power than machine learning techniques to achieve good performances but have better modeling power. [ZYC+18] One very relevant DL technique for image classification and segmentation is Convolutional neural networks (CNN's) which consist of multiple layers connected by neuron-like connections as in classic neural networks and can be described by three typical layers [LJC+17]. These can be seen in figure 2.7 and are shortly described below.

- **Convolution layers** - These layers detect edges, lines, and other visual components by performming mathematical operations called convolutions, hence the name. After performing a convolution a feature map is generated that is then passed to the next layer.

- **Pooling layers** - Pooling layers are used to reduce the dimensions of the feature maps, the most used approach is max pooling, where the maximum value of a window is selected and the window is replaced by that value.

- **Fully connected layers (FC)** - These layers accommodate all the features gathered from the last layers, connecting to every neuron from the previous layer and provide the final result.

Convolution and pooling layers sequences are usually repeated several times to increase the modeling power of the network. Due to this increase, CNN's are very prone to overfitting, which essentially means that the network can memorize the data instead of generalizing. To account for this, regularization techniques have been proposed. One popular technique is Dropout, it consists of randomly deactivating a group of neurons for each sample during training. Other ways of regularizing the network are data augmentation and the use of pre-trained networks that already have learned meaningful weights and now only need fine-tuning. To conclude, CNN's major advantage is that they do not require feature engineering, so full images can be passed to the network and then the best features will be learned instead of engineered [LJC+17].
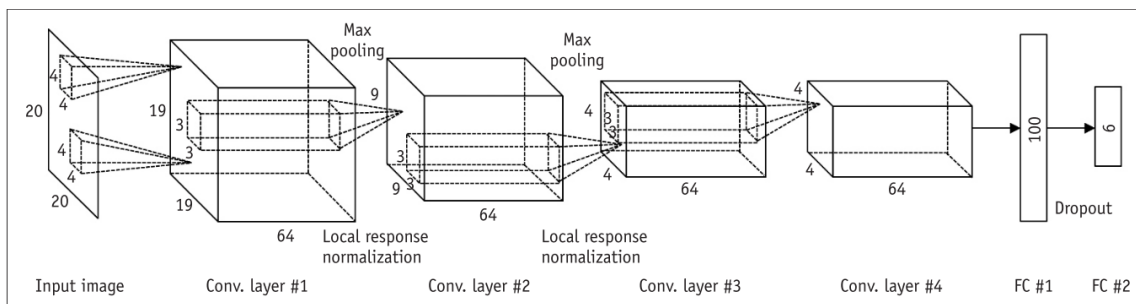
Figure 2.7: Convolution Neural Network example [LJC+17]

2.3.2 Metrics

In order to evaluate these models, several metrics are available [TJ18]. First, it is important to introduce 4 intuitive concepts[1]. The initial two are associated with correct classification, True Positive (TP) is the outcome where the model correctly predicts the positive class. Likewise, True Negative (TN) is the outcome where the model correctly predicts the negative class. Similarly, the other two concepts are related with an incorrect classification from the model, False Positive (FP) is the outcome where the model incorrectly predicts the positive class and False Negative (FN) is the outcome where the model incorrectly predicts the negative class.

Now addressing the most common performance metrics for binary classification, first we have the accuracy, which is the degree of similarity of the predicted results with the ground truth.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{2.3}$$

Sensitivity is the ability of correctly detecting a positive class, it is also called Recall and true positive rate (TPR).

---

[1]https://developers.google.com/machine-learning/crash-course/classification/true-false-positive-negative

$$Sensitivity = \frac{TP}{TP+FN} \tag{2.4}$$

Specificity is the ability of correctly detecting a negative class.

$$Specificity = \frac{TN}{TN+FP} \tag{2.5}$$

Precision is a metric for how precise the model is out of those predicted positive. It is a relevant metric for when the cost of predicting FP is large.

$$Precision = \frac{TP}{TP+FP} \tag{2.6}$$

F1 is a function of precision and sensitivity, it is useful when a balance between these two metrics is desired and the number of TN is not relevant for the problem.

$$F1 = 2 \times \frac{Precision \times Sensitivity}{Precision + Sensitivity} \tag{2.7}$$

False positive rate (FPR), is the rate of predicting the positive class when the outcome is the negative class and can be calculated through the two following equations.

$$FPR = \frac{FP}{FP+TN} \tag{2.8}$$

$$FPR = 1 - Specificity \tag{2.9}$$

Additionally, TPR and FPR can also be used to plot the ROC curve, by plotting the TPR on the y-axis and the FPR on the x-axis from 0 to 1. This curve explains the model capability to distinguish both classes. Finally, the area under the curve (AUC) is an extensively used metric that consists, as the name suggests, on the area under the ROC curve, where a 0.5 value is random guessing and 1.0 represents perfection.

Now changing the subject from supervised classification to supervised segmentation tasks, in here, the most important metric is intersection over union (IoU). It is a simple metric that given the ground truth mask and the prediction mask, can be computed dividing the area of overlap of the two masks by the area of union.

$$IoU = \frac{TP}{TP+FP+FN} \tag{2.10}$$

Finally, when these models are being trained a loss function is required, this is a function calculated between the ground truth and the prediction that acts as an error between the two instances. The result of the loss function is then used to slightly update the model weights in the direction

that will minimize the loss function [LJC$^+$17]. A loss function with value zero indicates a perfect model.

Several loss functions exist, however, the most commonly used for binary classification is cross-entropy loss, also called log loss, it rapidly increases as the model predicted probability diverges from the ground truth and slowly decreases in the opposite case. The equation 2.11 represents the cross entropy loss for the binary classification problem and equation 2.12 generalizes this loss for the multiclass classification.

$$CrossEntropyLoss(binary) = -(ylog(p) + (1-y)log(1-p)) \tag{2.11}$$

In 2.12, $y_{o,c}$ is a binary label with value 1 if the class label c is the correct classification for observation o and 0 otherwise, while $p_{o,c}$ stands for the predicted probability that observation o is of class c.

$$CrossEntropyLoss(multiclass) = -\sum_{c}^{C} y_{o,c}log(p_{o,c}) \tag{2.12}$$

When considering segmentation tasks, an alternative loss to the cross-entropy called Dice-Loss is also widely used. This is based on the Dice coefficient and was first introduced as a loss in [MNA16]. It is a measure of overlap between the ground truth and the prediction that doesn't require class weights when the background and foreground classes are widely imbalanced.

Moreover, in [FCX$^+$18a], a multi-label dice loss was introduced, adding a weight per class (W_c) to the Dice coefficient and summing the results for all classes. In the equation for this loss (2.13), y, p, C and N represent the ground truth, prediction, class and pixel number, respectively.

$$DiceLoss = 1 - \sum_{c}^{C} \frac{2W_c \sum_{i}^{N}(y \cap p)_{c,i}}{\sum_{i}^{N} y_{c,i} + \sum_{i}^{N} p_{c,i}}, \sum_{c}^{C} W_c = 1 \tag{2.13}$$

2.4 CAD Systems in Healthcare

Healthcare professionals usually have to analyze a huge collection of data from different body structures during their work routine which is a tiring and challenging task with a high degree of associated responsibility.

With this mind, Computer-aided diagnosis (CAD) systems have been widely developed during the last decades to help these professionals. These systems, as the name suggests, assist professionals in interpreting medical images and other types of data, providing 'second opinions', allowing a faster patient triage, and sometimes act as end-to-end solutions for an initial diagnose of several conditions.

Therefore, the main goals of these systems consist in reducing the burdens caused by the intense workload of a healthcare professional, reduce errors caused by fatigue, improve diagnostic accuracy, reduce inter-reader and intra-reader variability and of course automate the screening process, allowing massive screenings for various diseases that weren't possible before [FUN$^+$08].

The recent growth in the development of these systems can be justified by the successive improvements in machine learning/deep learning and computer vision fields in combination with the increasing availability of biomedical data. This new data is continuously being obtained by clinical imaging, electronic health records (EHRs), genomes and wearable devices, bringing exceptional opportunities and challenges for the development of new CAD systems [MWW$^+$17].

Furthermore, as in other research topics, it can be noticed that during the last years, deep learning CAD systems are gradually surpassing all of the state-of-the-art systems.

Early diagnosis of Alzheimer disease from brain Magnetic Resonance Imaging (MRI), diagnosis of breast nodules and lesions from ultrasound images, classification of cancer types using gene profiles, and knee cartilage MRIs to predict the risk of osteoarthritis are just a few of the examples of recent CAD systems using deep learning techniques, surveyed by [MWW$^+$17]. Particularly in the field of this work, ophthalmology, CAD systems for predicting cataracts from cross-sectional slit lamp images are being developed and the addressed fundus images are also being widely used to build these systems, not only for Glaucoma but also for other retinal disorders, such as age-related macular degeneration (AMD) and diabetic retinopathy (DR) [GORT18].

2.5 Summary and Conclusions

This chapter aimed at introducing the reader with the main concepts of this work, connecting these different topics between each other.

It can be understood that Glaucoma is a very relevant group of eye conditions that damage the optic nerve head, furthermore, fundus images contain the necessary morphological features to detect Glaucoma or at least aid the detection, and can be acquired in relatively affordable ways.

Additionally, the continuous growth in Artificial intelligence, namely machine learning/deep learning techniques is allowing the development of new and better CAD systems with biomedical data, such as fundus images.

In conclusion, CAD systems can allow faster and improved diagnosis of several conditions, namely Glaucoma.

Chapter 3

Glaucoma CAD Systems Based on Fundus Images

In this chapter, the state-of-the-art techniques and resources for Glaucoma detection with fundus images are exposed. The chapter starts with an overview of the fundamental technologies for the development of CAD systems 3.1. Moreover, section 3.2, describes the typical pipeline used when building CAD systems for Glaucoma detection, with an overview of the main techniques used in each step by the current approaches. Then, in section 3.3, several implementations of these CAD systems are reviewed and compared, additionally, some of the current implementations of these systems have already turned into commercial solutions and in section 3.4, an overview of the main ones is provided. Finally, on section 3.5, the main conclusions of this chapter are drawn and the development opportunity is briefly introduced.

3.1 Technologies

Several frameworks/libraries exist to aid the developers in the development of these systems. The main ones serve as abstractions of several computer vision and machine learning/deep learning techniques. The most relevant technologies are briefly introduced below.

3.1.1 OpenCV

OpenCV[1] is one of the most, if not the most popular open-source computer vision library, providing several modules to perform image operations, being therefore widely used to process and augment fundus images.

---

[1] https://opencv.org/

3.1.2 imgaug

The imgaug[2] library is an open source library for image augmentation. It uses OpenCV as its backbone and acts as a simple interface to build complex augmentation pipelines to augment images, key-points, bounding boxes, heat maps, and segmentation maps.

3.1.3 Scikit-Learn

Scikit-Learn[3] is a machine learning library in python, providing a very simple API for tasks such as classification, regression, clustering, dimensionality reduction, model selection and preprocessing. This library covers the basics of machine learning and is therefore very limited when it comes to deep learning.

3.1.4 TensorFlow

Tensorflow[4] is a more robust and complete machine learning library that has a slightly higher learning curve, but is a lot more powerful when compared to Scikit-Learn and widely used for deep learning. Furthermore, it has an extension called TensorFlow Lite[5] that allows the integration of developed models in mobile and embedded devices with low latency. Tensorflow Lite can be very useful for the development of a more affordable mobile-based CAD system for Glaucoma detection.

3.2 Typical Pipeline

The development of conventional machine learning based systems follows a common pipeline, and CAD systems for Glaucoma detection are no exception. This pipeline is expressed in figure 3.1 and explains the path between a fundus image and the final classification. In the following sections, every step of this procedure will be addressed in isolation, shortly explaining in what it consists and providing an overview of the most common techniques used in that section, considering the existing CAD systems for Glaucoma detection.

Fundus Image → Pre-Processing → Segmentation → Feature Extraction → Feature Selection → Classification

Figure 3.1: Typical pipeline followed by CAD systems for Glaucoma detection

---

[2] https://imgaug.readthedocs.io/en/latest/
[3] https://scikit-learn.org/stable/
[4] https://www.tensorflow.org/
[5] https://www.tensorflow.org/lite

3.2.1 Pre-processing

Fundus images often contain noise or artifacts that may affect or hinder the subsequent steps. Therefore they are usually pre-processed to enhance the desired structures for Glaucoma detection and as a way of regularizing and eradicating undesired noise from them [VRN16].

The main approaches consist in the normalization of the RGB values, Contrast limited adaptive histogram equalization (CLAHE) to improve the contrast [HKT$^+$18], blood vessels removal, illumination correction [VRN16] and conversion from the RGB to the HSV plane [SGA16]. The transformation of the fundus images from the Cartesian coordinate system to the polar coordinate system is also a pre-process approach that has shown good results in recent applications, since it can enlarge the cup region, balancing the cup proportion in the image which facilitates its segmentation [FCX$^+$18a]. Additionally, extracting the ROI from all the images by cropping them down is also helpful since as mentioned in section 2.2 this region is the important section of the fundus images when considering Glaucoma assessment.

Finally, mostly when considering deep learning approaches, resizing and data augmentation is also performed to increase image heterogeneity [LHK$^+$18].

3.2.2 Segmentation

Segmentation is a vast research topic, which also assumes great importance in some Glaucoma CAD systems, being sometimes the most demanding component of the pipeline.

The most desired structures to segment in these images are the optic disc and optic cup since they allow the calculation of relevant morphological features for Glaucoma screening [FCX$^+$18a].

The main techniques to achieve segmentation on fundus images where surveyed by [ABRL15] and [TJ18]. Firstly, there is the level-set approach that uses the concept of curves and surfaces for segmentation. Furthermore, thresholding, as the name suggests, simply uses a threshold and can be based on color, contrast, intensity values of the pixels and others. Then, tactics with a focus on the shapes and boundaries of the region that will be segmented are characterized as active shape modeling approaches and active contour modeling approaches, respectively. Moreover, when similar pixels are grouped together in specified ranges they form clusters that can be interpreted as segments, this is therefore called clustering based segmentation. Then there is also the component-based approach, that extracts features from the image based on image components such as intensity pixels and colors. Besides all this, U-Net, a deep learning technique that consists in a U-shape convolution neural network is now being used and achieving very good performances [LHK$^+$18, FCX$^+$18b].

Finally, all these methods can be combined, creating a hybrid segmentation approach.

3.2.3 Feature Extraction

First of all, it is important to notice that this step of the pipeline mainly concerns machine learning techniques, since the most used deep learning techniques, such as convolutional neural networks do not require feature extraction as addressed in section 2.3.

The previous step stated the main techniques used to perform segmentation in fundus images, and in this step, Glaucoma features are extracted/calculated recurring to these segmentations and the initial fundus image.

The morphological features addressed in 2.2.2 are some of the commonly extracted features but there are several other non-morphological features that are usually extracted from fundus images for Glaucoma detection, as surveyed by [HKT+18].

Firstly, there is wavelet decomposition, which consists of the decomposition of the image into coefficients. One type of wavelet decomposition is Discrete Wavelet Transform (DWT), which is a multi-scale analysis method where the analysis can be performed on various levels and scales [SDP+16], another type of wavelet decomposition is Empirical Wavelet Transform (EWT), this, unlike DWT, is a signal dependent decomposition technique, furthermore the EWT decomposes the signal into elements of different frequency bands [MPA17].

Another feature extraction technique is image descriptors, these describe visual features from the fundus images, one example is the GIST descriptor.

Then there are non-linear feature extraction methods, which are widely used to detect pixel variations in fundus images, an example being Correntropy, a non-linear kernel based measure of similarity [MPA17].

Moreover, multiple textural features can be extracted, such as textons, which are the most basic microstructures in normal images [ABK+17].

Finally, transformation feature extraction techniques like Gabor transform, modified census transformation (MCT) and radon transform (RT) [HKT+18] are also used.

3.2.4 Feature Selection

When feature extraction is accomplished, feature selection is usually also performed, in order to select the best subset of features to perform the classification, thus reducing dimensionality issues.

Common methods for feature selection include principal component analysis, linear discriminant analysis, independent component analysis, locality sensitive discriminant analysis, and marginal fisher analysis [HKT+18]. Additionally, the t-test algorithm that ranks features based on their t-value [MPA17] and genetic algorithms [SDP+16] are also used in these systems.

3.2.5 Classification

After gathering the best representation of the fundus images, several classification techniques are commonly used.

The most simple is to determine Glaucoma based on some kind of threshold, for example, just verifying if the CDR is above some pre-determined value [SGA16].

The most common approaches, however, use machine learning and deep learning classifiers. As surveyed by several authors [HKT+18, VRN16, SBS+18, TJ18], the most widely used classifiers until recently were SVM, KNN, Naive Bayes and Random Forests, with the SVM being the most used and usually best-performing technique. Nowadays this is rapidly changing and is

not the truth anymore. A number of approaches using deep learning classifiers like CNNs and pre-trained CNNs have now been proposed, and are obtaining increasingly better performances.

3.3 Current Implementations

The first relevant implementations of these systems started to arise in 2008 [HKT$^+$18], and since then a significant collection of approaches have been proposed.

In 2016, A. Singh et al.[SDP$^+$16] achieved a state-of-the-art performance at the time with 94.7% accuracy. They started by identifying the center of the disc, performing the disc segmentation afterward. Then, the intuition that blood vessels represent noisy pixels that affect the performance of the system led to the removal of these vessels from the obtained optic disk images. After this, feature extraction was accomplished using first level discrete wavelet decomposition that resulted in a feature vector of 18 features. For feature selection two approaches were tried, genetic algorithms and PCA, and the final feature vectors of each technique were then tested with several classifiers (SVM, KNN, Random Forest, Naive Bayes and ANN). Since the fundus images are from a local dataset of 63 images, which is rather small, leave-one-out cross-validation was also performed to account for overfitting issues. The best performing models were the SVM and KNN, with PCA feature selection that resulted in only 2 components. Both of these classifiers obtained the 94.7% accuracy in the described dataset.

Moreover, Sanjivani Shantaiya et al. [SGA16] used morphological features for Glaucoma detection, namely the CDR and NRR ratio in the ISNT quadrants. The images were pre-processed extracting the ROI, the green plane and the V-plane, after converting the images to HSV. Then the green plane was converted into grayscale and threshold segmentation using pixel intensity was performed to segment the cup. Regarding the segmentation of the disc, this was accomplished with both the V-plane and the green plane using threshold segmentation as well. Then the NRR was obtained simply by performing an AND operation between the disc and the cup. The areas of these structures were collected by counting the white pixels in the segmented map, in order to calculate the CDR and NRR area in the ISNT quadrants. Finally, the classification was obtained directly, detecting Glaucoma if the CDR is above 0.3 and the ISNT rule is violated. An accuracy of 98.8% was obtained on a rather small dataset.

Moving now to approaches published in the last year (2018), where the trend in the application of deep learning techniques to these systems is clearly noticed.

U.Raghavendra et al. [RFB$^+$18] proposed what they claim to be the first fully automated CNN architecture for Glaucoma detection. The architecture consists of 18 layers, starting with a convolution layer, followed by batch normalization, then Relu and finally max-pooling. This is repeated for four times with different filter sizes, forming the first 16 layers. To finalize the architecture a fully connected layer and a soft-max layer for classification are added. Regarding the optimization of the network, stochastic gradient descent momentum was used in combination with several learning rates. The private dataset used contained a total of 1426 fundus images (589 normal images and 837 Glaucoma images), and all the images were resized to 64x64 before being

fed to the network. The highest accuracy of 98.13% was obtained with a learning rate of 0.001 and without any pre-processing on the images.

Furthermore, a new segmentation approach of the optic disc was published by H.Fu et al. [FCX+18a]. The novelty of this work consists of the joint segmentation of the optic disc and optic cup since the majority of the approaches achieve this segmentation separately. To deal with this joint segmentation the work also proposes a new joint loss metric. The first step of this method is the transformation of the fundus images to the polar coordinate system, in order to enlarge the cup, after this the segmentation is performed with a U-Net, this consists of four key parts, the multi-scale input layer, the U-shape CNN a side output layer and the mentioned multilabel loss metric. Several tests were performed in different conditions and it was concluded that the polar transformation attains a relevant performance gain. Glaucoma screening was then performed calculating the CDR value and an AUC of 0.8997 in the private SCES dataset was obtained. The code for this work is available online [6].

A few months later, the same authors published another work [FCX+18b], iterating over the existing one. It consists of an ensemble model that combines four models into a final classification. First, the complete fundus image is classified with a CNN, then a U-Net is used to segment the disc in the fundus image, and from this U-Net other classification is obtained by attaching fully-connected layers after the encoding path of the U-Net. Having segmented the disc, this is then used to crop the initial fundus image ending with an image of the ROI. This image is then transformed to the polar coordinate system and both images are used in separate CNNs to obtain the remaining two classifications. All the previously mentioned CNNs follow the ResNet-50 architecture and are initialized with the weights of the ImageNet competition, being, therefore, pre-trained CNNs. This approach increased by almost 2% the result of the last method, obtaining an AUC of 0.9183 in the SCES dataset. The code for this work is also available online [7].

An even more recent approach (from August 2018) by Z.li et al. [LHK+18], also uses a CNN. Instead of developing a CNN from scratch, similar to the previous, this approach uses an existing architecture, namely the Inception-v3 architecture [SVI+15] that has already proven its success in other tasks. The images were resized to 229x229 and the RGB values were normalized between 0 and 1 before feeding the images to the network. A minibatch gradient descent of 32 was used in combination with the ADAM optimizer and a 0.002 learning rate. But the key aspect of this approach is its dataset since a total of 70000 fundus images were downloaded from an unlabeled dataset[8] and 48116 of those images were selected after verifying that they include a visible disc. After this, 21 trained ophthalmologists performed the classification of these images in an online platform and multiple interpretations of each image were obtained for quality control. The result was 39745 images, where 31745 were used as a training set and 8000 as a validation set. The system achieved an AUC of 0.986 with 95.6% sensitivity and 92% specificity, which demonstrates the relevance of the dataset size and provides further evidence about the success of deep learning

---

[6] https://github.com/HzFu/MNet_DeepCDR
[7] https://github.com/HzFu/DENet_GlaucomaScreen
[8] http://www.labelme.org/

techniques for Glaucoma detection.

3.4 Commercial Solutions

In the last couple of months, some implementations of the trending deep learning based CAD systems for Glaucoma diagnosis have already reached the product state and are now available as commercial solutions.

An example is Yanbao [ZFZ$^+$18], it is a mobile application already available in popular marketplaces [9], that allows the user to upload a fundus image to a server where 4 types of feedback are generated, CDR analysis, NRR analysis, glaucoma confidence level, and doctor diagnosis. The first three topics are reportedly returned in about 10 seconds and the doctor analysis is dependent on a professional expert that will interpret the images. In order to provide this feedback, the optic disc is first localized and used to crop the full fundus image and obtain the ROI image, then, joint segmentation of the optic disc and cup are performed with an enhanced U-net, trained and tested on the ORIGA dataset. These segmentations are then used to perform the CDR and NRR analysis by calculating several morphological features, afterwise, feature selection is accomplished and the selected features are used by an SVM classifier that outputs the glaucoma confidence level. The performances are then evaluated once again on the ORIGA dataset.

Another commercial approach that is gaining traction in Glaucoma screening is Retinalyze Glaucoma[10]. This is an algorithm within the Retinalyze[11] product spectrum, that has been released on the 1st of May 2018 and already performed 60.000 examinations. It computes the optic nerve head hemoglobin levels by analyzing the image. This is accomplished because regions of the optic nerve head with high hemoglobin levels mainly reflect red light while blue and green light is usually more reflected in regions with low hemoglobin. Then it uses optic nerve hemoglobin (ONH Hb) across the vertical disc diameter (8&20 Hb), estimated cup-disc ratio (C/D) and the glaucoma discriminant function (GDF), which combines the slope of Hb amount with the mean in 8&20 Hb. It was tested and compared with Optical Coherence Tomography (OCT) and confocal tomography (HRT III) and the authors show similar diagnosis capabilities [MHRUG$^+$16]. Furthermore, the algorithm is under active development and in recent presentations[12] deep learning techniques are applied for segmentation and classification tasks to identify the differences in the hemoglobin distribution in glaucomatous and non-glaucomatous eyes achieving better results than some well established clinical practices.

---

[9]http://url.cn/57tk9jT
[10]https://www.retinalyze.com/post/retinalyze-glaucoma-a-revolution-in-glaucoma-screening
[11]https://www.retinalyze.com
[12]https://www.retinalyze.com/post/prof-gonzalez-de-la-rosa-presents-new-groundbreaking-performance-of-the-laguna-onhe-retinalyze-glaucoma-algorithm

3.5   Summary and Conclusions

Several conclusions can be drawn from this literature review. First, a considerable variety of implementations have been and continue to be proposed, indicating the relevance in developing a CAD system for Glaucoma diagnosis, furthermore, the proposed techniques are rather different between each other.

Comparing the existing techniques directly it's also a rather challenging task since most of the approaches use their own private datasets.

Despite the aforementioned, it has been clearly acknowledged by the research community that deep learning approaches are capable of obtaining the best results and are evolving very quickly.

Finally, even knowing that these systems are constantly being improved and published, there are no found reports of offline mobile-based CAD systems, neither classification metrics of the existing systems on a mobile environment. Therefore, this is identified as a research and development opportunity.

Chapter 4

Problem Definition and Proposed Solution

The goal of this chapter is to precisely describe what are the motivations, goals, and context leading to the development of this work, as well as the proposed solution to solve the identified issues. At first, section 4.1, discloses the problem by exposing the relevance of an early Glaucoma diagnosis and the weaknesses of existing methods to perform this diagnostic. This powers the reasoning that culminates in the definition of the problem statement. Secondly, section 4.2, provides a detailed overview of the solution that will be developed, identifying the main requirements it should accomplish. Lastly, section 4.3 wraps this chapter, summarizing and pinpointing the main conclusions that can be derived from it.

4.1 Problem

In order to accurately describe the problem that this dissertation addresses, it is useful to further understand the context that surrounds Glaucoma. The following subsections will, therefore, aid the reader in that direction, by explaining prevalence, burdens and current limitations in the diagnosis of this group of diseases, culminating with a compact problem statement.

4.1.1 Glaucoma Prevalence

The representative majority of Glaucoma prevalence studies primarily targets, angle-closure Glaucoma (ACG) and primary open angle Glaucoma (POAG) since these are the most prevalent types of Glaucoma and therefore representative of the total Glaucoma prevalence.

In these studies [QB06, BTF+16, KCF+16], a correlation between age, ethnicity, gender and Glaucoma prevalence is usually made. Throughout the time it is possible to see changes in the gender and ethnic groups that are most affected by Glaucoma, but one constant factor is the age. Older populations are undoubtedly more affected by Glaucoma than younger populations.

The overall conclusion of these studies is that Glaucoma prevalence is continuously growing. Furthermore, a total of 80 million people is expected to have Glaucoma in 2020 [QB06]. The continuous grow in prevalence can be intuitively associated with the aging of the population.

4.1.2 Glaucoma Burdens and Challenges

Glaucoma burdens can be divided into both economic and individual burdens. Individual burdens of Glaucoma largely impact the patient quality of life, these impacts consist in the fear of blindness, depression, social withdrawal, the inability to perform certain activities as the disease progresses, inconvenience, cost of the treatment and others. The impact in the patient quality of life can occur even if he is unaware that he has the disease, and expand linearly with the amount of vision loss. Despite this, these individual burdens can be attenuated if Glaucoma starts to be treated as soon as possible [VLGK11, BAA16].

Now addressing the economic burdens, a 2006 study [LWD+06] performed in the United States classified Glaucoma in 6 stages of increasing severity, going from the earliest Glaucoma (stage 0) to blindness (stage 5). The study concluded that Glaucoma management costs increase with the severity of the disease as can be seen in figure 4.1. Although the study has been performed some time ago, and the values may now differ, this trend still holds and can also be observed in Europe [VLGK11].

One of the main challenges in Glaucoma management is, therefore, its early detection. This is of utmost importance since it can reduce both individual and economic problems. But the current methods to diagnose Glaucoma are both expensive, bulky and require trained professionals as mentioned in 2.1, increasing the demand for new ways to detect Glaucoma.
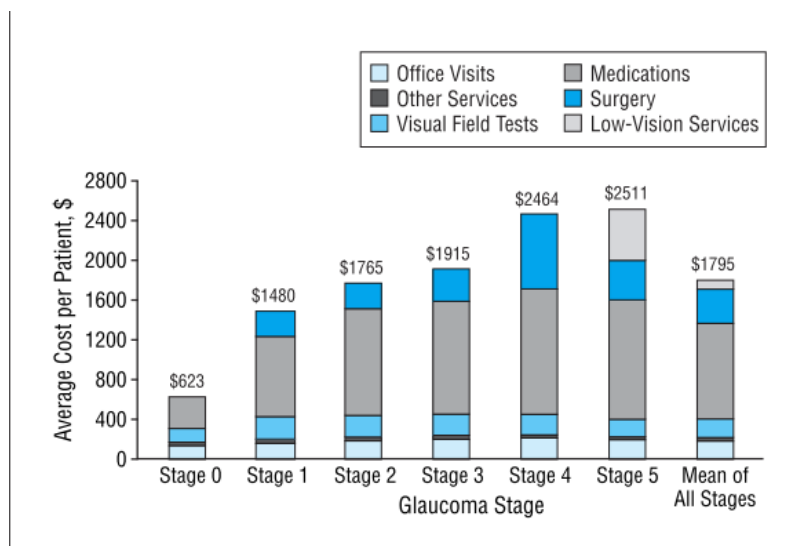
Figure 4.1: Economic burdens of Glaucoma in different stages (values correspond to direct annual costs). Copyright © by American Medical Association. All rights reserved. [LWD+06]

4.1.3   Issues with Current Glaucoma CAD Systems

CAD systems for Glaucoma detection are the most promising approach to increase the chance of an early Glaucoma diagnosis, considering that they can automate part of the Glaucoma screening process and reduce misclassified cases.

The research community has acknowledged the aforementioned and several CAD systems have been developed as addressed in more detail on the previous chapter 3.

These approaches are evolving quickly but there are still some issues/limitations with them. At our knowledge, it does not yet exist a complete standalone CAD system capable of acquiring fundus images and classifying them as Glaucomatous or not. Moreover, the current systems run on expensive computer hardware and are only capable of classifying high-quality fundus images. Finally, classification performance can still be improved.

4.1.4   Problem Statement

Considering the 3 preceding sections, it is clear that Glaucoma is a major health concern, it progressively affects more people economic and personal life, becoming more relevant as time goes by. To address this issue a solution to diagnose Glaucoma as early as possible is needed, CAD systems seem to be the answer but the current approaches still have relevant limitations.

4.2   Proposed Solution

The proposed solution aims at filling the gap in the Glaucoma diagnosis spectrum, by iterating over existing techniques for Glaucoma detection using fundus images.

Therefore the proposed solution consists in the development of an end-to-end CAD system that can be used to help professionals quickly diagnose patients and ultimately by non-experts in big Glaucoma screenings.

To achieve this, deep learning techniques will be used, following the main steps of the pipeline addressed in section 3.2 to produce an algorithm that is able to run remotely in an (Android) smartphone. The algorithm should be capable of processing and classifying fundus images as Glaucomatous or not with high efficiency, meaning that it should be fast during evaluation time on a mobile device, have good accuracy and high sensitivity. This is important to reduce, as much as possible, the number of false negatives detected since false negatives consist in a fundus image identified as non-Glaucomatous that was indeed Glaucomatous and that is the worst possible outcome.

Furthermore, despite the use of deep learning techniques, the developed solution should not act as a black box, by just receiving an image and outputting a Glaucoma confidence level. The developed solution should have a good degree of interpretability, providing an explanation of how the final decision was computed.

After the development of this algorithm, it will be integrated into the Fraunhofer AICOS Eye-FundusScope [1] system. This system consists of an Android application and a custom optical prototype that can be attached to an Android smartphone, and acquire fundus images. The acquired images by the prototype have lower quality when compared to the ones acquired by dedicated fundus cameras, however, new versions of the prototype are continuously being developed and the lower quality of this images might not be an issue in the near future. The developed solution, should be robust to high-quality fundus images and ultimately try to cope with the current quality restrictions of the images acquired with the EFS prototype.

If successful, besides turning the Fraunhofer AICOS EyeFundusScope system into a more complete screening tool while keeping the system low production costs, this work will fulfill its ultimate goal, develop a highly usable, efficient and affordable system for Glaucoma detection.

This will allow cheap and extensive Glaucoma screenings, resulting in the detection of this group of diseases at an earlier stage. Thus reducing Glaucoma burdens through adequate treatment, that will promptly follow the diagnose.

4.3 Summary and Conclusions

Glaucoma is a major health concern, with a growing prevalence mainly due to the aging of the population. This groups of diseases pose several limitations to the affected individual, but an early diagnosis can greatly decrease these limitations. Several CAD systems have been proposed to address this problem, but there are still significant limitations in the available solutions. To address this problem a new highly usable, efficient and affordable CAD system capable of detecting Glaucoma is desired. The solution proposed by this work aims at providing that solution by developing a mobile-based end-to-end CAD system, that, if successful, will be a very relevant contribution to boost Glaucoma management.

---

[1] https://www.fraunhofer.pt/content/dam/portugal/en/documents/Flyers/
EyeFundusScope%20-%20FhP-AICOS%20Project%20Brochure.pdf

Chapter 5

Data Gathering

The present chapter brings to light the foundation of the proposed solution, the data. In section 5.1, publicly available fundus images datasets are addressed. Then in section 5.2 the details of a private dataset developed in the context of this dissertation are unveiled.

Furthermore, the insights of the previous sections are merged and in section 5.3, the constructed datasets used to power this work are disclosed. To wrap everything up, section 5.4, presents the main conclusions of the chapter.

5.1 Fundus Image Datasets for Glaucoma Detection

For the development of CAD systems, properly annotated data is of ultimate importance, since the success of the machine learning/deep learning techniques used in the construction of these systems is greatly correlated with dataset size and quality.

These datasets are not easy to assemble, since they require big collections of representative fundus images, meaning that there should be a relatively good distribution of both classes (Glaucoma and no-Glaucoma), and, the annotation of these images needs to be made by trained professionals.

Luckily, there are various fundus images datasets that were made publicly available for research and educational purposes. Below an overview of publicly available datasets that can be used for Glaucoma detection is presented:

- **ORIGA-light** - The ORIGA-light dataset was initially published in 2010 [ZYL$^+$10], and remains one of the most used datasets by the research community, mainly because it has a satisfactory size and is very well annotated for Glaucoma diagnosis, which is not the case for the majority of the publicly available datasets. This dataset consists of 650 images, with annotations of the eye side, CDR, ISNT rule, RNFL, PPA, Notch in the NRR, Disc Hemorrhage, optic disc and cup boundaries, and a binary Glaucoma label. Despite being a public dataset, getting access to the ORIGA dataset is not easy since it is difficult to contact

the authors and to the best of my knowledge the dataset is not freely available on the web. Nonetheless, the dataset was acquired through the author of a recent work [SHO18], where the dataset was used. In this version, not all dataset annotations were present, counting just with the CDR, eye side, optic disc and cup boundaries, and Glaucoma label.

- **iChallenge-GON** - This dataset was made publicly available in a context of an online competition[1] with an on-site part that took place at the MICCAI 2018 conference. The dataset consists of a total of 1200 images equally split in 3 datasets (training, offline validation and on-site test) that were stratified to keep the same Glaucoma percentage in all of them. From the 1200 images, 120 are Glaucomatous and the test labels are not publicly available since this is a competition, resulting in a total of 800 publicly available images. Annotations for the disc, cup, fovea and a Glaucoma label are provided.

- **Drishti-GS** - The Drishti-GS dataset[2] [SKD+14] consists in 101 images. The ground truth of this dataset includes the cup to disc ratio, a glaucoma label, and a segmentation map for both the optic disc and cup.

- **RIM-ONE** - The RIM-ONE dataset[3] counts with 3 different releases, published from 2011 to 2015 that are essentially three different datasets.

    RIM-ONE r1 - The first release [FAS+11], published in 2011, contains 169 ROI fundus images. The images are classified into four different glaucoma levels (none, early, moderate and deep) which is not common, since the majority of the datasets only classify Glaucoma as present or not present. Additionally, some of the images are classified in a fifth category, ocular hypertension. Disc boundary annotations are also provided.

    RIM-ONE r2 - The second release consists of 455 ROI fundus images, containing annotations of the optic disc boundary and a label indicating the presence of Glaucoma in each fundus image.

    RIM-ONE r3 - The third and last release [FdlR15] was published in 2015, containing 159 stereo fundus images, with annotations of the cup and disc by two experts and two labels, healthy fundus image and Glaucoma/Suspect fundus image. This dataset consists of stereo images, two images of the same eye taken from two slightly different angles, enabling the recovery of part of the 3D perspective which provides some depth information, leading to easier and more accurate annotations of the optic disc and cup structures.

- **RIGA** - The RIGA (Retinal images for Glaucoma analysis) dataset[4] [ASA+17] was released in 2018 and consists of 750 fundus images. These images were derived from three different sources, the Messidor dataset (460 images), Bin Rushed Ophthalmic Center (195 images) and Magrabi Eye Center (95 images), the images were then manually annotated by six

---

[1]http://ai.baidu.com/broad/introduction?dataset=gno
[2]http://cvit.iiit.ac.in/projects/mip/drishti-gs/mip-dataset2/Home.php
[3]http://medimrg.webs.ull.es/research/downloads/
[4]https://deepblue.lib.umich.edu/data/concern/data_sets/3b591905z

ophthalmologists. The annotations consist in the optic disc and cup boundaries, an important highlight is that this dataset does not contain labels to classify each image as Glaucomatous or not and it is therefore only useful for segmentation tasks.

In figure 5.1 sample images of each dataset can be found.
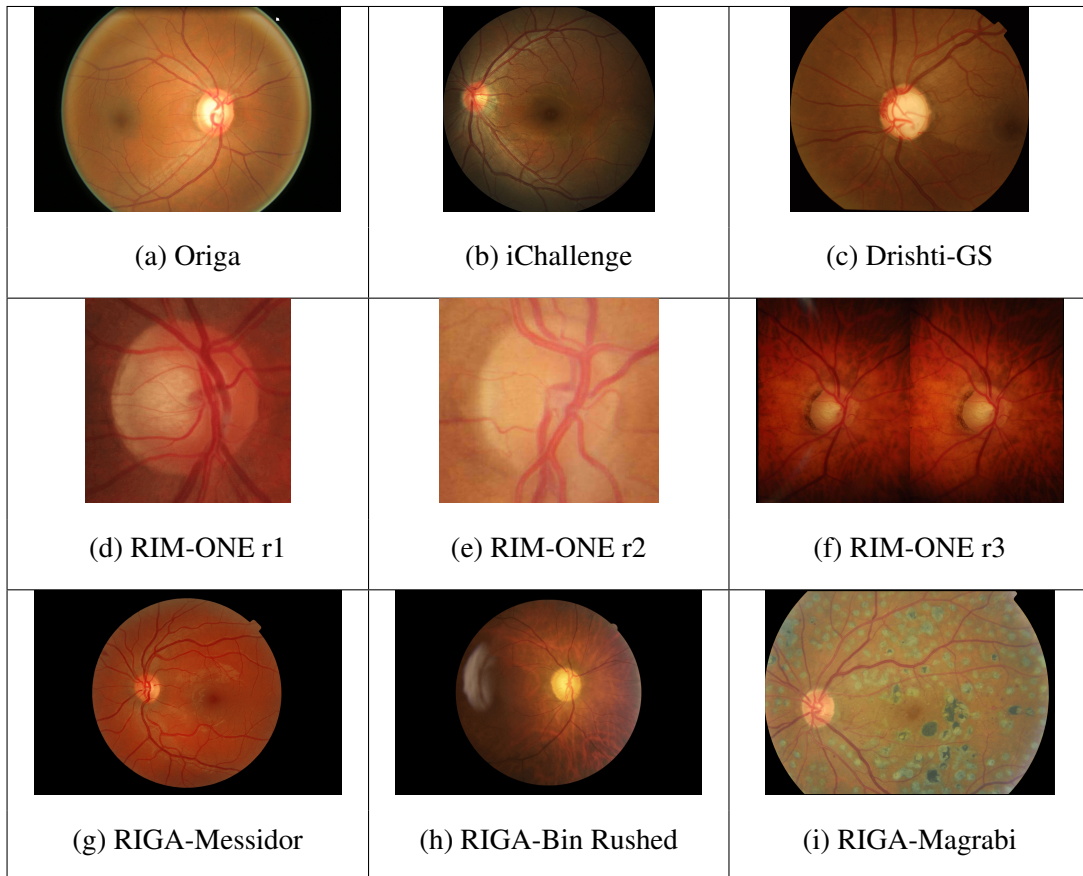
Figure 5.1: Sample fundus images from public datasets

5.2 SCAS Dataset

Since this dissertation is associated with the EyeFundusScope system developed by Fraunhofer, the need for the creation of a dataset acquired with the prototype emerged. In order to suppress this need, a dataset was created during the 'Semana das ciências aplicadas à saúde' (SCAS) at Coimbra's Alma Shopping, along 4 full days. This dataset was created with the cooperation of David Azevedo, a Fraunhofer collaborator that helped to perform the on-site acquisitions.

In the next sections a more detailed explanation of the EyeFundusScope system, how it was used and the main characteristics of this dataset can be found.

5.2.1 EyeFundusScope System

The EyeFundusScope system (figure 5.3) was initially built to explore Diabetic Retinopathy and consists of a smartphone coupled to a 3D printed optical device and an android mobile application that manages the acquisition process.

In fact, there are two main applications used with the prototype, an official application, and a test application. The official application is targeted at the general public and currently functions as a CAD system for Diabetic Retinopathy, however, since it is a more stable and controlled version it doesn't allow full customization of the camera and prototype parameters. That's where the test application comes in, it is a simple and straightforward application with the essential functionality to perform an acquisition with the prototype that allows the full customization of the parameters.

Considering that the conditions of the environment where the acquisitions were going to be performed were not priorly known, the test application was selected to construct this dataset, allowing an on-site adaption and customization of the camera and prototype parameters. These parameters encompass exposure, focus, white balance related parameters as well as parameters deeply related to the prototype workflow that cannot be disclosed. They are very relevant to ensure that the collected fundus images are properly illuminated, with low amounts of noise, good contrast and small or no reflections.

To better suit the test application to the acquisition environment needs, the ability to edit and save patient information, namely the id and the eye that was being acquired, was added to the test application in the context of this work. In figure 5.2, the main screens of the test application are illustrated.
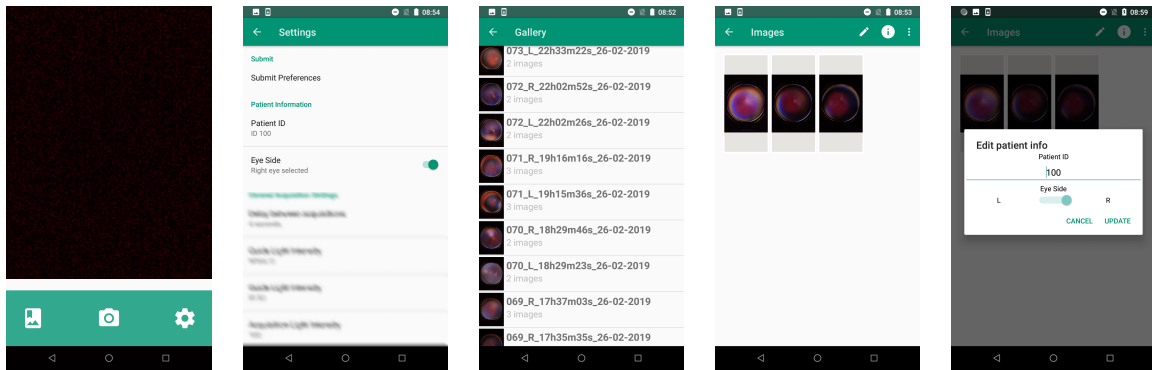
Figure 5.2: EyeFundusScope improved test application

5.2.2 Acquisition

In order to facilitate the acquisition process, the patient pupil should be as dilated as possible. The pupil can be intuitively seen as the window to the retina, so a larger window means an easier alignment of the prototype for each patient and consequently a better image. Since the prototype is intended to be used in large screening programs with restricted resources, eye drops for pupil

Figure 5.3: EyeFundusScope prototype

dilation should not be used, therefore the patient was brought to a relatively dark room (figure 5.4), where he stayed for a couple of minutes, allowing his pupil to dilate.

During this time a detailed explanation of the acquisition process was given to the patient and the informed consent (annex A) was signed.

After that the patient was told to look at a bright fixed spot on the wall (red circle in figure 5.4), making it easier to acquire an optic disc centered fundus image.
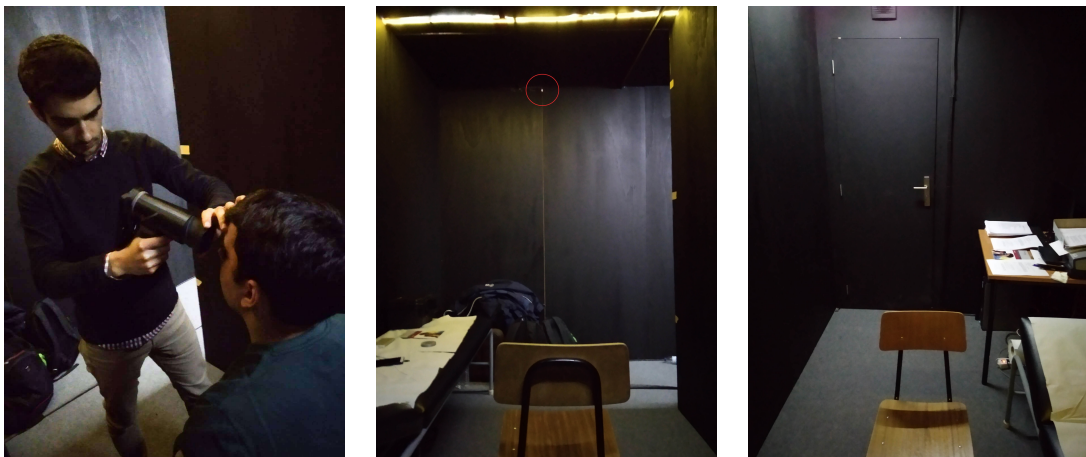
Figure 5.4: Acquisition Environment

Performing the acquisition on both eyes took approximately 3-5 minutes for each patient and 2-3 photos were taken to each eye.

5.2.3 Dataset Overview

A total of 100 patients participated in this study, the majority of them above 40 years old, in order to maximize the possibility of finding positive glaucoma cases. The result was a collection of 525 images, however, some of them were invalid, due to:

- Images acquired by accident (i.e: images of the ground, hands, and others);

- Images with very large reflections, turning the majority of the image into white;

- Images without the optic disc.

After the removal of all the images in one or multiple of the above categories, a total of 434 images were selected and in the context of this dissertation, each image was considered as an individual case. Some samples of the SCAS dataset can be observed in figure 5.5.
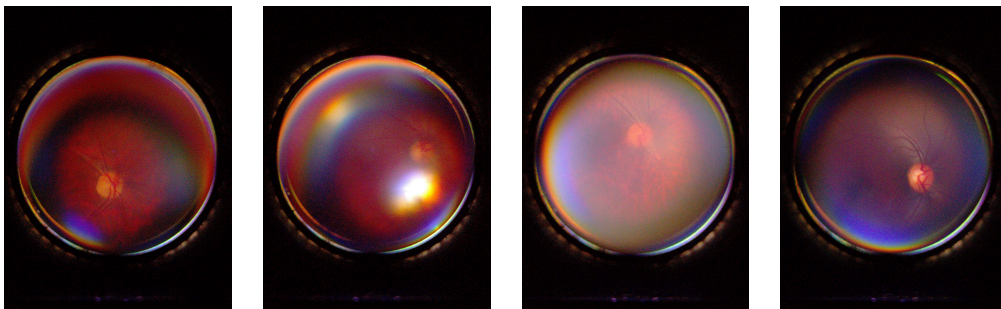
Figure 5.5: SCAS image samples

5.2.3.1 Issues

Performing the acquisition was not trivial, due to the handheld nature of the prototype and the inferior camera sensor present in the smartphone when compared to the ones present in full fundus cameras. It is challenging to get a good alignment at all times and even more challenging to tune the prototype parameters appropriately, for the environment and for each patient, since the number of diopters and other ocular conditions that the patient might have required some adjustments to the prototype.

These adversities were reflected in the quality of the acquired images, the selected parameters were not optimal in a large percentage of the acquisitions, which resulted in very noisy images, making it difficult/impossible to see the optic cup in the images and posing some challenges for the segmentation and classification tasks with this images.

Furthermore, only 3 patients reported that they had Glaucoma, which unfortunately results in a heavily unbalanced dataset, less useful for Glaucoma classification.

5.2.3.2 Ground Truth

Before performing the acquisitions, it was intended to elaborate the ground truth with ophthalmologists annotations. These annotations would concern the optic disc and cup boundary and a label indicating if the image was indeed Glaucomatous or not. However, given the limitations of the acquired images, it was not viable to annotate the optic cup boundary in a big portion of the images, or ask the ophthalmologists to annotate the images as glaucomatous or not, since the patients questionnaire already made clear that the percentage of Glaucoma cases was very low, therefore not justifying the cost of having ophthalmologists annotating the images.

The dataset ground truth was then built with the optic disc boundary annotation. This annotation was performed by myself using a publicly available annotation tool [FdlR15], that can be seen in figure 5.6 with an annotation example.
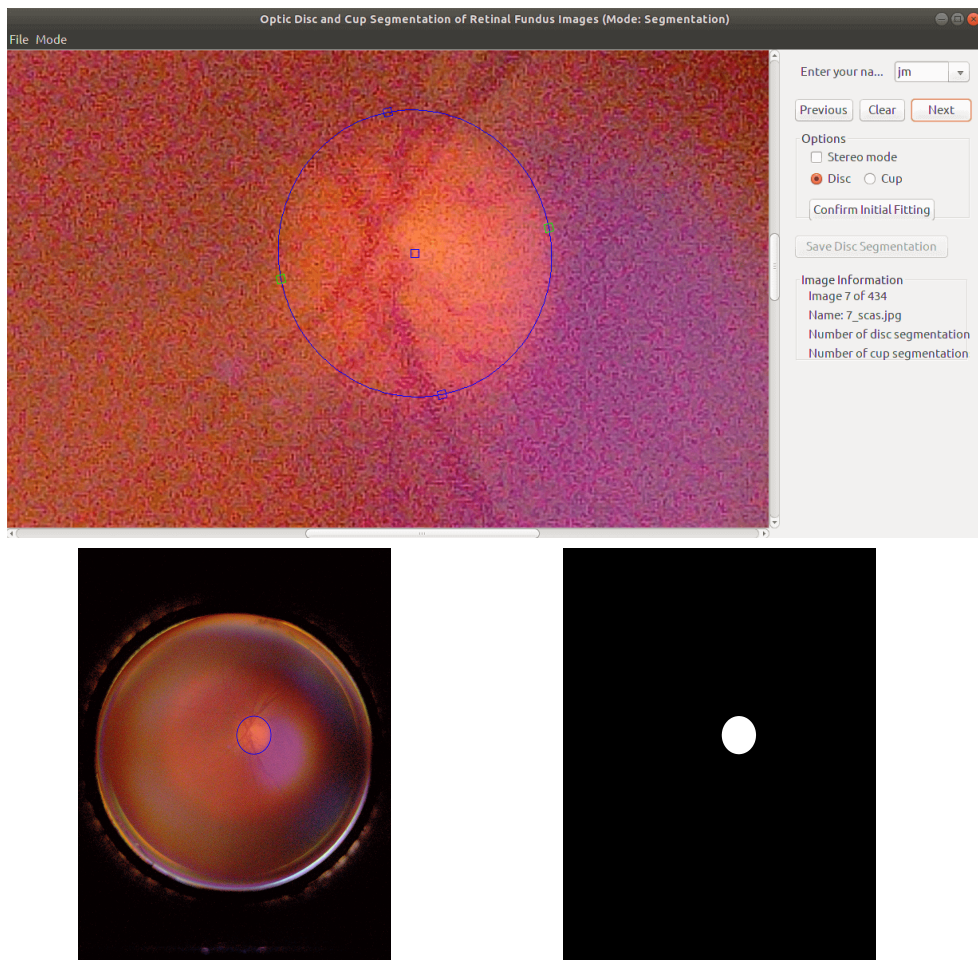
Figure 5.6: SCAS annotation

5.3   Dataset Construction

Considering the previously addressed datasets (sections 5.1 and 5.2) briefly described in table 5.1, their annotations, and the desire to explore both segmentation and classification tasks, three datasets where created.

Dataset	Glaucoma	No-Glaucoma	Total
Origa	168	482	650
Drishti-GS	70	31	101
RIM-ONE r1	40	118	158
RIM-ONE r2	200	255	455
RIM-ONE r3	148	170	318
iChallenge	80	720	800
RIGA	749		749
SCAS	434		434

Table 5.1: Image count per dataset

Firstly, a dataset for segmentation tasks of both the optic disc and optic cup was created merging the Origa, Drishti-GS, iChallenge, RIM-ONE r3, and RIGA datasets, since they all had in common optic disc and cup annotations. In the datasets where several annotations were provided for the same image, the considered ground truth was the region of agreement between the annotators. Considering, for instance, the RIGA dataset, where 6 ophthalmologists annotated all the images, the considered annotation was the result of the AND operation of all the 6 annotations.

Secondly, in order to validate part of the developed solution with images acquired by EyeFundusScope, the SCAS dataset was used.

Lastly, for the classification task, Origa, Drishti-GS, RIM-ONE (r1,r2,r3) and iChallenge got merged, given that they have glaucoma annotations. Furthermore, only two classes were considered, Glaucoma and No-Glaucoma, since the majority of the datasets only included a binary classification for Glaucoma. In situations where the glaucoma classification was not binary (RIM-ONE r1) or the Glaucoma cases where mixed with Glaucoma suspects (RIM-ONE r3), the samples were all considered as Glaucoma, since in CAD systems for glaucoma detection the goal is to achieve a good sensitivity hence every image that poses a doubt about its classification is considered as Glaucomatous in the ground-truth of the classification dataset.

Additionally when considering the RIM-ONE r3 dataset, which consisted of stereo images, each of these images was divided into two, and considered as a separate case. With this in mind the number of images in this dataset, present in both the segmentation and classification datasets, increased from 159 to 318. To summarize, a quick overview of these three datasets can be found on table 5.2.

Dataset	Description	Total
Segmentation	Origa + Drishti-GS + RIM-ONE r3 + iChallenge + RIGA	2618
Segmentation-SCAS	SCAS	434
Classification	Origa + Drishti-GS + RIM-ONE (r1,r2,r3) + iChallenge	2482

Table 5.2: Final segmentation and classification datasets

5.4 Summary and Conclusions

The created datasets for segmentation and classification are rather small when considering the usual Deep Learning practice, nonetheless, when compared with the datasets used by some of the state-of-the-art approaches in CAD systems for Glaucoma diagnosis, the gathered datasets have comparable or sometimes better size. Additionally, the fact that several datasets were merged, increases the variability within the dataset, contributing to a better generalization capacity in Machine Learning/Deep Learning approaches [AFLC+19]. It is also important to notice that there are several private/paid datasets that are being used in the development of CAD systems for Glaucoma detection. In fact, most of the published systems use their own private datasets [HKT+18].

As a final conclusion, unfortunately, the nonexistence of a dataset of EFS images with glaucoma, optic disc, and optic cup annotations, turns impossible the full validation of the developed solution solely in the EFS context, but on the other side, the accomplished merge of public datasets result in a robust dataset.

Data Gathering

Chapter 6

Image Pre-processing

In order to enhance the images that build the 3 datasets used in this work (section 5.3) and make them more suitable for the learning process, several image processing techniques were performed. In the present chapter, a detailed overview of each of the experimented techniques is provided.

6.1   Crop

Given that the images come from several different datasets having therefore different aspect ratios, it is important to normalize this situation. This is accomplished by cropping the image and not resizing it since the shape of the structures in the region of interest should be kept because several morphological features are calculated based on widths and areas of this structures and a resize would introduce errors.

The goal was to keep a 1:1 aspect ratio since this is the ratio used in the majority of deep learning architectures inputs, and will also be used in the approaches used by this work, to enable the use of pre-trained networks. Considering the aforementioned two types of crops were executed. First, a center crop was performed when it was necessary to change the image aspect ratio to 1:1.

The second type of crop consisted of cropping the ROI region using the annotated disc boundary since this region contains the most relevant features for Glaucoma assessment. An example of both types of crops can be seen in figure 6.1.

6.2   Data Normalization and Standardization

This is a simple and common pre-processing technique used in machine learning/deep learning approaches [HLVW17, MZZS18]. Data normalization consists in keeping the intensity values of the images between 0 and 1, this is done by simply dividing the original the image intensities in all channels by 255. Additionally, data standardization consists of re-scaling the data so that it ends with 0 mean and a standard deviation of 1 [Bro19].

| Original image | Center Crop | ROI Crop |
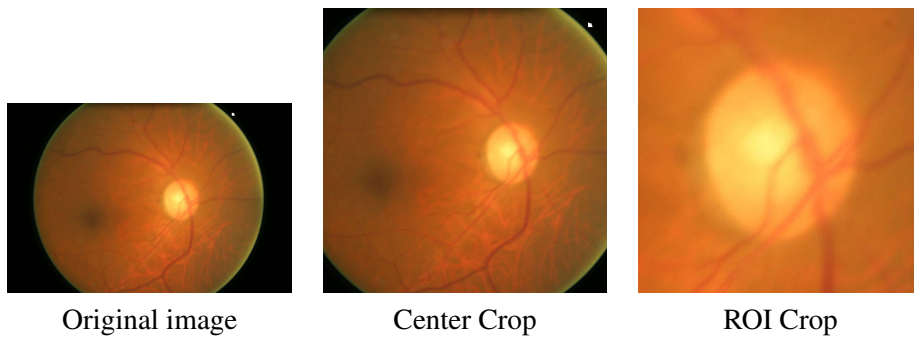
Figure 6.1: Crop showcase

Data normalization and standardization improves the convergence of the models because the data is distributed in a more consistent way, avoiding that specific samples with different ranges have a greater effect than the desired in the model weights.

An example of this transformation can be found in figure 6.2.

Figure 6.2: Normalization example, with normal image on the left and feature-wise mean subtracted image on the right

6.3 Contrast Limited Adaptive Histogram Equalization (CLAHE)

CLAHE [Zui94], first introduced in 1994, its a technique to enhance the contrast of an image, based on histograms, it has its roots in medical imaging but has already been used in a variety of other contexts.

Histogram equalization, its based on the assumption that a good image will have its pixel intensities distributed by the full histogram and not confined to a single region, it is therefore accomplished by effectively distributing the brighter intensities through the full histogram, allowing for areas with low contrast to be more noticeable, figure 6.3 provides an illustration.

There is however a problem with this approach, the brighter pixels of the images might also be caused by undesired areas, especially noisy areas, resulting in the enhancement of the contrast of these undesired areas. To solve this issue, Adaptive Histogram Equalization (AHE) was introduced, in this approach instead of using the full image context, the images are divided into a grid of contextual regions of a defined size and histogram equalization is performed on each contextual
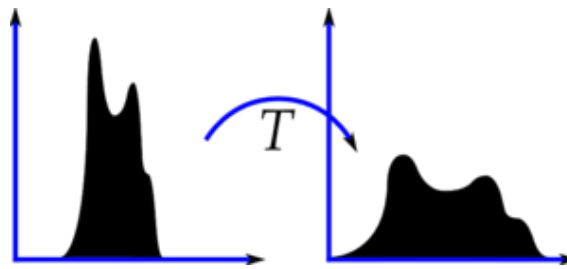
Figure 6.3: Histogram Equalization (extracted from OpenCV documentation)

region with bilinear interpolation in the boundaries of the regions to avoid the visualization of artifacts. Despite this, a contextual area with noise would still increase the noise representation and that's were CLAHE finally comes in, introducing a clip limit parameter that is used to clip pixels in a histogram bin that doesn't respect this limit and distribute them between other bins before applying the histogram equalization.

In the context of fundus images, CLAHE has already been used in several approaches [ZFZ$^+$18, SSS$^+$19, ABRL15, TJ18], in this work, CLAHE was used with a contextual area of size 8x8, this was the one used in the original paper and usually gives good results, with regards to the clip limit, 2 was the selected value. Moreover, since CLAHE is based on the pixel intensities of one channel, it should be applied to intensity-related channels, for this, instead of applying CLAHE to each RGB channel a different approach was followed. The RGB fundus images were transformed to the LAB color space, composed of 3 channels, lightness (L), red to green value (A) and blue to yellow value (B). CLAHE was then applied to the lightness channel, and after that, the image was re-converted to RGB color space.

In figure 6.4 an example of the application of this algorithm to both high-quality fundus images and EFS fundus images can be found.

6.4 Polar Transformation

A point in the polar coordinate system is represented by two coordinates, a distance (r) from a reference point and an angle (θ) from a reference direction [Wik19]. In the context of fundus images, this transformation is useful if the reference point is the disc center and the reference direction is the positive direction of the x-axis, given that the disc center is the origin. This results in a simpler separation between the disc and the cup, that are the most important structures for the Glaucoma diagnosis. Besides this, the cup proportion with respect to the image background is also enhanced.

The conversion between the polar coordinates system and vice versa, can be accomplished with equations 6.1 and 6.2.

An example of this transformation can be found in figure 6.5.
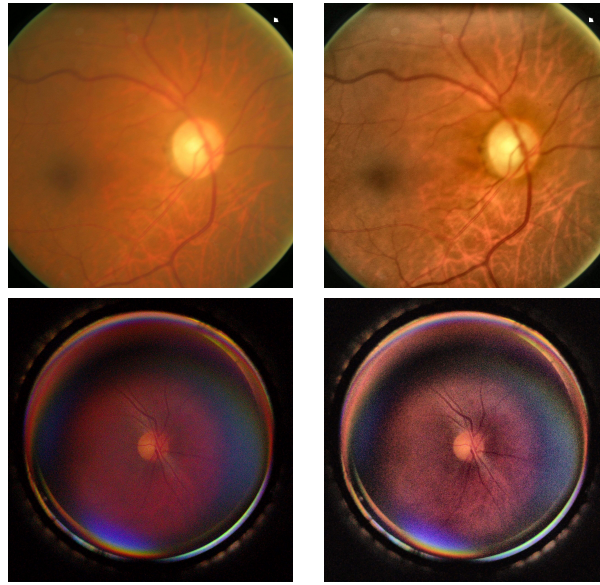
Figure 6.4: CLAHE example (left: normal; right: enhanced)

$$Cartesian = \begin{cases} x = rcos\theta \\ y = rsin\theta \end{cases} \quad (6.1) \quad Polar = \begin{cases} r = \sqrt{x^2 + y^2)} \\ \theta = atan2(y,x) \end{cases} \quad (6.2)$$
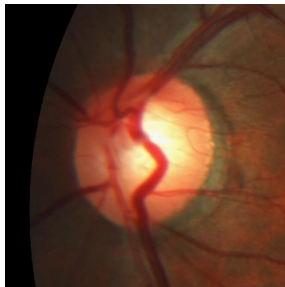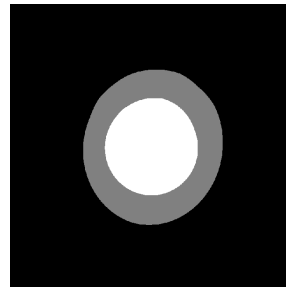
Cartesian coordinate system image   Cartesian coordinate system mask
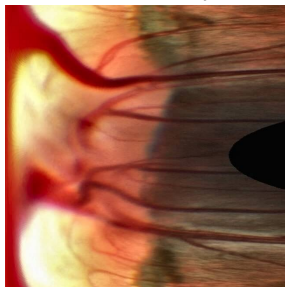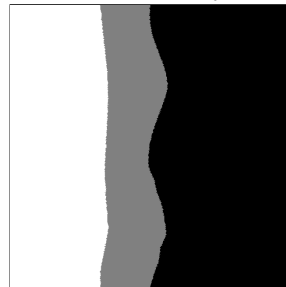
Polar coordinate system image   Polar coordinate system mask

Figure 6.5: Polar transformation example

This transformation has already been used for segmentation [FCX⁺18a] and classification tasks [FCX⁺18b] in Glaucoma diagnosis with fundus images, as addressed in section 3.3, and will also be explored in this work.

6.5 Denoising/Smoothing

Considering that one of the main issues of the SCAS dataset is the huge amount of noise present in the images (section 5.2.3.1), some denoise/smoothing techniques were performed on these images to combat this issue. This was only performed on the SCAS images since the images from public datasets do not present any noticeable noise.

First, gaussian blur was performed, where a Gaussian kernel is used through the image, however, the increased blur made the image seem rather unfocused.

Then, median blur was used; this takes the median of all the pixels in a window with a specified dimension and replaces the center pixel with the result. It gave better visual results than a gaussian blur, nonetheless, this technique does not preserve retinal vessels with small areas.

Bilateral filtering [TM98] was then tested. This technique its recognized for its simplicity and effectiveness and was introduced for edge-preserving smoothing, which is very useful to mitigate the issue with the retinal vessels previously described. It uses two filters, a spatial filter that measures spatial distances to enforce that only nearby pixels are considered and a range filter, that measures intensity differences, to ensure that just pixels with intensities similar to the central pixel are blurred. This allows the edges to be preserved, since edge pixels have significant variations in their intensities, thus not being considered for blurring. This method has already been applied to fundus images in the past [HZZ+17] and provided satisfactory results with the SCAS images as well.

Additionally, OpenCV (3.1.1) provides an implementation for Non-local Means Denoising[1], it builds from the assumption that noise is usually considered a random variable with zero mean and in order to find it, instead of considering a local area like in the previous methods, a pixel can be averaged between multiple similar patches of the images and the result will hopefully be the value of the pixel without noise, since noise is assumed to have zero mean. This was also tested and the results were good, nonetheless, the bilateral filter continued to do a better job at preserving retinal vessels with small areas. In figure 6.6 a sample of the results obtained by the addressed techniques can be found.

| Original | Bilateral Filter | Gaussian Blur | Median Blur | Non-local Means |
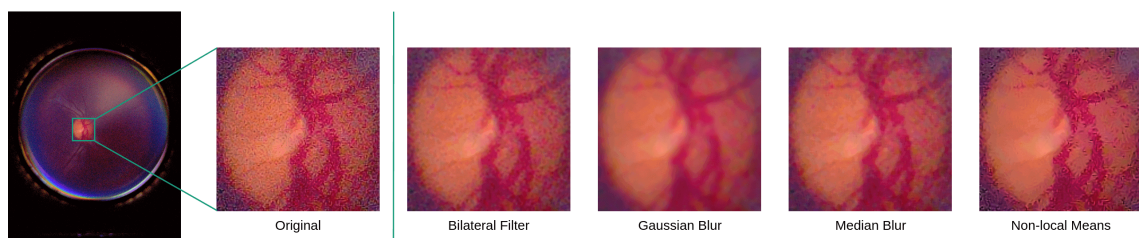
Figure 6.6: Denoise techniques performed on SCAS dataset

---

[1]https://docs.opencv.org/3.2.0/d5/d69/tutorial_py_non_local_means.html

6.6 Augmentation

A good augmentation pipeline, that enhances the existing datasets in a representative way by systematically creating transformed images is of ultimate importance, especially when considering deep learning approaches where a greater amount of data is needed to achieve a good generalization and decrease overfitting issues.

For this, real-time data augmentation approaches were used, since they are robust and easy to implement, not requiring additional disk space to save the augmented images and providing different augmented images in every batch of data.

Given that the goal is to create new data, but at the same time ensure that the data is representative for the task, in this augmentation pipeline no geometric augmentations were performed, since for instance a random horizontal/vertical flip or other geometric changes might affect the calculation of morphological features in the image and therefore the final classification decision, because the ISNT quadrants might be swapped or the shape of some structures like, the disc and cup can be changed, inducing errors in the classification.

The augmentation pipeline was built using the imgaug library (section 3.1.2). It consists of 4 main steps that are applied in random order and from augmentation to augmentation some of them might even be skipped. The first step consists in the application of blur to the image, this blur can be either a gaussian blur an average blur or a median blur. Secondly, contrast normalization is performed by moving the pixel values away or closer to 128. Then brightness changes are performed per channel. Finally, the images might be sharpened with a sharpening kernel.

In figure 6.7, an example of the results of this pipeline can be found.
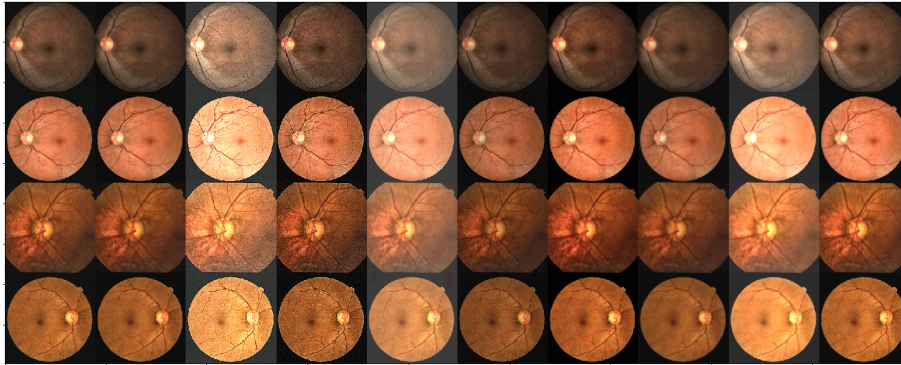
Figure 6.7: Augmentation example

6.7 Summary and Conclusions

All the addressed techniques in this chapter were already explored in the context of Glaucoma CAD systems backed by fundus images, and are therefore relevant and useful for this problem.

In the following chapters, several combinations of these preprocessing techniques will be experimented and discussed with a variety of deep learning approaches.

Chapter 7

Segmentation Approaches for Glaucoma Diagnosis

Accurate segmentation of the optic disc and the optic cup is a very relevant task in a CAD system for Glaucoma diagnosis since it allows the calculation of several morphological features that can aid the classification decision. Lately, deep learning based approaches are surpassing the traditional segmentation approaches as addressed in section 3.2.2 and, in this chapter, several implemented deep learning techniques are explored and discussed with a focus on mobile inference.

In section 7.1, a careful description of the models architectures is provided. Then an overview of the training process can be found on section 7.2, followed by a discussion of the final results in section 7.3 and the main issues faced by the models in section 7.4.

Finally, the morphological features that were calculated with the segmentation results are explained and the main conclusions from this chapter are disclosed in section 7.5 and 7.6, respectively.

7.1 Architectures

7.1.1 Building Blocks

In order to construct the networks, several major building blocks addressed and tested in the literature were used. Below a description of each one is unveiled.

- **U-shape architecture** - This architecture consists of an encoding path, where the convolution and pooling operations are performed and a decoding path, where the pooling operators are replaced by up-sampling operators. Additionally, the number of filters used is the same or at least similar in both the encoding and decoding paths at the same depth level, resulting in a symmetric architecture, hence the '*U*' in the name. This type of architecture was

first applied with success on segmentation tasks in U-net [RFB15], and since that has been widely used.

- **Skip Connections** - Skip connections consist of copying and concatenating feature maps from the encoding path to the decoding path at the same depth level. This is useful for segmentation tasks since the information from the earlier layers contains more spatial context from the initial images, and adding that information to the last layers will positively aid the up-sampling operation and consequently enhance the segmentation results [DVC+16].

- **Multiple inputs/outputs** - Multiple inputs of different scales, have similar advantages to the skip connections, providing more context in advanced layers on the encoding path. On the other side, multiple outputs can alleviate the vanishing gradients problem, since these outputs help back-propagate the loss to earlier layers in the decoder. For this, the final output should consist of an operation that combines the multiple outputs, for instance, an average operation [FCX+18a].

- **Relu 6** - This consists of using Relu as the activation function between layers with a max value of 6, which results in a more robust non-linearity for low-precision computation [SHZ+18]. The previously mentioned makes the use of Relu 6 very appealing for models that are intended to run in embedded devices and mobile environments.

- **Spatial Pyramid Pooling (SPP)** - SPP [HZRS15b] was originally proposed to address the fact that arbitrary input sizes could be used with convolutional layers, yet, when considering fully connected layers this is not the case. Therefore SPP acts as a new layer between convolutional and fully connected layers that can map any input size to a fixed output size, solving the mentioned issue. This is achieved by dividing the image in a number of spatial bins with size proportional to the input size. The number of bins is fixed and independent of the input. Additionally, the division of the image in spatial bins is performed at different levels, in example, a layer with 16 bins divides the image in a 4x4 grid, another layer of 4 bins will divide the image in a 2x2 grid and a layer with 1 bin will consider the full image. In each of these layers, pooling is performed in all bins, the layer is flattened and then all the layers are concatenated. Since the number of bins is fixed, this concatenation will always lead to a fixed length representation. Furthermore, this module also acts as a global context prior and it is therefore very useful for segmentation tasks. An example is Pyramid Scene Parsing (PSP) [ZSQ+17], which has demonstrated outstanding performances on several segmentation problems using a variation of the addressed SPP module. In PSP, pooling is also applied at different scales but instead of flattening and concatenating the result, each layer goes through a pointwise convolution (1x1 kernel) to reduce the number of filters and keep the weight of the original feature map since afterwise, the layers at different scales are upsampled to the original resolution and concatenated between themselves and the original feature map, as illustrated in figure 7.1. For this work, variations of the PSP spatial pyramid pooling module were used.

| (a) Input Image | (b) Feature Map | (c) Pyramid Pooling Module | (d) Final Prediction |
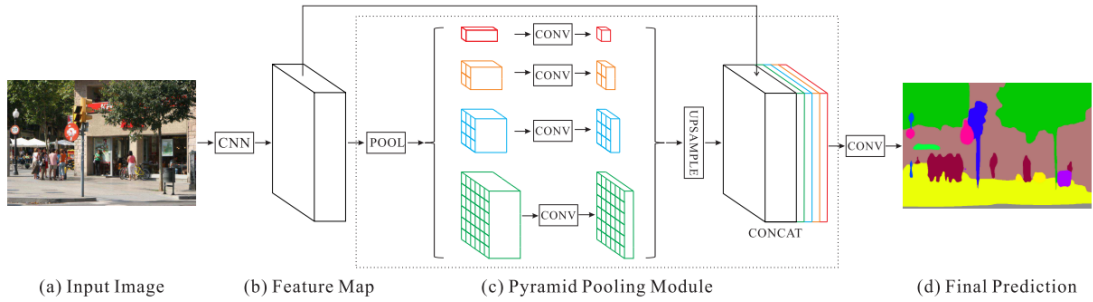
Figure 7.1: SPP module in Pyramid Scene Parsing [ZSQ+17]

- **Atrous Spatial Pyramid Pooling (ASPP)** - This module is inspired in the previously addressed SPP module and was used in several iterations of DeepLab models [CPK+16, CPSA17, CZP+18], that is currently the state-of-the-art model in several segmentation benchmarks. ASPP is based on atrous convolutions (also called dilated convolutions), these are convolutions with a dilation rate, where a value of 1 is equivalent to a normal convolution, and a greater value allows the enlargement of the receptive field or in other words, the field of view of the filters, allowing a better incorporation of context in the downscaling process. In figure 7.2 an example of two dilation rates can be found. In the ASPP module, instead of using pooling operations at different rates, like in SPP, dilated convolutions are used at various rates. This can alleviate the loss of information related to object boundaries caused in pooling or strided convolutions.
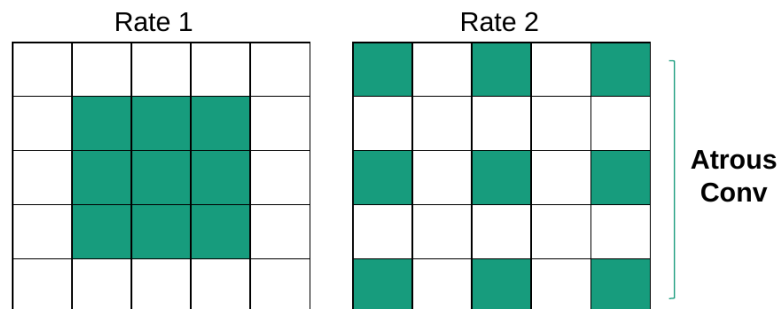
Figure 7.2: Example atrous/dilated convolutions at different rates

7.1.2 GFI-SPP

This was the first developed architecture to perform the segmentation of both the optic disc and the optic cup. An illustration is provided in figure 7.3. It consists of a symmetric encoder and decoder path and a spatial pyramid pooling module.

The encoding path consists of 4 depth levels and is similar to typical CNNs architectures. In each level, two padded convolutions with relu6 activation's and 3x3 kernels are performed,

the number of filters starts at 32 and is then doubled on every transition to next depth level, this transition is performed through average pooling with a pool size of 2. Average pooling was selected in detriment of max pooling since the literature reports better results with it on similar tasks [ZSQ⁺17]. Furthermore, two inputs are used in the network in the first two depth levels and skip connections between the encoder and decoder path are present at every depth level.

When it comes to the transition from the encoder to the decoder path, an SPP model was used, in this implementation, it consists of 4 different pooling levels, the first considering the full image and performing average pooling, and the following three dividing the image into 4 bins, 16 bins, and 49 bins, and then performing the average pooling. This results in 1x1, 2x2, 4x4 and 7x7 filters that then go through pointwise convolutions in order to keep global context and later are upsampled and concatenated.

Now we reach the decoding path, it is very similar to the encoding one, however, in here, the transition to an upper level leads to the decrease in the number of filters in each convolution by a factor of 2 and is done with nearest interpolation. Finally, in the first two depth levels of the decoder, two outputs are generated and then averaged to form the final segmentation prediction.

Wrapping up, this architecture consists of a total of 1,963,126 parameters and 61 layers.
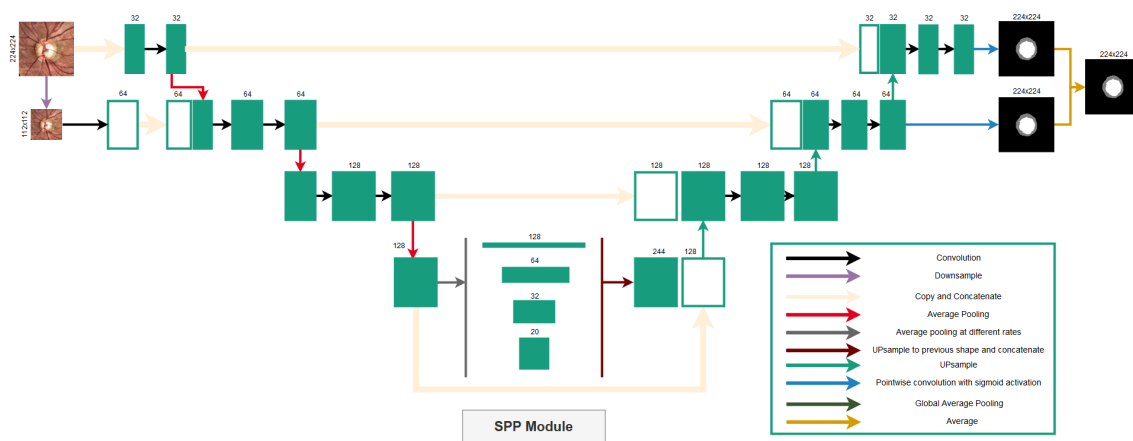
Figure 7.3: GFI-SPP architecture diagram

7.1.3 GFI-ASPP

This architecture is very similar to GFI-SPP (7.1.2), in fact, as the name suggests, the only difference is the module used in the transition between the encoder and the decoder path.

In here, an ASPP module was chosen, it consists of 4 parallel padded atrous convolutions with dilation rates of 1, 2, 4 and 7. Additionally, image-level features obtained by performing a global average pooling on the original feature map are also used. This is done because when the dilation rate increases, the number of filters weights that are applied to the valid feature region (region without the padded zeros) declines. The inclusion of image-level features helps to mitigate this

problem and include more global context in the network as addressed in [CPSA17]. Finally, the result of each level of the ASPP module and the image-level features are concatenated.

This architecture consists of a total of 1,940,662 parameters and 57 layers and is represented in figure 7.4.
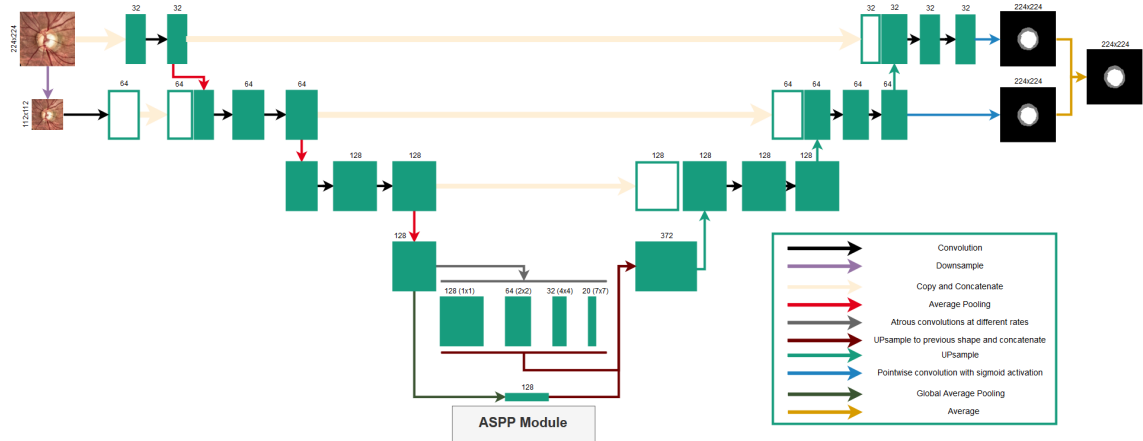
Figure 7.4: GFI-ASPP architecture diagram

7.1.4 Variants

Having the previously addressed architectures in mind (sections 7.1.2 and 7.1.3), three slightly different architecture variations were implemented:

- **GFI-SPP-Depth and GFI-ASPP-Depth** - These two variations consist of the replacement of all convolution blocks in the encoder and decoder paths by depthwise separable convolution blocks illustrated in figure 7.5. These blocks are inspired in the MobileNet architectures [HZC+17, SHZ+18], and help further reduce both the time and space complexity of the network while having minor or no impact on the performance, turning these architectures more suitable for mobile inference. With this change, the number of parameters in GFI-SPP-Depth and GFI-ASPP-Depth drops to 1,175,325 and 1,152,131, respectively.

- **GFI-SPP-Depth-simple** - This network is a simplified version of GFI-SPP-Depth, intended to be used just for disc segmentation since this is a rather simpler task than the joint disc/cup segmentation or the cup segmentation alone. Two considerable adjustments have been made, first, the starting number of filters in the first depth layer is 16 instead of 32, and is then decreased and increased in the same way as the parent architecture. Secondly, the SPP-model as also been massively simplified, including only a global average pooling operation. The transition between the encoder and decoder path consists therefore in the concatenation of the original feature map, the SPP model result and the result of a convolution block applied to the original feature map also introduced in this architecture.

49

| Depthwise Convolution | → | Batch Normalization | → | Relu 6 | → | Pointwise Convolution | → | Batch Normalization |
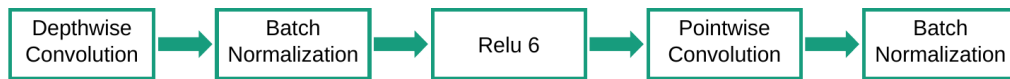
Figure 7.5: Depthwise separable convolution block

7.2 Training

With regards to the training process, firstly a validation approach was selected and used throughout all training sessions, in order to allow a consistent comparison between the trained models. The validation approach consists of a train/val/test split with 80/10/10 proportions of the original dataset.

When considering strategies of experimentation to continuously tune the model hyperparameters and increase the model performance, three options are often used, best guess, that consist in tuning one hyperparameter at a time, based on a 'guess'. Nested validation, that optimizes one hyperparameter, and then moves to another assuming that the previous one is optimal and lastly, grid-search, that consists in performing all possible combinations and selecting the best model. When training the segmentation models all of these approaches were used with a greater focus on nested-validation, due to time constraints.

Now, starting with the training process itself, first, the ADAM optimizer [KB14] was selected and never changed, since it usually achieves good results and is widely used in the literature. Loss wise, experiments with both cross-entropy loss and dice loss (2.3.2) were done, and it was verified that optimizing the dice-loss results in better end values for both the dice and cross-entropy losses, leading to an improvement of up to 0.02 and 0.37 in the final values of dice and cross-entropy losses, respectively. The dice-loss used was multi-label and the class weight (W_c) was 0.5 for both the optic disc and cup since both of these structures are equally important and should contribute equally to the loss value calculation. However in networks were just the disc is being segmented, the W_c of the disc is 1, as expected. At the end of each training session, the IoU was also calculated for both the disc and the cup and the hyperparameter optimization decisions were mainly based on this metric value.

Furthermore, callbacks were used in every training session, to save the model every time the validation loss decreases and to reduce the learning rate or perform an early stopping if the validation loss is not decreased for several epochs. With these callbacks the number of epochs was set to 200, considering that in the superiority of the cases, the early stopping occurred first, giving a complete perception of what the model could reach.

Considering the Segmentation dataset (section 5.3), several learning rates were explored and 0.0001 was selected since both lower and higher rates (values experimented with variation by a factor of 10 in both directions) conducted the model to local minimums in the majority of the cases with final IoU scores up to 0.17 lower. With regards to the batch size, a size of 32 was impossible to use due to hardware limitations, therefore 16 was the selected batch size for the experiments.

When it comes to augmentation since these models are built for segmentation, geometric transformations were tried, given that the limitation of these transformations addressed in section 6.6 only affects classification networks. These transformations consisted of small flips, shifts, and rotations, nonetheless, the augmentation pipeline specified in section 6.6 produced slightly better results more often and was then used in the final models.

Now addressing the preprocessing done to the images before applying augmentation and starting the train, initially, the center cropped fundus images with data standardization were used to reach the aforementioned parameters. Then experiments with the ROI cropped images (section 6.1) showed improvements up to 0.05 in the loss values and 0.02 in IoU and applying CLAHE (section 6.3) afterwise proved to increase the performance, even more, resulting in about 0.05 improvements in the average optic disc/cup IoU. Additionally, given that the green channel on these images has very good contrast and is sometimes used alone for segmentation tasks, training just with the green channel was also tried, however, the results were poor when compared to the previous approach. Next, Polar images (section 6.4) were tested on the ROI cropped images with CLAHE, and the obtained results were very similar to the previously obtained without the polar transformation, this transformation was then discarded, in order to simplify the pipeline. Furthermore, in all the mentioned tests, the images were all being resized to 224x224 before being provided to the networks, other dimensions where tried, like 400x400 images (with a batch size of 8 due to hardware restrictions), but no significant improvements were verified, and therefore 224x224 was the kept input size.

Since the images cropped on the ROI gave considerably better results for the joint segmentation of the optic disc and cup, networks just for disc segmentation were trained to power the crop of the ROI region. Given that this is a rather easier task, GFI-SPP-Depth-simple architecture was the chosen one. In the disc segmentation using CLAHE doesn't lead to better performances since the optic disc contrast, contrary to the cup, is already very good, so CLAHE wasn't used in disc-only segmentation tasks. Lastly, it was verified that the data standardization wasn't bringing significant performance increases, so to ease the integration of the pipeline in other system and prevent carrying the mean and standard deviation arrays, necessary to perform the same standardization every time, this preprocessing technique was also discarded.

In figure 7.6, an example of the losses convergence of one of the networks during train is shown, the plots are very similar in other architectures, with rather fast convergence in the begin and then, stabilization with the model converging at a very low rate until early stopping is eventually performed.

Now considering the Segmentation-SCAS dataset (section 5.3), in here, only disc segmentation models were experimented, since the dataset doesn't have cup labels. The training with this dataset was accomplished after the aforementioned experiments and similar hyperparameters were used. However, in these training sessions, the weights were initialized with the results from the training with the high-quality images. With this in mind, a lower learning rate of 0.00001 was used, and a batch size of 32 was adopted since with this dataset the hardware restrictions didn't pose any issue.
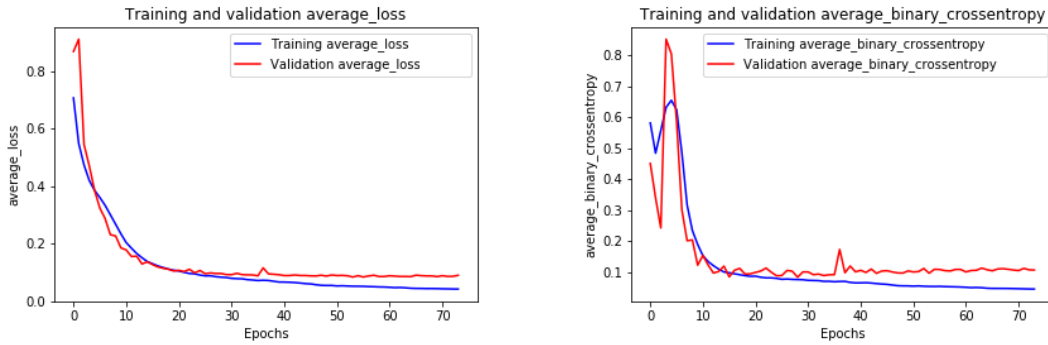
Figure 7.6: GFI-ASPP-Depth training losses

Despite, the disc segmentation being considerably easier to segment than the cup, with the Segmentation-SCAS dataset, the GFI-SPP-Depth-simple architecture didn't achieve results comparable to the more complete networks, such as GFI-ASPP-Depth, with differences greater than 0.3 in IoU values. This can be justified by the added difficulty of identifying the disc in this dataset, resulting in the necessity of a network with more modeling power.

With regards to the preprocessing, the center cropped images were used initially, then denoising using bilateral filtering was applied and increased the IoU up to 0.03. Afterward, CLAHE was also tried, however in this case, no improvements were verified for the same reason stated above and because the SCAS images are rather noisy, which sometimes result in contrast enhancement of noisy regions, affecting the segmentation performance.

An illustration of the loss values during the training is presented in figure 7.7, in here the convergence, in the beginning, is even faster when compared to figure 7.6, which can be justified by the fact that this approach already starts with meaningful weights considering that transfer learning is being used.
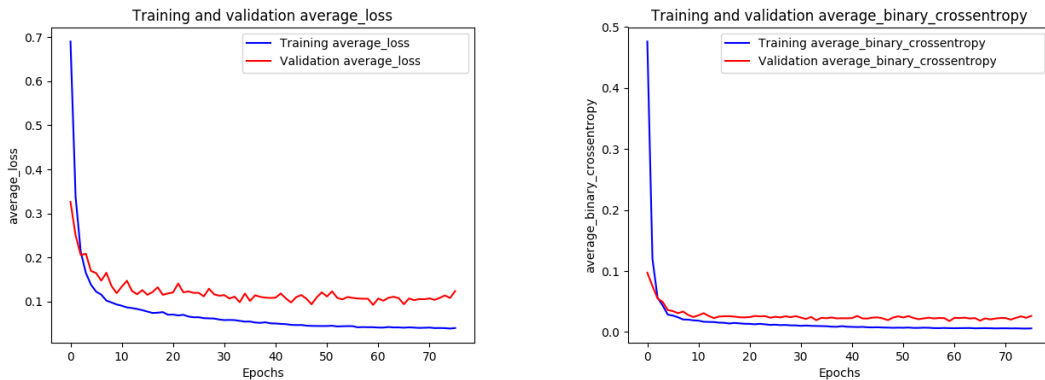
Figure 7.7: GFI-ASPP-Depth (Disc-SCAS) training losses

7.3   Results

In tables 7.1, 7.2 and 7.3 the best results of each developed model are revealed for the train,val and test sets, respectively. In each table, three highlighted models can be found, they represent the selected models for three different tasks, joint optic disc and cup segmentation (GFI-ASPP-Depth), disc segmentation (GFI-SPP-Depth-Simple (Disc)) and disc segmentation on the SCAS images 5.2 (GFI-ASPP-Depth (Disc-SCAS)).

Model	Multi-label Dice Loss	Log Loss	IoU Disc	IoU Cup
GFI-SPP	0.05	0.07	0.92	0.81
GFI-SPP-Depth	0.049	0.05	0.91	0.84
GFI-ASPP	0.05	0.07	0.92	0.81
GFI-ASPP-Depth	0.03	0.04	0.93	0.86
GFI-SPP-Depth-simple	0.08	0.09	0.89	0.73
Model	**One-label Dice Loss**	**Log Loss**	**IoU Disc**	**IoU Cup**
GFI-SPP-Depth-simple (Disc)	0.05	0.02	0.90	-
GFI-ASPP-Depth (Disc-SCAS)	0.03	0.005	0.92	-

Table 7.1: Train set segmentation results

Model	Multi-label Dice Loss	Log Loss	IoU Disc	IoU Cup
GFI-SPP	0.08	0.12	0.91	0.74
GFI-SPP-Depth	0.08	0.10	0.91	0.75
GFI-ASPP	0.08	0.10	0.91	0.75
GFI-ASPP-Depth	0.08	0.10	0.91	0.74
GFI-SPP-Depth-simple	0.10	0.12	0.88	0.71
Model	**One-label Dice Loss**	**Log Loss**	**IoU Disc**	**IoU Cup**
GFI-SPP-Depth-simple (Disc)	0.06	0.02	0.89	-
GFI-ASPP-Depth (Disc-SCAS)	0.10	0.02	0.80	-

Table 7.2: Val set segmentation results

Model	Multi-label Dice Loss	Log Loss	IoU Disc	IoU Cup
GFI-SPP	0.08	0.10	0.91	0.75
GFI-SPP-Depth	0.08	0.08	0.91	0.75
GFI-ASPP	0.08	0.08	0.91	0.75
GFI-ASPP-Depth	0.08	0.09	0.91	0.75
GFI-SPP-Depth-simple	0.10	0.09	0.89	0.71
Model	**One-label Dice Loss**	**Log Loss**	**IoU Disc**	**IoU Cup**
GFI-SPP-Depth-simple (Disc)	0.06	0.02	0.89	-
GFI-ASPP-Depth (Disc-SCAS)	0.07	0.01	0.86	-

Table 7.3: Test set segmentation results

Now, drawing some observations from the results obtained in each set, it is easily noticeable that the results between the validation and test splits are very similar and sometimes equal in the bulk of the models, this is very good since this sets indicate the generalization capacity of the models. However, when looking to the differences between the train set and the others, it is possible to see some considerable differences between the obtained losses, these can be justified by the difference of up to 0.12 in the cup IoU, indicating some slight overfitting in the cup segmentation during training. Nonetheless, this can be eventually justifiable due to the relatively small dataset size for deep learning tasks and the difficulty of segmenting the cup.

Regarding how these results were achieved, this can somewhat be inferred from the training section 7.2, but will be made explicit in the next paragraphs.

All the models used the online augmentation pipeline addressed in section 6.6 after resizing the images to 224x224 and performing data normalization (section 6.2), data standardization was not performed. In respect to the loss used, multi-label dice loss was selected, moreover, the ADAM optimizer was used on all models with a learning rate of 0.0001 on the models using the high-quality images and a learning rate of 0.00001 on the model using the SCAS images, since transfer learning was being performed. Additionally, the learning rate was reduced during the learning process by a factor of 2 if the validation loss was not improved in the last 10 epochs. Furthermore, the model was trained for 200 epoch with early stopping if, no improvements are verified in the validation loss after 20 epochs.

Now, considering GFI-ASPP-Depth in particular, the input images fed to the network consist in the ROI images with the application of CLAHE (section 6.3) and the model was trained with a batch size of 16 for a total of 74 epochs, since early stopping occurred. In the GFI-SPP-Depth-simple case, a batch size of 16 was again used, but the input images were the center cropped full fundus images, the model was trained for a total 94 epochs with early stopping occurring again. Lastly, GFI-ASPP-Depth (Disc-SCAS) was initialized with the weights of the same model trained in the segmentation dataset of high-quality fundus images. In here a batch size of 32 was used and the SCAS images were first center cropped and denoised using bilateral filtering (section 6.5), the model was early stopped at epoch 76.

The results between the models are similar, therefore, the selected models were the ones with lower complexities, both spatial and time wise.

Example results for each of the three selected models can be found in figure 7.8.

7.3.1 State-of-the-art Comparison

To further evaluate the robustness of the developed models, the achieved results are compared with state-of-the-art approaches in table 7.4, based on the IoU metric.

One issue when comparing these models is that the used datasets often differ quite a lot, making a direct comparison unfair in some cases. In order to mitigate this issue, the M-net model [FCX$^+$18a] without polar transformation (PT), was trained with the segmentation dataset used in the implemented models. Then the results of this model can be directly compared with the GFI-ASPP-Depth model that is the best all around developed model performance wise.
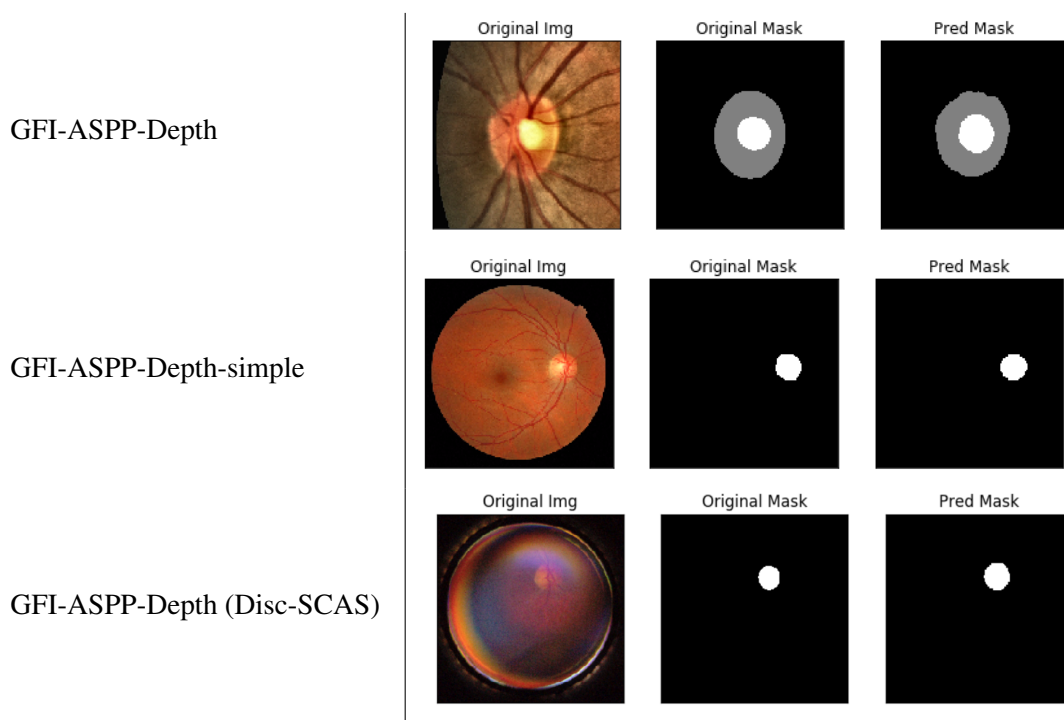
Figure 7.8: Segmentation predictions examples (images from the validation sets)

Furthermore, the state-of-the-art results extracted from [FCX$^+$18a], compare conventional methodologies with deep learning approaches and are comparable between themselves. Hence, by comparing the obtained M-net result with the result in the referred paper and then with the other state-of-the-art methods, it is possible to establish a point of comparison between all models.

Method	IoU Disc	IoU Cup
State-of-the-art results (Adapted from [FCX$^+$18a])		
R-Bend [JSK11]	0.871	0.605
ASM [YLO$^+$11]	0.882	0.687
Superpixel [CLX$^+$13]	0.898	0.736
LRR [XDL$^+$14]	-	0.756
QDSVM [CTWL17]	0.89	-
U-net [RFB15]	0.885	0.713
M-net [FCX$^+$18a]	0.917	0.715
M-net + PT [FCX$^+$18a]	0.929	0.770
Achieved results		
M-net (self-trained) [FCX$^+$18a]	0.87	0.70
GFI-ASPP-Depth	0.91	0.74

Table 7.4: Performance comparison with state-of-the-art

The results indicate that the developed model outperforms the existing methods while being considerably faster and smaller, given that for instance in comparison with the M-net model the number of parameters in GFI-ASPP-Depth is around 7 times smaller.

7.4   Issues

When observing the obtained results is possible to see a clear difference between the optic disc and optic cup IoU metric, this is an expected issue due to the increased difficulty in segmenting the optic cup since the accurate boundary of this structure is sometimes almost impossible to determine. This occurs due to the very modest intensity changes in the transition between the NRR and the optic cup.

The aforementioned is not an issue just for Deep Learning classifiers, the annotations of the optic cup for the same image by different ophthalmologists usually differ quite a lot, showing that this is also a hard task for humans. In figure 7.9 an example of this situation is given, the images in the RIGA dataset contain annotations from 6 different ophthalmologists and it is noticeable the difference in the cup boundary annotations from different ophthalmologists.
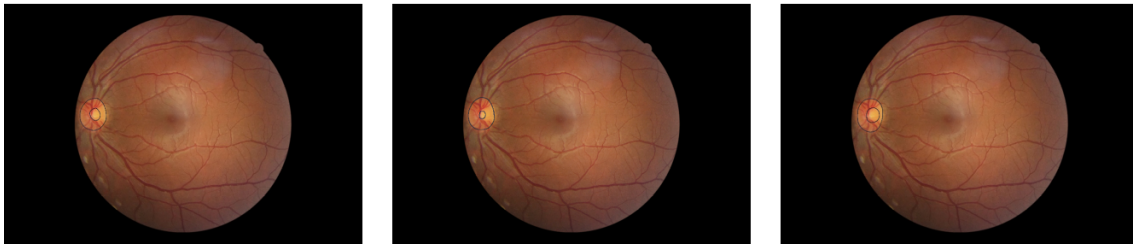
Figure 7.9: Riga image sample annotated by 3 different ophthalmologists

Furthermore, when considering the SCAS dataset, the heavy noise and the presence of artifacts not usually encountered in high-quality fundus images justify the reduced disc IoU in this model (GFI-ASPP-Depth (Disc-SCAS)), since this characteristics pose some great challenges to the model.

When performing a more careful analysis of the model predictions it is clear that the model is not robust to very low contrasts in the optic disc region and some artifacts in these images, like reflections. This is illustrated in figure 7.10, with an example of an image with various reflections and an image with very low contrast, alongside with the respective ground truths and segmentation results.

This issue is justified by the similarity that some of these reflections have with the optic disc and the fact that the majority of the dataset images don't have big reflections and usually present good contrast in the optic disc region, turning the robustness of the algorithm to these situations into a very challenging task.

7.5   Morphological Feature Calculation

The previously addressed segmentation models for the joint optic disc and cup segmentation can now be used to generate segmentation maps that power the calculation of several morphological features, addressed in section 2.2.2, for Glaucoma diagnosis. In figure 7.11, the majority of the
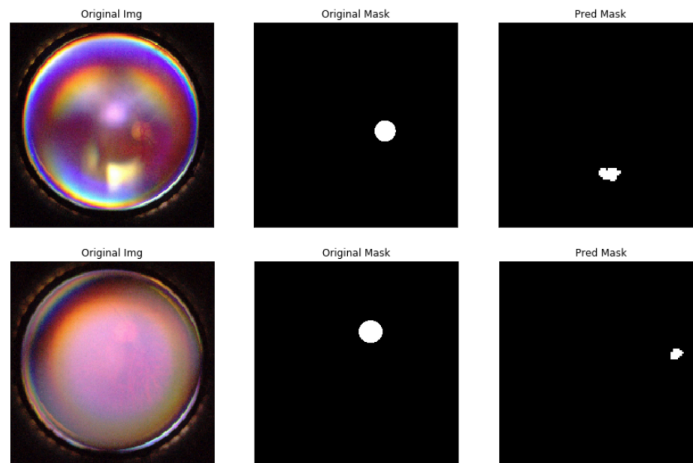
Figure 7.10: Segmentation issues in the SCAS dataset

values to calculate these features are illustrated, these are, the VDD (Vertical disc diameter), VCD (Vertical cup diameter) and the NRR (Neuro-Retinal Rim) widths in each of the four quadrants of a fundus image.

A total of eight morphological features were calculated, first the CDR was estimated based on the optic disc and cup areas, additionally, CDR was also calculated following the clinical convention by using the vertical diameters and was called VCDR. Furthermore, RDAR (Rim to Disc Area Ratio) that is similar to DDLS and was used in [FCX$^+$18a] was calculated using the NRR and disc areas as the name suggests. Moreover the NRR widths in each quadrant are the next four morphological features and lastly, the value indicating if the ISNT rule is followed or not is also computed.
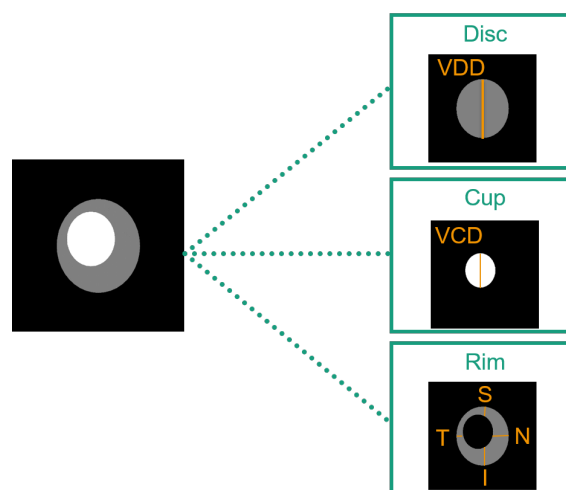
Figure 7.11: Concepts for Glaucoma morphological feature calculation

7.6 Summary and Conclusions

Several custom architectures were developed from the ground up and explored with a variety of preprocessing techniques and hyperparameters in this chapter.

The chapter has three main outputs, these are, GFI-ASPP-Depth, GFI-SPP-Depth-Simple (Disc) and GFI-ASPP-Depth (Disc-SCAS), a total of three segmentation models for three different tasks, joint optic disc, and cup segmentation, disc segmentation and disc segmentation with SCAS images, respectively.

The obtained results are very promising, considering that it was possible to achieve similar/better performances than the ones in the literature, while massively reducing time and space complexity.

Chapter 8

Classification Approaches for Glaucoma Diagnosis

This chapter follows an identical structure to chapter 7, however, in here, Glaucoma classification techniques are addressed.

Once again, deep learning models are used and in sections, 8.1 and 8.2 both the developed architectures and the details of the training process are revealed, respectively. Afterward, the final model results are disclosed and compared to the state-of-the-art approaches in section 8.3, followed by the main identified issues in section 8.4.

To wrap everything up, a brief summary of the chapter with some conclusions is provided in section 8.5.

8.1 Architectures

8.1.1 Building Blocks

Considering the network construction phase, solely deep learning techniques were used, namely CNNs, since in the machine learning spectrum a lot has already been explored, as addressed in the section 3.3, and nowadays the results of deep learning techniques have proven to be capable of improving the state-of-the-art and are therefore worthwhile being explored.

Given the relatively low amount of training data for deep learning approaches, pre-trained networks were used in an attempt to transfer the knowledge of other tasks to the Glaucoma classification task and mitigate this issue.

Three main building blocks were employed to construct the networks, a backbone network, fully connected layers, and dropouts, below a description of each one of them can be found.

- **Backbone network** - This backbone consists of a network trained for other tasks, such as ImageNet, and powers the transfer learning process. Only the convolutional blocks of these networks, known as the feature extractors, are used as the backbone.

- **Fully connected layers** - Also called dense layers, these layers have already been introduced in the brief deep learning section (2.3, and, are an essential block in CNNs. They connect every neuron in one layer to every neuron in other layer and are the last layers in CNNs.

- **Dropout** - This consists in random deactivation of neurons during training, acting as a network regularizer, preventing overfitting situations.

8.1.2 GFI-Classification

Considering that this work is intended to be integrated into a mobile environment, the selected backbone for this architecture was MobileNetV2 [SHZ⁺18], since it has proven to be a robust model with close to state-of-the-art performances in tasks like ImageNet while being considerably smaller and more efficient for mobile inference scenarios. This is achieved due to depthwise convolution blocks, inverted residual connections to help the flow of gradients through the network and expansion and projection layers.

After the MobileNetV2 block, pooling is performed to flat the features maps and then, two fully connected layers, interleaved by heavy dropouts are performed.

The result is an architecture with 2,299,521 parameters and 162 layers, a simple illustration of the network is available in figure 8.1.
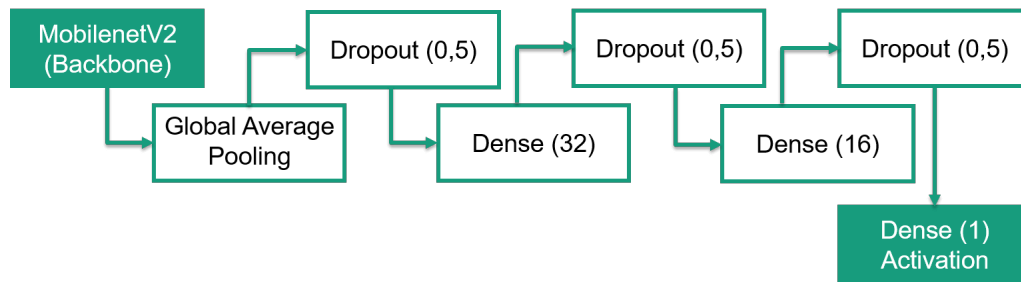
Figure 8.1: GFI-classification (MobileNetV2) architecture diagram

8.1.3 Variants

Several variations of GFI-Classification were created, either by changing the backbone or by changing the fully connected layers.

With regards to the backbone changes, 4 new different backbones have been used, the first two are VGG16 and VGG19 [SZ15], these architectures consist of convolution layers with 3x3 filters with max-pooling layers in-between some of them and finally 3 fully connected layers. The difference between VGG16 and VGG19 is only the number of weight layers each one has, 16 and

19, respectively. The novelty of this network was the increased depth when compared to other approaches at the time.

InceptionV3 [SVI$^+$15] is another used backbone, the first version consists of using multiple sized convolutional filters at the same depth level, and the following versions until V3, enhance the model by factorizing large convolutional filters to lower ones and using a regularization component called label smoothing, that can be added to the loss and helps to prevent overfitting.

Lastly, ResNet50 backbone [HZRS15a], is also used, this architecture helps to solve the problem of training very deep architectures by introducing shortcut connections in a plain network, turning it into a residual network. These connections are useful in very deep networks since they allow a proper propagation of the gradients, ResNet50, in particular, is a residual network with 50 weighted layers.

Regarding the changes in the fully connected layers, the details of these changes and the reasoning behind them will be discussed in the next section (8.2).

8.2 Training

Similarly to the training process followed by the segmentation models (section 7.2), in here, the validation approach consists in a stratified train/val/test with 80/10/10 dimensions, the hyperparameters where mainly optimized one at a time (nested validation), the ADAM optimizer was once again used on all experiments and the number of epochs was likewise set to 200, with callbacks to reduce the learning rate and perform early stop if the validation loss hasn't improved for a considerable amount of epochs.

Now, addressing specific training details of the classification models, first, the optimized loss was binary cross-entropy, and this was the main metric used for comparing intermediate results, additionally, for each training session, accuracy, sensitivity, specificity, F1 score and AUC were also computed. With regards to the dataset imbalance issue, since Glaucoma images constitute just about 1/3 of the dataset, class weights were used in all experiments in order to penalize the loss at a greater rate when it miss-classifies a Glaucoma case and mitigate the class imbalance issue. In all experiments, the backbones were first initialized with ImageNet weights and the input images resized to 224x224.

Then, a learning rate of 0.0001 and a batch size of 32 were used with the MobileNetV2 backbone to explore the impact of the final dense layers. Initially, two consecutive dense layers of 512 and 256 with intermediate dropouts were trained in various settings. However, the large number of parameters of these layers led to a considerable gap between the training and validation results indicating some overfitting to the train set. Then the number of units in the two dense layers was reduced to 32 and 16, which alleviated the previous issue. Tests with just one dense layer and with other dropout rates were also achieved, but the previous results were more consistent. Therefore, the 32 and 16 dense layers with 0.5 dropouts are the ones present in the GFI-Classification architecture (section 8.1.2).

After that, the GFI-Classification architecture was used to explore whether freezing the weights of the first layers would benefit the network or not. The layers were totally frozen and partially frozen at several different levels, but as initially expected, re-training all the layers was the best approach. This can be explained by the fact that classifying retinas as Glaucomatous or not is a very different task than ImageNet. Freezing the layers led to increases in the validation loss up to 0.20 in these experiments.

After deciding that all the layers would be trained, a bigger batch size of 64 was selected along with a smaller learning rate of 0.00001, to prevent losing the advantage of having meaningful weight initialization. Then, analogous to the segmentation approaches, several preprocessing techniques were tested, first the network was trained with raw center cropped images to create a baseline, then, ROI cropped images (section 6.1) were used, which resulted in a significant decrease in the validation loss, reaching in some experiments 0.21, additionally, CLAHE (section 6.3) was applied contributing to a slight decrease of 0.01. Moreover, polar transformations (section 6.4) were also performed, however, it led to a big increase in validation loss and was therefore discarded.

With respect to the followed augmentation approaches, geometric transformations, including flips and rotations conducted to a slightly better validation loss, but they greatly affected the ability to predict the Glaucoma class (sensitivity) in up to 0.15 in some cases, which can be justified by how these geometric transformations destroy relevant morphological information as addressed in section 6.6. The final used augmentation was, therefore, the pipeline addressed in section 6.6 which conducted to the best results, exposed in the next section 8.3.

Experiments with the morphological features calculated with the segmentation masks (in section 7.5) were also achieved, by adding the computed features to the last dense layer of the network, but it didn't result in any improvement, in fact, the results got significantly worse and the direct integration of the features in the network was not explored any further due to time constraints.

In figure 8.2, the evolution of the binary cross-entropy loss and accuracy during training is illustrated, showing a good convergence rate.
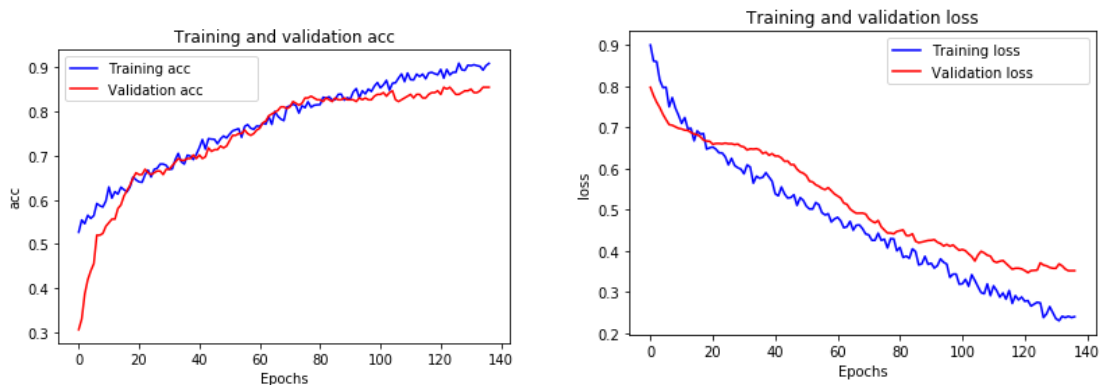
Figure 8.2: GFI-classification accuracy and loss during train

Lastly, the best hyperparameters were kept and GFI-Classification was also trained with the VGG16, VGG19, InceptionV3 and ResNet50 backbones.

8.3 Results

In this section a complete view of the obtained classification results is provided. In figure 8.3, a confusion matrix, illustrates the GFI-Classification inference results per class alongside a ROC curve, these results were obtained on the test split.
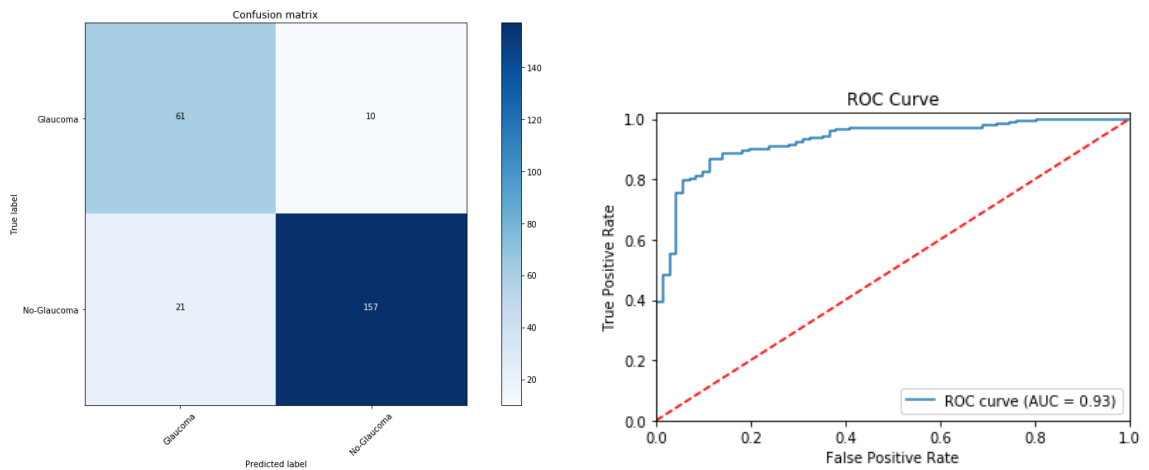
Figure 8.3: GFI-classification confusion matrix and ROC curve (test split)

Since the model should have a high sensitivity as addressed in section 4.2, different decision thresholds were tried besides the default 0.5, in order to better tune the sensitivity/specificity trade-off. In figure 8.4, a confusion matrix on the same split is illustrated, this time, obtained with a threshold of 0.25.

Figure 8.4: GFI-classification 0.25 threshold

In order to have a more complete view of the performance of different models, the tables 8.1, 8.2 and 8.3 provide all the computed metrics for each model in the train, validation and test split. In these tables, Acc, Sens, and Spec stand for accuracy, sensitivity, and specificity, respectively.

Model	Loss	Acc	Sens	Spec	Precision	F1	AUC
GFI-Classification	0.15	0.95	0.97	0.94	0.88	0.92	0.99
GFI-Classification (Thresh=0.25)	0.15	0.84	1.0	0.78	0.64	0.78	0.99
GFI-Classification (Dhristhi-GS)	0.21	0.90	0.94	0.88	0.75	0.84	0.98
GFI-Classification-VGG16	0.12	0.95	1.0	0.93	0.85	0.92	0.99
GFI-Classification-VGG19	0.09	0.97	0.99	0.96	0.92	0.95	1.0
GFI-Classification-InceptionV3	0.05	0.99	0.99	0.99	0.99	0.99	1.0
GFI-Classification-ResNet50	0.07	0.98	0.98	0.99	0.97	0.97	1.0

Table 8.1: Train set classification results

Model	Loss	Acc	Sens	Spec	Precision	F1	AUC
GFI-Classification	0.35	0.85	0.81	0.87	0.71	0.75	0.93
GFI-Classification (Thresh=0.25)	0.35	0.79	0.94	0.73	0.57	0.71	0.93
GFI-Classification (Dhristhi-GS)	0.51	0.79	0.81	0.74	0.87	0.84	0.87
GFI-Classification-VGG16	0.37	0.86	0.84	0.87	0.71	0.77	0.92
GFI-Classification-VGG19	0.39	0.85	0.72	0.90	0.75	0.73	0.91
GFI-Classification-InceptionV3	0.35	0.89	0.75	0.94	0.85	0.80	0.92
GFI-Classification-ResNet50	0.37	0.85	0.77	0.89	0.73	0.75	0.92

Table 8.2: Val set classification results

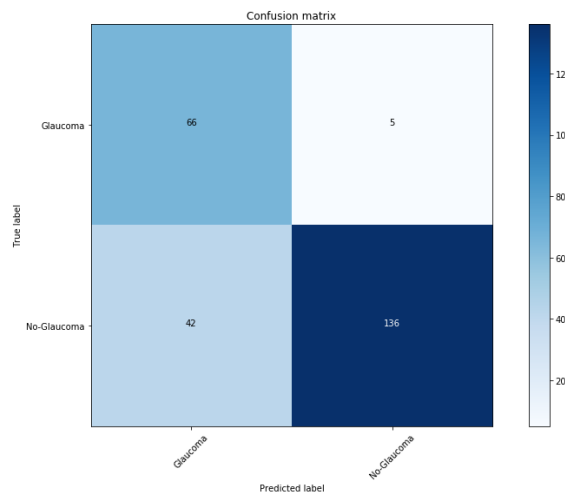| Model	Loss	Acc	Sens	Spec	Precision	F1	AUC
GFI-Classification	0.37	0.87	0.85	0.88	0.74	0.79	0.93
GFI-Classification (Thresh=0.25)	0.37	0.81	0.92	0.76	0.61	0.73	0.93
GFI-Classification (Dhristhi-GS)	-	-	-	-	-	-	-
GFI-Classification-VGG16	0.28	0.86	0.81	0.88	0.73	0.77	0.95
GFI-Classification-VGG19	0.27	0.87	0.83	0.89	0.76	0.79	0.95
GFI-Classification-InceptionV3	0.30	0.90	0.80	0.94	0.85	0.82	0.95
GFI-Classification-ResNet50	0.32	0.88	0.83	0.90	0.77	0.80	0.93

Table 8.3: Test set classification results

A total of 7 models were evaluated, first the GFI-Classification introduced in section 8.1.2, then, 4 of them consist of backbone variations with self-explanatory names, moreover, GFI-Classification (Thresh=0.25) is the GFI-Classification model with a threshold of 0.25 when performing the binary classification and finally GFI-Classification (Dhristhi-GS) consists of the GFI-Classification model trained on every dataset except the Dhristhi-GS, since this dataset was solely used for validation purposes in order to further evaluate the generalization capabilities of the developed model, by testing on a dataset completely uncorrelated with the training data.

All the models were trained with a batch size of 64, a learning rate of 0.00001, resized 224x224 ROI cropped images with CLAHE transformation as input and online augmentation using the pipeline addressed in section 6.6. Furthermore, callbacks to reduce the learning by a factor of 2 after 10 epochs and interrupt the training session after 20 epochs, if no improvements are achieved in the validation loss, were also used. With regards to the backbones, ImageNet weights were used on all networks and all the layers were trained.

Now comparing the results exposed on the tables, it is possible to identify almost perfect results in the training set, especially when considering models with bigger backbones, such as GFI-Classification-InceptionV3. This could mean a heavy overfit, however, when looking to the validation and test set results, it is possible to see that they are quite equal and, despite being considerably worse than the training results, they demonstrate a consistent generalization capacity of the models in unseen data, with results that are still very pleasant.

GFI-Classification is in bold, since it was identified as the best model for the task at hand, given the comparable results to the models with other backbones and superior results when concerning size and time complexities, mainly due to the MobileNetV2 backbone. Additionally, the GFI-Classification (Dhristhi-GS) shows that the model is robust to the outside world, since, despite the slightly lower achieved results, the metrics are still very compelling for a previously unseen dataset that has the particularity of having 2/3 of its images being Glaucomatous. The previously mentioned makes this validation approach turn the training set into an even more unbalanced one, which may justify, in part, the performance decrease.

When considering the model with a bigger threshold, it was not selected since the increase in sensitivity was not considered sufficient to justify the decrease in specificity, given that besides being important to accurately predict the positive class, false-positives should also be carefully managed to prevent filling hospitals with incorrect screening results.

8.3.1 State-of-the-art Comparison

The main state-of-the-art approaches are surveyed in table 8.4 ordered from top to down chronologically.

Similar to what was reported with the segmentation methods on section 7.3.1, in here, the different approaches differ quite a lot between themselves in terms of the followed validation strategy as well, using different metrics and datasets, making it challenging to perform accurate comparisons.

Despite the aforementioned, it is noticeable that the developed model (DL 7) is on par with other approaches, and accomplishes very similar results to the most recent method (DL 6), which was built by following a similar strategy, merging multiple public and private datasets and using pre-trained networks.

Moreover, the developed model is the only one who was conceived with mobile inference in mind and is, therefore, the most efficient model for this specific task. Other models like the DL 6 have around 30 million parameters while the developed model has just 2,299,521 parameters and more optimized operations.

Method	Datasets	Acc	Sens	Spec	AUC
State-of-the-art results (Adapted from [AFLC$^+$19])					
ML 1 - [MPA17]	RIM-ONE (250+/255-) Private (30+/30-)	0.81-0.98	-	-	-
ML 2- [ABK$^+$17]	Private (559+/143-)	0.95	-	-	-
DL 1- [ABANATZ17]	RIM-ONE (200+/255-)	0.88	0.85	0.89	-
DL 2- [FCX$^+$18b]	ORIGA (168+/482-) SCES (46+/1636-)	-	0.84	0.92	0.91
DL 3- [LHK$^+$18]	Private (48116)	-	0.95	0.92	0.98
DL 4- [CBB$^+$18]	Private (5633+/9189-)	-	0.88	0.95	0.91
DL 5- [MFT$^+$18]	Private (1364+/1768-)	-	-	-	0.96
DL 6- [AFLC$^+$19]	Private (113+/1333-) RIM-ONE r1,r2,r3 (311+/455-) Drishti-GS (70+/31-)	0.88	0.87	0.89	0.94
Achieved results					
DL 7- **GFI-Classification**	Classification (706+/1776-) Dataset (5.3)	0.87	0.85	0.88	0.93

Table 8.4: Performance comparison with state-of-the-art methods. In the methods column, *ML* and *DL*, indicate whether the method consists in a machine learning or deep learning approach, respectively. Considering the datasets column, the number of No-Glaucoma cases are marked with '-' and the number of Glaucoma cases are marked with '+'. Furthermore, the metrics Acc, Sens and Spec stand for accuracy, sensitivity, and specificity, respectively.

8.4 Issues

The main concern in the classification step of this work is the lack of annotated data for Glaucoma taken by the EyeFundusScope prototype addressed in section 5.2.1. This makes the fine-tuning of the developed models for the more restricted conditions of the prototype images impossible and invalidates the validation of the models with these images as well.

Other issues were also faced, like, the small amount of data for deep learning tasks and the class imbalance in the dataset. These issues, however, were partially solved by using pre-trained networks and class weights, respectively.

8.5 Summary and Conclusions

A variety of architectures were built, explored and validated in this chapter for the Glaucoma classification task.

The use of pre-trained architectures with ImageNet weights has proven to be a good way to suppress the low amount of data issue.

This chapter closes with the identification of a model that fulfills the expectations, by providing acceptable results when compared with the existing approaches in literature while being suitable for mobile environments.

Chapter 9

Developed CAD System

In this chapter, the implemented CAD system is described by combining the insights and solutions developed and addressed in the last chapters.

Starting with an overview of the full pipeline for Glaucoma assessment in section 9.1, after, the integration of the mentioned pipeline in the EFS system is revealed and discussed in section 9.2.

Next, one of the main requirements of the pipeline, the interpretability, is examined in section 9.3, and finally, the chapter closes with an overview of the main chapter topics and conclusions in section 9.4.

9.1 Pipeline

By gathering the insights of the available data and its characteristics (chapters 5 and 6), as well as the developed segmentation and classification models (chapters 7 and 8), a pipeline for Glaucoma assessment was constructed.

It was developed for high-quality fundus images; images acquired with dedicated fundus cameras and not the EFS system. The reason for this is that, as illustrated in the previous chapters, the prototype images have different characteristics that at the present moment don't allow the generalization of the developed models to these images. Moreover, the lack of annotated data for Glaucoma assessment with prototype images also prevented the exploration of classification approaches with them.

Now addressing the pipeline itself, in figure 9.1, a diagram with an overall illustration is presented.

The pipeline starts with an acquired full fundus images, this image is then center cropped and the GFI-SPP-Depth-simple (DISC) model, introduced in section 7.1.4, is employed to segment the disc in step '(1)', this segmentation is then used to transform the full fundus image in a ROI image during step '(2)', and then apply the CLAHE transformation.
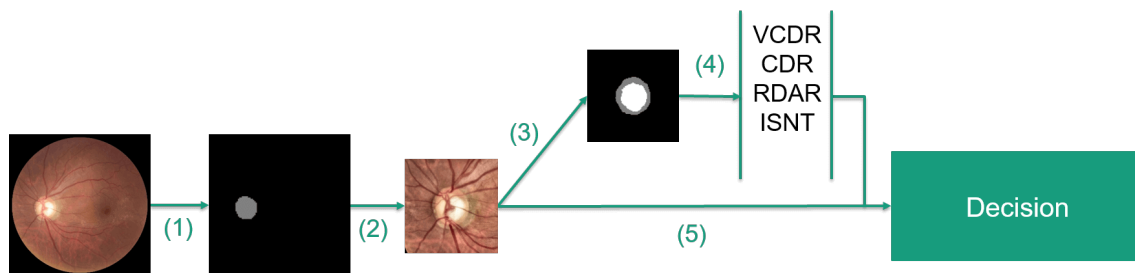
Figure 9.1: Classification pipeline

The CLAHE ROI image is then utilized for two different tasks that build the pipeline outputs. First, in step '(3)', GFI-ASPP-DEPTH, discussed in section 7.1.4, is used to obtain a fine optic disc and cup segmentation. This powers the calculation of several morphological features (section 7.5), that contribute to the interpretability of the decision.

Additionally, step '(5)', consists of running the images outputted from step '(2)' through the GFI-Classification network, addressed in section 8.1.2, this results in a Glaucoma confidence level.

Finally, the decision is built, merging the Glaucoma confidence level with the calculated morphological features.

9.2 Pipeline Integration in the EyeFundusScope System

In this section the previously addressed pipeline is integrated into the official Android EFS application, that was briefly introduced in 5.2.1. This application is focused on Diabetic Retinopathy, and now, the functionality for Glaucoma screening was added. The acquisition flow of the application consists in acquiring the image, answering a small questionnaire and then several services run in the background to perform image quality checks and run the assessment pipelines for Diabetic Retinopathy and now, Glaucoma.

The integration of the pipeline in this application, consisted in the creation of a background service that used TensorFlow Lite (section 3.1.4) to run the trained models on a mobile environment and OpenCV (section 3.1.1), with a Java wrapper, to manage the images throughout the pipeline. This service runs for every acquired image and continuously notifies the user about the progress of the Glaucoma evaluations in their acquisitions.

Furthermore, the acquisition flow was slightly changed, by re-designing the questionnaire activity, including new questions besides the existing ones that were only related to Diabetic Retinopathy. These are questions that can be easily answered in a screening environment and assess some Glaucoma risk factors addressed in section 2.1, like age, family history, and Cardiovascular incidents. They are not used in the current pipeline but are persisted in the patient profile to enable further studies with new techniques.

In figure 9.2, screenshots of the acquisition process focused on Glaucoma screening are exposed.
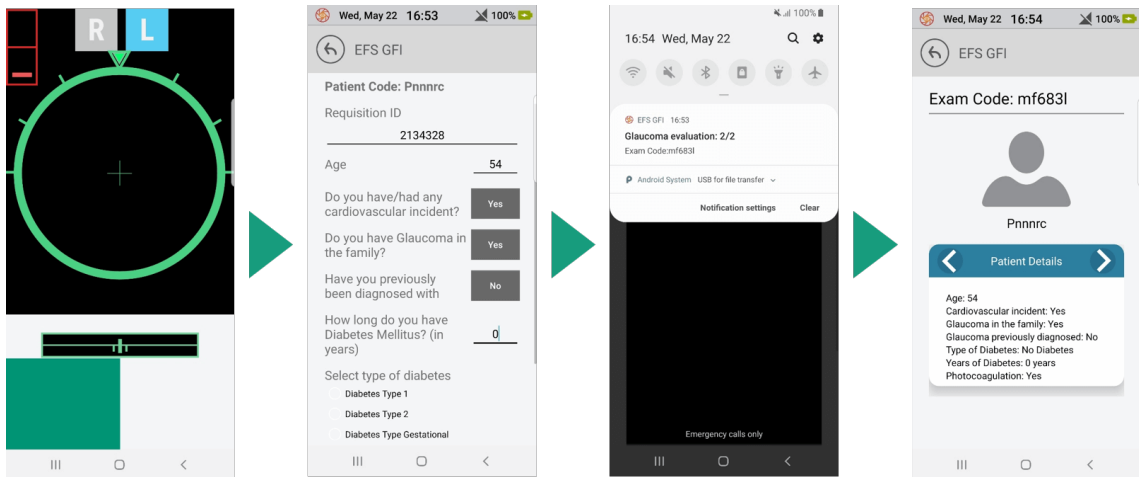
Figure 9.2: Android integration: Acquisition process

After the background service that performs Glaucoma assessment is completed, the user is able to see their acquisitions and for each one of them, the intermediate results of the pipeline are shown, namely, the first disc segmentation, the cropped image and the joint optic disc/cup segmentation.

Finally, the result of the morphological feature calculation, as well as the glaucoma confidence level, are accessible in a dialog. This is illustrated in the figure 9.3.
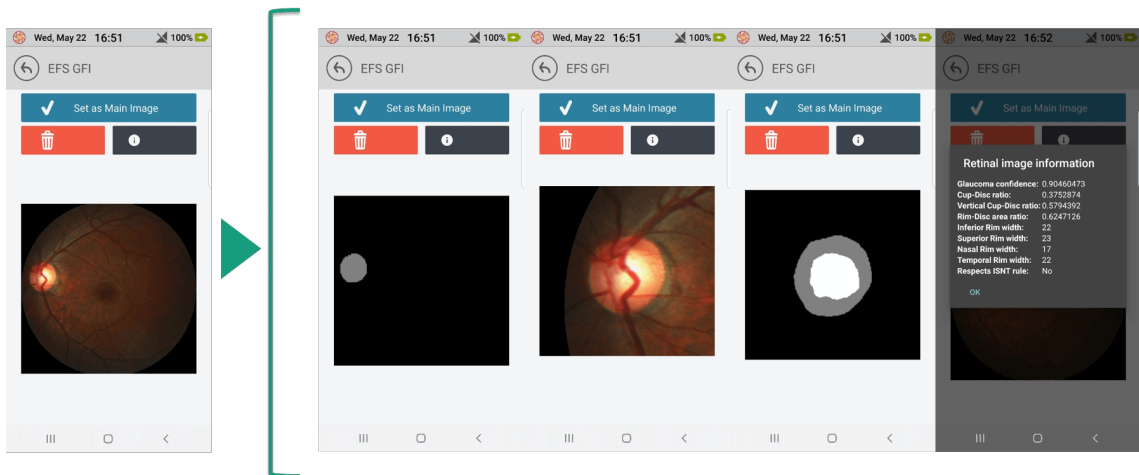
Figure 9.3: Android integration: Pipeline inference results

9.2.1 Results

After converting the developed segmentation and classification models to the TensorFlow Lite format, several strategies exist to further optimize the mobile inference times and the size of the final model.

First, the model can be kept with floating point precision or be quantized which reduces the precision to 8 bits. Quantization improves both size and time requirements of a model, however, it has a small negative effect on the model performance.

Furthermore, the model can then be run on the CPU or on the GPU, the GPU delegate has been recently introduced, it allows a speed up in time performance and since GPUs always compute at 16-bit or 32-bit floating point numbers, quantization is not advantageous anymore. Moreover, the number of threads where the model runs can also be tweaked.

In table 9.1, several times and size metrics of the models integrated into the Glaucoma assessment pipeline are exposed. The time metrics were obtained by performing 3 runs in isolation on the same mobile device and averaging times.

| Model	CPU (ms)	GPU (ms)	Size (mb)
GFI-SPP-Depth-simple	≈ 250	≈ 215	1.1
GFI-SPP-Depth-simple_quantized	≈ 280	≈ 220	0.3
GFI-ASPP-Depth	≈ 580	≈ 450	4.6
GFI-ASPP-Depth_quantized	-	-	1.7
GFI-Classification	≈ 80	≈ 25	9.0
GFI-Classification_quantized	≈ 110	≈ 120	2.3

Table 9.1: Metrics of the models used in the Glaucoma assessment pipeline (Obtained on a Samsung Galaxy S8)

By analyzing the table it is concluded that running the models on GPU is preferred than quantizing and running on the CPU since the times are better and no performance losses will occur. Moreover, multi-threaded execution was also tried, but this leads to performance variability, depending on what is being concurrently executed on the device and no improvements in the results were verified in the performed experiences, so, no more exhaustive tests were done. The selected models are highlighted in the table.

9.2.2 Issues

Tensorflow Lite is a very powerful tool, however, it still has some limitations/issues that can affect the development process. For instance, several operations are not yet supported and the error messages are not clear about what is wrong. This might be a big issue when first converting a model to the Tensorflow Lite format, for example, during this work, when first converting GFI-ASPP-Depth, an unclear error appeared. This was due to the type of interpolation used in the up-sampling layers, since bilinear was not yet supported and changing it to nearest interpolation solved the problem. Nonetheless, this was a network detail and it is not trivial to find what is causing the issue, leading to significant increases in the implementation time.

Furthermore, some models can be converted to the Tensorflow Lite format, but then have problems running in the mobile environment, this happened with GFI-ASPP-Depth_quantized, it was successfully converted to the desired format, but when running in an Android environment,

errors related to memory allocation in the Java heap occurred, that's the reason why this model doesn't have times in table 9.1.

Finally, the biggest issue of this pipeline implementation is that despite being integrated into the EFS prototype, the pipeline is intended to be used with professional fundus images, not fulfilling the ultimate goal of building a complete end-to-end system by using images acquired by the prototype.

9.3 Interpretability

As mentioned in section 4.2, the developed solution shouldn't act as a black-box, it should achieve a good degree of interpretability. This is, however, the main Achilles heel of deep learning models, since, despite having exhibited superior performances in plenty of tasks, it's difficult to make sense of these types of models when compared to machine learning approaches [ZZ18]. This mainly occurs because the feature representation is not engineered by the developer, but by the network itself, turning its translation into representative values challenging. Nowadays, deep learning interpretability is a hot topic of research [ZZ18].

Also, this topic is not only important to further understand the inner workings of a model and ease the development process, but it is also essential or even mandatory for regulations and other bureaucracies, needed in case these methods are intended to be integrated into commercial solutions, especially in a medical context.

The proposed solution has 3 interpretability measures, some of them have already been partially addressed, but below a complete description is provided.

- **Pipeline intermediate results** - These results are evidenced in figure 9.3 and were integrated into the system to allow quick confirmation of the final results, since completely wrong segmentations may eventually happen, and having the possibility to visualize those results enables the user to discard some final results right away.

- **Morphological features** - These have already been addressed multiple times along this document, explaining their meaning (section 2.2.2) and their calculation (section 7.5). In the developed solution, the joint segmentation of the optic disc and cup is only performed to power the calculation of these features and they help turn the model more interpretable since they have a defined and studied meaning among the ophthalmologist community. Despite not being directly used in the model that outputs the Glaucoma confidence level, the combination of the two output values turns the decision more trustworthy.

- **Activation Maps** - The used activation maps are gradient class activation maps (Grad-CAM) [SCD+17]. These are a generalization of class activation maps (CAM) [ZKL+15] that can be computed in CNNs with fully connected layers. In Grad-CAM, the relevant image regions for the decision are identified by using the gradient information flowing into the last convolution layer, since this layer is the last one that retains spatial information.

71

These maps were calculated in the GFI-Classification network and an illustration of the results with images from both classes can be found in figure 9.4. In the figure, the images are overlayed with the calculated Grad-CAM maps and blue tones indicate the area was not important to the classification, while redder tones indicate a big influence of that region for the final decision. By analyzing the images it is noticeable that the network is looking to the same structures that ophthalmologists inspect, like the optic disc, cup, retinal vessels topology, and more interesting, in the last two images of the Glaucoma cases a focus on the PPA region (addressed in section 2.2.2) is visible, which is an identified Glaucoma risk factor. All these insights greatly contribute to further increase the confidence in the network predictions. These maps are not available in the pipeline integration in the EFS system due to Tensorflow Lite limitations, since the library is only targeted at inference. Nonetheless, they are available in a desktop environment and can be very useful to make the pipeline more believable and help in regulatory processes.
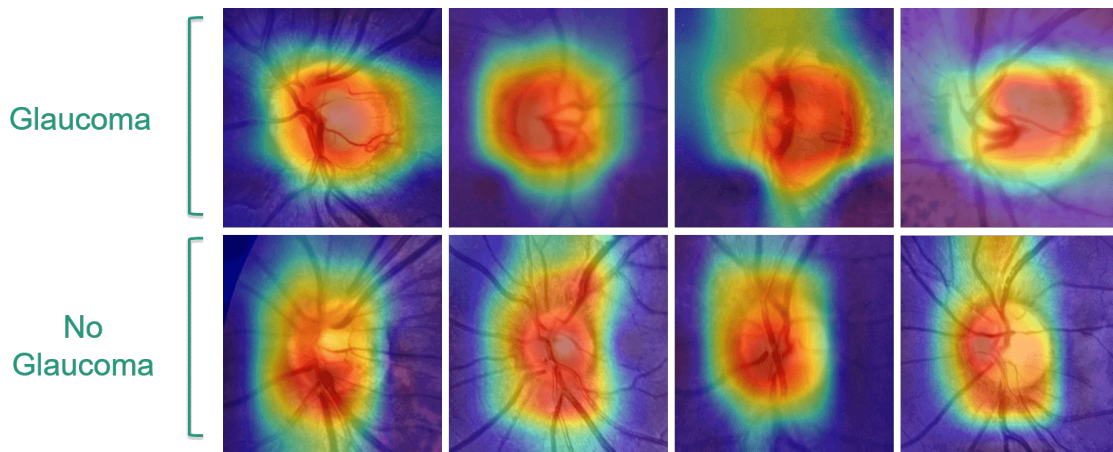
Figure 9.4: Grad-cam activation maps [SCD^{+}17] of the GFI-Classification network

9.4 Summary and Conclusions

This chapter described the final developed pipeline for Glaucoma assessment, the implementation details of this pipeline in an Android environment and the all-around solution that was built to perform Glaucoma screenings within this work.

The main goals initially set for this solution in section 4.2 have been fulfilled with the developed solution, that achieved high sensitivities, good interpretability and low latency in mobile environments.

The only missed piece is the complete integration of the pipeline with the EFS prototype, however, this was not possible since no Glaucoma annotated images of the prototype were available during the development of the present work.

Chapter 10

Conclusions and Future Work

This chapter wraps this document, reflecting about the main difficulties (section 10.1), contributions (section 10.2) and conclusions (section 10.3) of this work and then addressing the most relevant topics to enhance the developed solution in the future work section 10.4.

10.1   Main Difficulties

The first big challenge of this work consisted in gathering all the necessary public datasets. The majority of them were freely available and had easy access, nonetheless, the Origa dataset, that is probably the most popular dataset in Glaucoma research was not the case. This forced me to contact the authors of the original paper and several other authors from papers published in the last two years that have used the Origa dataset in their work. Luckily I got one answer with the dataset after two weeks of contacts and now, looking back, this was a key part for the success of this work. With regards to private datasets, despite having acquired the SCAS dataset (section 5.2), the context where it was acquired didn't allow the gathering of sufficient Glaucoma positive cases, which resulted in a considerable limitation of this work.

After collecting all the desired data, the main adversity consisted of developing segmentation networks from scratch, the possibilities are endless and the added performance concern, due to the fact that the models should run on mobile devices with reasonable inference times, turned the work even harder. It required a lot of exploration and research about the best performing blocks that could be useful for the work needs. With respect to the classification networks, the task was slightly easier due to the use of pre-trained networks that allowed relatively good performances right from the start.

Another struggle of this work, that is common to the deep learning research community, was the proper tuning of the network hyperparameters, deciding the optimal input image to feed to the network in combination with the parameters to start the training session is not trivial and things

can get out of hand really quickly, furthermore, the huge training times of these models, reaching sometimes more than one day in the segmentation networks, turns the process a lot less agile.

Lastly, as mentioned in section 9.2.2, the conversion of the models to the Tensorflow Lite format was not always trivial since some models wouldn't be able to convert and finding the root of the issue was a time-consuming task as well.

10.2 Main Contributions

The main contribution of this work is the fact that this Glaucoma assessment pipeline runs offline in Android devices in milliseconds, which to the best of my knowledge is the only approach in the Glaucoma CAD system spectrum to achieve these statements.

Moreover, this was achieved while having comparable and sometimes even better results in the segmentation and classification tasks of the developed system, contributing to the enhancement of the state-of-the-art in these topics as well.

Additionally, the focus on interpretability, with the topics addressed in section 9.3, is also a novelty, since the vast majority of techniques in the literature do not even address interpretability and the ones who do, have a more incomplete solution than the one implemented in this work.

Lastly, in order to share the findings of this work with the scientific community, a paper was submitted to the journal "Computer Methods and Programs in Biomedicine" and can be found in appendix B.

10.3 Conclusions

The main conclusion of this work is that CNNs are indeed capable of powering a robust Glaucoma CAD system that can be used in massive screening campaigns, and this structures can be used within mobile and embedded devices to further power this campaigns.

The achieved results were quite good considering the restrictions, the system isn't, however, a complete end-to-end CAD system as desired, this is due to the use of high-quality fundus images and not EFS acquired fundus images in the Glaucoma assessment pipeline. Nonetheless, this was a factor out of my control, since no Glaucoma annotated data of the prototype was available, making it impossible to build and validate complete models with these images.

Despite the aforementioned, the quality of the acquired images by the prototype is continuously increasing and a new version of the prototype with code *LISBOA-01-0247-FEDER038400* [1], is already under active development, this will decrease the gap between the quality of the images acquired by the prototype and images acquired by dedicated cameras. The previously mentioned can mean that the developed pipeline will, in the best scenario, work directly with images acquired by the EFS system images.

---

[1] https://lisboa.portugal2020.pt/np4/%7B\protect\T1\textdollarclientServletPath%7D/?newsId=47&fileName=Lista_Op_PO_LISBOA_30_04_2019_site.pdf

10.4   Future Work

In the future, the first big task to consider should be the acquisition of a big dataset of EFS images with Glaucoma annotations, to power new studies with these images and refine the existing system to be a complete end-to-end system. This would also require the exploration of new preprocessing techniques in terms of denoising and reflection removal to turn the prototype images more appropriate for the task at hand.

Additionally, multimodal and longitudinal studies could also be considered in future approaches, by using data from multiple sources and time frames to augment the existing system.

Moreover, in the addressed segmentation approaches, the resulting masks were directly used to perform the following operations in the pipeline, however, ellipse fitting could be first applied to these masks. This might positively influence the calculation of the morphological features by reducing the boundary variation in the predicted masks.

Furthermore, despite the good results on the segmentation models when considering the Glaucoma assessment pipeline, these results are still not quite enough to perform real-time segmentation. This could enable other use cases for the developed models, like, guiding the acquisition process by continuously segmenting relevant structures and providing feedback to the user about when he should take the picture or even take the picture automatically. With this in mind, further research to increase even more the performance of the developed segmentation models will bring increased value to the solution.

Conclusions and Future Work

References

[ABANATZ17] Baidaa Al-Bander, Waleed Al-Nuaimy, Majid A. Al-Taee, and Yalin Zheng. Automated glaucoma diagnosis using deep learning approach. *2017 14th International Multi-Conference on Systems, Signals and Devices, SSD 2017*, 2017-Janua:207–210, 2017.

[ABK+17] U. Rajendra Acharya, Shreya Bhat, Joel E.W. Koh, Sulatha V. Bhandary, and Hojjat Adeli. A novel algorithm to detect glaucoma risk using texton and local configuration pattern features extracted from fundus images. *Computers in Biology and Medicine*, 88(May):72–83, 2017.

[ABRL15] Ahmed Almazroa, Ritambhar Burman, Kaamran Raahemifar, and Vasudevan Lakshminarayanan. Optic Disc and Optic Cup Segmentation Methodologies for Glaucoma Image Detection: A Survey. *Journal of Ophthalmology*, 2015, 2015.

[ABWAN+18] Baidaa Al-Bander, Bryan M. Williams, Waleed Al-Nuaimy, Majid A. Al-Taee, Harry Pratt, and Yalin Zheng. Dense fully convolutional segmentation of the optic disc and cup in colour fundus for glaucoma diagnosis. *Symmetry*, 10(4), 2018.

[AFLC+19] Alfonso Antón, Gianluca Fatti, María J. Ledesma-Carbayo, Alejandra Herranz, Bart Liefers, Andrés Santos, Juan J. Gómez-Valverde, and Clara I. Sánchez. Automatic glaucoma classification using color fundus images based on convolutional neural networks and transfer learning. *Biomedical Optics Express*, 2019.

[ASA+17] Ahmed Almazroa, Weiwei Sun, Sami Alodhayb, Kaamran Raahemifar, and Vasudevan Lakshminarayanan. Optic disc segmentation for glaucoma screening system using fundus images. *Clinical Ophthalmology*, 11:2017–2029, 2017.

[BAA16] Nadeem Hafeez Butt, Muhammad Hammad Ayub, and Muhammad Hassaan Ali. Challenges in the management of glaucoma in developing countries. *Taiwan Journal of Ophthalmology*, 6(3):119–122, 2016.

[Bro19] Jason Brownlee. How to improve neural network stability and modeling performance with data scaling. https://machinelearningmastery.com/how-to-improve-neural-network-stability-and-modeling-performance-with-data-scaling/, 2019. (Accessed on 06/12/2019).

[BTF+16] Rupert R.A. Bourne, Hugh R. Taylor, Seth R. Flaxman, Jill Keeffe, Janet Leasher, Kovin Naidoo, Konrad Pesudovs, Richard A. White, Tien Y. Wong, Serge Resnikoff, and Jost B. Jonas. Number of people blind or visually impaired by glaucoma worldwide and in world regions 1990 - 2010: A meta-analysis. *PLoS ONE*, 11(10):1–16, 2016.

REFERENCES

[CBB+18]   Mark Christopher, Akram Belghith, Christopher Bowd, James A. Proudfoot, Michael H. Goldbaum, Robert N. Weinreb, Christopher A. Girkin, Jeffrey M. Liebmann, and Linda M. Zangwill. Performance of Deep Learning Architectures and Transfer Learning for Detecting Glaucomatous Optic Neuropathy in Fundus Photographs. *Scientific Reports*, 8(1):16685, 2018.

[CLX+13]   Jun Cheng, Jiang Liu, Yanwu Xu, Fengshou Yin, Damon Wing Kee Wong, Ngan Meng Tan, Dacheng Tao, Ching Yu Cheng, Tin Aung, and Tien Yin Wong. Superpixel classification based optic disc and optic cup segmentation for glaucoma screening. *IEEE Transactions on Medical Imaging*, 32(6):1019–1032, 2013.

[CPK+16]   Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. DeepLab: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully Connected CRFs. *IEEE transactions on pattern analysis and machine intelligence*, 40(4):834–848, 2016.

[CPSA17]   Liang-Chieh Chen, George Papandreou, Florian Schroff, and Hartwig Adam. Rethinking Atrous Convolution for Semantic Image Segmentation. *CoRR*, 2017.

[CTWL17]   Jun Cheng, Dacheng Tao, Damon Wing Kee Wong, and Jiang Liu. Quadratic divergence regularized SVM for optic disc segmentation. *Biomedical Optics Express*, 8(5):2687, 2017.

[CZP+18]   Liang Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 11211 LNCS:833–851, 2018.

[DVC+16]   Michal Drozdzal, Eugene Vorontsov, Gabriel Chartrand, Samuel Kadoury, and Chris Pal. The importance of skip connections in biomedical image segmentation. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 10008 LNCS:179–187, 2016.

[FAS+11]   F. Fumero, S. Alayon, J. L. Sanchez, J. Sigut, and M. Gonzalez-Hernandez. RIM-ONE: An open retinal image database for optic nerve evaluation. *IEEE Symposium on Computer-Based Medical Systems*, pages 1–6, 2011.

[FCX+18a]   Huazhu Fu, Jun Cheng, Yanwu Xu, Damon Wing Kee Wong, Jiang Liu, and Xiaochun Cao. Joint Optic Disc and Cup Segmentation Based on Multi-Label Deep Network and Polar Transformation. *IEEE Transactions on Medical Imaging*, 37(7):1597–1605, 2018.

[FCX+18b]   Huazhu Fu, Jun Cheng, Yanwu Xu, Changqing Zhang, Damon Wing Kee Wong, Jiang Liu, and Xiaochun Cao. Disc-aware Ensemble Network for Glaucoma Screening from Fundus Image. *CoRR*, pages 1–9, 2018.

[FdlR15]   F. Fumero, J. Sigut, S. Alayon, M. González-Hernández and M. González de la Rosa. Interactive Tool and Database for Optic Disc and Cup Segmentation of Stereo and Monocular Retinal Fundus Images. *Short Papers Proceedings - WSCG 2015, , Pilsen, Czech Republic,*, pages 91–97, 2015.

REFERENCES

[FUN+08]   Hiroshi Fujita, Yoshikazu Uchiyama, Toshiaki Nakagawa, Daisuke Fukuoka, Yuji Hatanaka, Takeshi Hara, Gobert N. Lee, Yoshinori Hayashi, Yuji Ikedo, Xin Gao, and Xiangrong Zhou. Computer-aided diagnosis: The emerging of three CAD systems induced by Japanese health care needs. *Computer Methods and Programs in Biomedicine*, 92(3):238–248, 2008.

[GORT18]   Parampal S. Grewal, Faraz Oloumi, Uriel Rubin, and Matthew T.S. Tennant. Deep learning in ophthalmology: a review. *Canadian Journal of Ophthalmology*, 53(4):309–313, 2018.

[Hen06]   Jeffrey D. Henderer. Disc damage likelihood scale. *British Journal of Ophthalmology*, 90(4):395–396, 2006.

[HKT+18]   Yuki Hagiwara, Joel En Wei Koh, Jen Hong Tan, Sulatha V. Bhandary, Augustinus Laude, Edward J. Ciaccio, Louis Tong, and U. Rajendra Acharya. Computer-aided diagnosis of glaucoma using fundus images: A review. *Computer Methods and Programs in Biomedicine*, 165:1–12, 2018.

[HLVW17]   Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q. Weinberger. Densely connected convolutional networks. *Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017*, 2017-Janua:2261–2269, 2017.

[HZC+17]   Andrew G. Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, and Hartwig Adam. MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications. *CoRR*, 2017.

[HZRS15a]   Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep Residual Learning for Image Recognition. *CoRR*, 2015.

[HZRS15b]   Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Spatial Pyramid Pooling in Deep Convolutional Networks for Visual Recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 37(9):1904–1916, 2015.

[HZZ+17]   Yunlong He, Yuanjie Zheng, Yanna Zhao, Yanju Ren, Jian Lian, and James Gee. Retinal Image Denoising via Bilateral Filter with a Spatial Kernel of Optimally Oriented Line Spread Function. *Computational and Mathematical Methods in Medicine*, 2017:1–13, 2017.

[Jan14]   Marina Ronzhina Jan Odstrcilik, Radim Kolar, Ralf-Peter Tornow, Jiri Jan, Attila Budai, Markus Mayer, Martina Vodakova, Robert Laemmer, Martin Lamos, Zdenek Kuna, Jiri Gazarek, Tomas Kubena, Pavel Cernosek. Thickness related textural properties of retinal nerve fiber layer in color fundus images. *Computerized Medical Imaging and Graphics*, 38(6):508–516, 2014.

[Jon05]   Jost B. Jonas. Clinical implications of peripapillary atrophy in glaucoma. *Current Opinion in Ophthalmology*, 16(2):84–88, 2005.

[Joy17]   Joy Chan. What is Glaucoma? https://drjoychan.com/2017/02/08/what-is-glaucoma/, February 2017.

REFERENCES

[JSK11]   Gopal Datt Joshi, Jayanthi Sivaswamy, and S. R. Krishnadas. Optic disk and cup segmentation from monocular color retinal images for glaucoma assessment. *IEEE Transactions on Medical Imaging*, 30(6):1192–1205, 2011.

[KB14]   Diederik P Kingma and Jimmy Ba. Adam: A Method for Stochastic Optimization. *CoRR*, 2014.

[KCF+16]   Venediktos V. Kapetanakis, Michelle P.Y. Chan, Paul J. Foster, Derek G. Cook, Christopher G. Owen, and Alicja R. Rudnicka. Global variations and time trends in the prevalence of primary open angle glaucoma (POAG): A systematic review and meta-analysis. *British Journal of Ophthalmology*, 100(1):86–93, 2016.

[LHK+18]   Zhixi Li, Yifan He, Stuart Keel, Wei Meng, Robert T. Chang, and Mingguang He. Efficacy of a Deep Learning System for Detecting Glaucomatous Optic Neuropathy Based on Color Fundus Photographs. *Ophthalmology*, 125(8):1199–1206, 2018.

[LJC+17]   June Goo Lee, Sanghoon Jun, Young Won Cho, Hyunna Lee, Guk Bae Kim, Joon Beom Seo, and Namkug Kim. Deep learning in medical imaging: General overview. *Korean Journal of Radiology*, 18(4):570–584, 2017.

[LWD+06]   Paul P. Lee, John G. Walt, John J. Doyle, Sameer V. Kotak, Stacy J. Evans, Donald L. Budenz, Philip P. Chen, Anne L. Coleman, Robert M. Feldman, Henry D. Jampel, L. Jay Katz, Richard P. Mills, Jonathan S. Myers, Robert J. Noecker, Jody R. Piltz-Seymour, Robert R. Ritch, Paul N. Schacknow, Janet B. Serle, and Gary L. Trick. A multicenter, retrospective pilot study of resource use and costs associated with severity of disease in glaucoma. *Archives of Ophthalmology*, 124(1):12–19, 2006.

[MFT+18]   Keita Mitsuhashi, Yuri Fujino, Masaki Tanito, Naoto Shibata, Hiroshi Murata, Ryo Asaoka, and Masato Matsuura. Development of a deep residual learning algorithm to screen for glaucoma from fundus photography. *Scientific Reports*, 8(1):1–9, 2018.

[MHRUG+16]   Carmen Mendez-Hernandez, Ignacio Rodriguez-Uña, Manuel Gonzalez-de-la Rosa, Paula Arribas-Pardo, and Julian Garcia-Feijoo. Glaucoma diagnostic capacity of optic nerve head haemoglobin measures compared with spectral domain OCT and HRT III confocal tomography. *Acta Ophthalmologica*, 94(7):697–704, 2016.

[MHS+11]   Chisako Muramatsu, Yuji Hatanaka, Akira Sawada, Tetsuya Yamamoto, and Hiroshi Fujita. Computerized detection of peripapillary chorioretinal atrophy by texture analysis. *Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBS*, pages 5947–5950, 2011.

[MNA16]   Fausto Milletari, Nassir Navab, and Seyed Ahmad Ahmadi. V-Net: Fully convolutional neural networks for volumetric medical image segmentation. *Proceedings - 2016 4th International Conference on 3D Vision, 3DV 2016*, pages 565–571, 2016.

[Mor11]   Jafet Morales. *Assessment of Iris Reflection Artifacts and Alignment in Fundus Images*. PhD thesis, Faculty of the Graduate School of St. Mary's University, 07 2011.

REFERENCES

[MPA17]   Shishir Maheshwari, Ram Bilas Pachori, and U. Rajendra Acharya. Automated Diagnosis of Glaucoma Using Empirical Wavelet Transform and Correntropy Features Extracted from Fundus Images. *IEEE Journal of Biomedical and Health Informatics*, 21(3):803–813, 2017.

[MWW+17]   Riccardo Miotto, Fei Wang, Shuang Wang, Xiaoqian Jiang, and Joel T. Dudley. Deep learning for healthcare: review, opportunities and challenges. *Briefings in Bioinformatics*, pages 1–11, 2017.

[MZZS18]   Ningning Ma, Xiangyu Zhang, Hai Tao Zheng, and Jian Sun. Shufflenet V2: Practical guidelines for efficient cnn architecture design. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 11218 LNCS:122–138, 2018.

[Pra18]   Prasad Netralaya. What is Glaucoma? Check for this one symptom. http://prasadnetralaya.com/blog/2018/03/29/what-is-Glaucoma-check-for-this-one-symptom, 2018.

[QB06]   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.

[RFB15]   Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 9351:234–241, 2015.

[RFB+18]   U. Raghavendra, Hamido Fujita, Sulatha V. Bhandary, Anjan Gudigar, Jen Hong Tan, and U. Rajendra Acharya. Deep convolution neural network for accurate diagnosis of glaucoma using digital fundus images. *Information Sciences*, 441:41–49, 2018.

[SBS+18]   Tanzila Saba, Syedia Tahseen Fatima Bokhari, Muhammad Sharif, Mussarat Yasmin, and Mudassar Raza. Fundus image classification methods for the detection of glaucoma: A review. *Microscopy Research and Technique*, pages 1–17, 2018.

[SCD+17]   Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. *Proceedings of the IEEE International Conference on Computer Vision*, 2017-Octob:618–626, 2017.

[Sci14]   Clinical Sciences. The ISNT Rule and Differentiation of Normal From Glaucomatous Eyes. *JAMA Ophthalmology*, 124(11):1579–1583, 11 2014.

[SDP+16]   Anushikha Singh, Malay Kishore Dutta, M. ParthaSarathi, Vaclav Uher, and Radim Burget. Image processing based automatic diagnosis of glaucoma using wavelet features of segmented optic disc from fundus image. *Computer Methods and Programs in Biomedicine*, 124:108–120, 2016.

[SGA16]   Sanjivani Shantaiya, Shruti Gorasia, and Rida Anwar. Early Detection of Glaucoma Using Retinal Fundus Images. *Imperial Journal of Interdisciplinary Research (IJIR)*, 2(6):1525–1528, 2016.

REFERENCES

[SHO18]   Clark H. Stevenson, Sheng Chiong Hong, and Kelechi C. Ogbuehi. Development of an artificial intelligence system to classify pathology and clinical features on retinal fundus images. *Clinical and Experimental Ophthalmology*, 2018.

[Shr16]   Rida Anwar Shruti Gorasia. A Review Paper on Detection of Glaucoma using Retinal Fundus Images. *International Journal for Research in Applied Science & Engineering Technology*, 4(I):166–170, 2016.

[SHZ+18]   Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. MobileNetV2: Inverted Residuals and Linear Bottlenecks. *CoRR*, 2018.

[SKD+14]   Jayanthi Sivaswamy, S. R. Krishnadas, Gopal Datt Joshi, Madhulika Jain, and A. Ujjwaft Syed Tabish. Drishti-GS: Retinal image dataset for optic nerve head(ONH) segmentation. *2014 IEEE 11th International Symposium on Biomedical Imaging (ISBI)*, pages 53–56, 2014.

[SSS+19]   Sonali, Sima Sahu, Amit Kumar Singh, S. P. Ghrera, and Mohamed Elhoseny. An approach for de-noising and contrast enhancement of retinal fundus image using CLAHE. *Optics and Laser Technology*, 110:87–98, 2019.

[SVI+15]   Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jonathon Shlens, and Zbigniew Wojna. Rethinking the Inception Architecture for Computer Vision. *CoRR*, abs/1512.00567, 2015.

[SZ15]   Karen Simonyan and Andrew Zisserman. VERY DEEP CONVOLUTIONAL NETWORKS FOR LARGE-SCALE IMAGE RECOGNITION. *arXiv 1409.1556*, 2015.

[TJ18]   Niharika Thakur and Mamta Juneja. Survey on segmentation and classification approaches of optic cup and optic disc for diagnosis of glaucoma. *Biomedical Signal Processing and Control*, 42:162–189, 2018.

[TLP11]   Ravi Thomas, Klaus Loibl, and Rajul Parikh. Evaluation of a glaucoma patient. *Indian Journal of Ophthalmology*, 59(7):43, 2011.

[TM98]   C. Tomasi and R. Manduchi. Bilateral filtering for gray and color images. In *Proceedings of the Sixth International Conference on Computer Vision*, ICCV '98, pages 839–, Washington, DC, USA, 1998. IEEE Computer Society.

[VLGK11]   Rohit Varma, Paul P. Lee, Ivan Goldberg, and Sameer Kotak. An assessment of the health and economic burdens of glaucoma. *American Journal of Ophthalmology*, 152(4):515–522, 2011.

[VRN16]   M. Caroline Viola Stella Mary, Elijah Blessing Rajsingh, and Ganesh R. Naik. Retinal Fundus Image Analysis for Diagnosis of Glaucoma: A Comprehensive Survey. *IEEE Access*, 4:4327–4354, 2016.

[Wik19]   Wikipedia. Polar coordinate system. https://en.wikipedia.org/wiki/Polar_coordinate_system, 2019. (Accessed on 06/12/2019).

REFERENCES

[XDL+14]   Yanwu Xu, Lixin Duan, Stephen Lin, Xiangyu Chen, Damon Wing, and Kee Wong. Optic Cup Segmentation for Glaucoma Detection Using Low-Rank Superpixel Representation. *Med Image Comput Comput Assist Interv*, 17(Pt 1):788–795, 2014.

[YLO+11]   Fengshou Yin, Jiang Liu, Sim Heng Ong, Ying Sun, Damon W.K. Wong, Ngan Meng Tan, Carol Cheung, Mani Baskaran, Tin Aung, and Tien Yin Wong. Model-based optic nerve head segmentation on retinal fundus images. *Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBS*, pages 2626–2629, 2011.

[ZFZ+18]   Xin Zhao, Zhun Fan, Beiji Zou, Xuanchu Duan, Bin Xie, Yuxiang Mai, and Fan Guo. Yanbao: A Mobile App Using the Measurement of Clinical Parameters for Glaucoma Screening. *IEEE Access*, 6:77414–77428, 2018.

[ZKL+15]   Bolei Zhou, Aditya Khosla, Àgata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. *CoRR*, abs/1512.04150, 2015.

[ZSQ+17]   Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. Pyramid scene parsing network. *Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017*, 2017-Janua:6230–6239, 2017.

[Zui94]   Karel Zuiderveld. Graphics gems iv. In Paul S. Heckbert, editor, *Graphics Gems IV*, chapter Contrast Limited Adaptive Histogram Equalization, pages 474–485. Academic Press Professional, Inc., San Diego, CA, USA, 1994.

[ZYC+18]   Rui Zhao, Ruqiang Yan, Zhenghua Chen, Kezhi Mao, Peng Wang, and Robert X. Gao. Deep learning and its applications to machine health monitoring. *Mechanical Systems and Signal Processing*, 115:213–237, 2018.

[ZYL+10]   Zhuo Zhang, Feng Shou Yin, Jiang Liu, Wing Kee Wong, Ngan Meng Tan, Beng Hai Lee, Jun Cheng, and Tien Yin Wong. ORIGA-light: An online retinal fundus image database for glaucoma analysis and research. *2010 Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBC'10*, pages 3065–3068, 2010.

[ZZ18]   Quanshi Zhang and Song-Chun Zhu. Visual Interpretability for Deep Learning: a Survey. *CoRR*, 19(1423305):27–39, 2018.

REFERENCES

Appendix A

SCAS informed consent

CONSENTIMENTO INFORMADO, LIVRE E ESCLARECIDO PARA PARTICIPAÇÃO EM INVESTIGAÇÃO

de acordo com a Declaração de Helsínquia[1] e a Convenção de Oviedo[2]

Por favor, leia com atenção a seguinte informação. Se achar que algo está incorreto ou que não está claro, não hesite em solicitar mais informações. Se concorda com a proposta que lhe foi feita, queira assinar este documento.

Título do estudo: Avaliação de um sistema móvel de aquisição de imagens da retina e obtenção de dados para suporte a estudos de investigação exploratórios acerca da previsão do risco cardiovascular ou doenças oculares.

Enquadramento: Este estudo de investigação está a ser promovido pela Associação Fraunhofer Portugal Research – Fraunhofer Portugal Research Center for Assistive Information and Communication Solutions (Fraunhofer Portugal AICOS) no âmbito do projeto EyeFundusScope que tem como coordenador científico e investigador principal o Doutor Filipe Soares. Este estudo contribuirá também para a recolha de dados a serem utilizados em dissertações de Mestrado em Engenharia Informática da Faculdade de Engenharia da Universidade do Porto associadas à Fraunhofer Portugal AICOS.

Explicação do estudo: O estudo de investigação acima mencionado tem como finalidade avaliar um sistema móvel para aquisição de imagens da retina em termos de qualidade de imagem e adequabilidade ao fim de diagnóstico de doenças oculares. Um outro objetivo do estudo é obter dados que suportem estudos exploratórios, que irão investigar a relação entre as imagens da retina e o risco cardiovascular. A sua participação neste estudo consiste na recolha de imagens da retina usando um protótipo desenvolvido pelo centro de investigação Fraunhofer Portugal AICOS. A recolha das imagens da retina tem uma duração de cerca de 5 minutos para os dois olhos, é indolor e não representa qualquer risco ou desconforto. As imagens serão gravadas, de forma anonimizada, para posterior análise; o sistema não fornece no imediato qualquer resultado de análise das imagens e estes resultados não serão lhe comunicados. Pedimos também que responda a um questionário/entrevista, que terá uma duração prevista de 5 minutos, com o objetivo de recolher dados demográficos, clínicos e laboratoriais relacionados com o risco cardiovascular.

Condições e financiamento: Não serão necessárias deslocações adicionais no âmbito deste estudo. Não haverá lugar a qualquer pagamento pela sua participação. Este estudo é financiado pelos programas Norte 2020, Portugal 2020 e pela União Europeia. A sua participação neste estudo é voluntária, podendo em qualquer altura cessá-la sem qualquer tipo de consequência.

Confidencialidade e anonimato: As informações recolhidas são confidenciais e a sua identificação não será revelada. Os dados recolhidos serão utilizados exclusivamente no âmbito dos objetivos do presente estudo. Os dados serão recolhidos e analisados pelos investigadores identificados abaixo ou por outro elemento com autorização delegada pelo investigador principal e serão apresentados em publicações científicas e conferências ou outro tipo de evento científico ou de divulgação do projeto, assim como nos diversos meios de comunicação social. A sua identidade não será revelada em qualquer situação. Os dados pessoais, assim como as respostas aos questionários/entrevistas, serão mantidos numa base de dados, de forma anonimizada.

Gostaríamos de contar com a sua participação. Agradecemos muito o seu contributo, fundamental para a nossa investigação!

Filipe Soares

Investigador Sénior na Fraunhofer Portugal AICOS

Sandra Ferreira

Presidente da Associação de estudantes da Escola Superior de Saúde de Coimbra

---

[1]   http://portal.arsnorte.min-saude.pt/portal/page/portal/ARSNorte/Comiss%C3%A3o%20de%20%C3%89tica/Ficheiros/Declaracao_Helsinquia_2008.pdf

[2]   http://dre.pt/pdf1sdip/2001/01/002A00/00140036.pdf

Coordenador científico e Investigador Principal:

Nome: Filipe Soares

Instituição: Associação Fraunhofer Portugal Research – Fraunhofer Portugal Research Center for Assistive Information and Communication Solutions

E-mail: filipe.soares@fraunhofer.pt

Assinatura: _Filipe Cruz Gomes Soares_____

-o-o-o-o-o-o-o-o-o-o-o-o-o-o-o-o-o-o-

Declaro ter lido e compreendido este documento, bem como as informações verbais que me foram fornecidas pela/s pessoa/s que acima assina/m. Foi-me garantida a possibilidade de, em qualquer altura, recusar participar neste estudo sem qualquer tipo de consequências. Desta forma, aceito participar neste estudo e permito a utilização dos dados que de forma voluntária forneço, confiando em que apenas serão utilizados para esta investigação e nas garantias de confidencialidade e anonimato que me são dadas pelo/a investigador/a.

Nome: _____

Assinatura: _____

Data: ____ /_____ /_____

SE NÃO FOR O PRÓPRIO A ASSINAR POR IDADE OU INCAPACIDADE

(se o menor tiver discernimento deve também assinar em cima, se consentir)

NOME: _____

BI/CD Nº: _____ DATA OU VALIDADE _____ /_____ /_____

GRAU DE PARENTESCO OU TIPO DE REPRESENTAÇÃO: _____

ASSINATURA: _____

ESTE DOCUMENTO É COMPOSTO DE 2 PÁGINAS E FEITO EM DUPLICADO:

UMA VIA PARA O/A INVESTIGADOR/A, OUTRA PARA A PESSOA QUE CONSENTE

SCAS informed consent

Appendix B

Submitted paper

Offline computer-aided diagnosis for Glaucoma detection using fundus images targeted at mobile devices

José Martins[a], Jaime S. Cardoso[b], Filipe Soares[a,*]

[a]*Fraunhofer Portugal AICOS, Rua Alfredo Allen 455/461, 4200-135 Porto, Portugal*
[b]*INESC TEC and Faculty of Engineering of the University of Porto, Portugal*

Abstract

Background and Objective: Glaucoma, an eye condition which leads to permanent blindness, is typically asymptomatic and therefore difficult to be diagnosed in time. However, if diagnosed in time, Glaucoma can effectively be slowed down by using adequate treatment; hence, an early diagnosis is of utmost importance. Nonetheless, the conventional approaches to diagnose Glaucoma adopt expensive and bulky equipment that requires qualified experts, making it difficult, costly and time-consuming to diagnose large amounts of people. Consequently, new alternatives to diagnose Glaucoma that suppress these issues should be explored.

Methods: This work proposes an interpretable computer-aided diagnosis (CAD) pipeline that is capable of diagnosing Glaucoma using fundus images and run offline in mobile devices. Several public datasets of fundus images were merged and used to build Convolutional Neural Networks (CNNs) that perform segmentation and classification tasks. These networks are then used to build a pipeline for Glaucoma assessment that outputs a Glaucoma confidence level and also provides several morphological features and segmentations of relevant structures, resulting in an interpretable Glaucoma diagnosis. To assess the performance of this method in a restricted environment, this pipeline was integrated into a mobile application and time and space complexities were assessed.

Results: Considering the test set, the developed pipeline achieved 0.91 and 0.75 of Intersection over Union (IoU) in the optic disc and optic cup segmentation, respectively. With regards to the classification, an accuracy of 0.87 with a sensitivity of 0.85 and an AUC of 0.93 were attained. Moreover, this pipeline runs on an average Android smartphone in under two seconds.

Conclusions: The results demonstrate the potential that this method can have in the contribution to an early Glaucoma diagnosis. The proposed approach achieved similar or slightly better metrics than the current CAD systems for Glaucoma assessment while running on more restricted devices. This pipeline can, therefore, be used to construct accurate and affordable CAD systems that could enable large Glaucoma screenings, contributing to an earlier diagnose of this condition.

Keywords: Computer-aided diagnosis (CAD), Fundus Images, Glaucoma, Deep Learning

1. Introduction

Glaucoma is a group of chronic eye diseases and a leading cause of irreversible blindness worldwide [1].

The rise of intraocular pressure (IOP) inside the eye is considered the main cause of Glaucoma. This pressure damages the optic nerve and is usually related to an inability of the eye to properly manage the balance between the amount of fluid that is produced and the amount that is drained [2]. Although it is confirmed that the raised intraocular pressure is a causal risk factor for Glaucoma and the only one that can be treated, it is neither sufficient nor necessary for the diagnosis of the disease [3].

In the majority of Glaucoma cases, no early symptoms or pain occur, and for that reason, Glaucoma is often called a 'silent thief of sight'.

Furthermore, the usual approaches to diagnose Glaucoma are performed by ophthalmologists and

---

*Corresponding author

Email address: `filipe.soares@fraunhofer.pt` (Filipe Soares)

consist of a comprehensive eye examination. This encompasses an external examination of the eye, ocular mobility, examination of the pupil, slit lamp examination, tonometry, gonioscopy and perimetry [3]. These methods require trained professionals and expensive equipment, therefore, due to the high costs incurred by them, some are only performed if Glaucoma is already suspected, which contributes to an increasing number of undiagnosed Glaucoma cases.

This number of undiagnosed cases is a big concern in the healthcare community since, despite being irreversible, Glaucoma can be effectively slowed down with appropriate treatment, pharmaceutical or surgical [4], therefore reducing the risk of total blindness. Hence, achieving an early diagnosis of Glaucoma is of utmost importance. This results in the demand for new diagnose techniques that aid or replace the existing ones, turning Glaucoma assessment into a more efficient process.

To respond to this demand, fundus images can be used. These images, illustrated in figure 1, have the necessary morphological features to diagnose several eye conditions, like Glaucoma, the topic of this work.
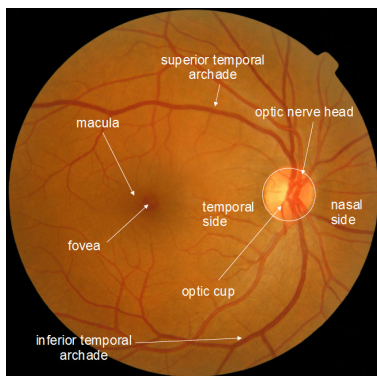

Figure 1: Fundus image [5].

The region of interest (ROI), illustrated in figure 2, shows the most relevant structures for Glaucoma diagnosis present in fundus images. These structures are the optic disc and optic cup.

This work contributes to improve Glaucoma management by proposing an interpretable pipeline for Glaucoma detection powered by fundus images that can run offline on mobile devices and achieves better or comparable results to the state-of-the-art techniques in classification and segmentation tasks.
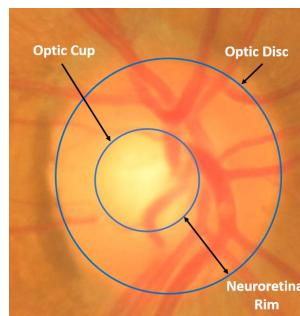

Figure 2: Fundus Image ROI [6].

2. Related work

The first relevant implementations of CAD systems for Glaucoma detection with fundus images started to arise in 2008 [2]. Since then a significant collection of approaches have been proposed.

In 2016, A. Singh et al.[7] achieved an accuracy of 0.947. They started by identifying the center of the disc, performing the disc segmentation afterward. Then, the intuition that blood vessels represent noisy pixels that affect the performance of the system led to the removal of these vessels from the obtained optic disc images. After this step, feature extraction was accomplished using first level discrete wavelet decomposition that resulted in a feature vector of 18 features. For feature selection, two approaches were tried, genetic algorithms and Principal Component Analysis (PCA). The final feature vectors of each technique were then tested with several classifiers (Support Vector Machine (SVM), K-Nearest Neighbors (KNN), Random Forest, Naive Bayes, and Artificial Neural Networks (ANN)). Since the fundus images are from a local dataset of 63 images, which is rather small, leave-one-out cross-validation was also performed to account for overfitting issues. The best performing models were the SVM and KNN, with PCA feature selection that resulted in only 2 components. Both of these classifiers obtained 0.947 accuracy.

More recently, Deep Learning approaches are proving to be capable of surpassing the existing techniques and are being heavily applied.

H. Fu et al. [1] proposed a new segmentation approach of the optic disc and cup. The novelty of this work consists in the joint segmentation of the optic disc and optic cup since most approaches achieve this segmentation separately. To deal with this joint segmentation the work also proposes a

2

new joint loss metric. The first step of this method is the transformation of the fundus image to the polar coordinate system, in order to enlarge the cup representation. Next, the segmentation is performed with a CNN, this consists of four key parts, the multi-scale input layer, U-shape architecture, a side output layer, and the mentioned joint loss metric. Several tests were performed in different conditions and it was concluded that the polar transformation attains a relevant performance gain. Glaucoma screening was then performed calculating the cup to disc ratio value and an AUC of 0.899 in the private SCES dataset was obtained.

A more recent approach by Z. li et al. [8] uses the Inception-v3 architecture, that has already proven its success in other tasks. The images were resized to 229x229 and the RGB values were normalized between 0 and 1 before feeding the images to the network. A minibatch gradient descent of 32 was used in combination with the ADAM optimizer and a 0.002 learning rate. The key aspect of this approach is its dataset since a total of 70000 fundus images were downloaded from an unlabeled dataset and 48116 of those images were selected after verifying that they include a visible optic disc. After this, 21 ophthalmologists performed the classification of these images in an online platform and multiple interpretations of each image were obtained for quality control. The result was 39745 images, where 31745 were used as a training set and 8000 as a validation set. The system achieved an AUC of 0.986, 0.956 sensitivity and 0.920 specificity, which demonstrates the relevance of the dataset size and provides further evidence about the success of deep learning techniques for Glaucoma detection.

These approaches, however, do not target mobile devices and despite the better results achieved in the last approaches with deep learning techniques, these solutions have little or no concerns in respect to the interpretability of the decision.

3. Proposed methodology

The proposed methodology explores deep learning techniques to build a Glaucoma assessment pipeline with a focus on interpretability and space and time complexities.

3.1. Datasets

To construct a robust dataset for the two main tasks of this work, segmentation and classification, several publicly available datasets, listed in table 1, were merged.

Dataset	G	NG	Total
Origa [9]	168	482	650
Drishti-GS [10]	70	31	101
RIM-ONE r1 [11]	40	118	158
RIM-ONE r2 [12]	200	255	455
RIM-ONE r3 [12]	148	170	318
iChallenge [13]	80	720	800
RIGA [14]	749		749

Table 1: Image count per dataset. G and NG represent Glaucoma and No-Glaucoma, respectively.

Firstly, a dataset for segmentation tasks of both the optic disc and optic cup was created merging the Origa, Drishti-GS, iChallenge, RIM-ONE r3, and RIGA datasets, resulting in a total of 2618 images. In the datasets where several annotations were provided for the same image, the considered ground truth was the region of agreement between the annotators.

Lastly, for the classification task, Origa, Drishti-GS, RIM-ONE (r1,r2,r3) and iChallenge were merged, resulting in 2482 images. Furthermore, only two classes were considered, Glaucoma and No-Glaucoma, since most datasets only included a binary classification for Glaucoma.

Additionally, when considering the RIM-ONE r3 dataset, which consisted of stereo images, each of these images was divided into two, and considered as a separate case. With this in mind, the number of images in this dataset increased from 159 to 318.

3.2. Augmentation

Real-time data augmentation was used since it is robust and easy to implement, not requiring additional disk space to save the augmented images and providing different images in every batch of data.

The augmentation pipeline consists of four main steps that are applied in random order and from augmentation to augmentation some of them might even be skipped. The first step consists in the application of blur to the image, this blur can be either a gaussian, average or a median blur. Secondly, contrast normalization is performed by moving the pixel values away or closer to 128. Then, changes in the brightness and sharpness of the images are also employed.

3.3. Segmentation

When considering Glaucoma, the optic disc and optic cup are the most important structures to segment, since they can power the calculation of several morphological features and the crop of the ROI in the full fundus image.

To target this task, an architecture entitled GFI-ASPP-Depth was developed. It has a U-shaped structure where the encoding path consists of four depth levels.

In each depth level, two depthwise separable convolution blocks, inspired in the MobileNet architectures [15, 16], are employed with 3x3 kernels, the number of filters starts at 32 and is then doubled on every transition to the next depth level.

The transition to the next depth level is performed through average pooling with a pool size of 2. Average pooling was selected in detriment of max pooling since the literature reports better results on similar tasks [17]. Furthermore, two inputs are used in the network in the first two depth levels and skip connections between the encoder and decoder path are present at every depth level to provide more spatial context to the decoding path.

When it comes to the transition from the encoder to the decoder path, an ASPP module [18, 19, 20] was chosen, consisting of four parallel padded atrous convolutions with dilation rates of 1, 2, 4 and 7. Additionally, image-level features obtained by performing a global average pooling on the original feature map are also used. This is done because when the dilation rate increases, the number of filters weights that are applied to the valid feature region (region without the padded zeros) declines. The inclusion of image-level features helps to mitigate this problem and include more global context in the network [19].

Then, the result of each level of the ASPP module and the image-level features are concatenated and the decoding path starts. This is very similar to the encoding path, however, in here, the transition to an upper level leads to the decrease in the number of filters in each convolution by a factor of 2 and is done with nearest interpolation. Finally, in the first two depth levels of the decoder, two outputs are generated and then averaged to form the final segmentation prediction.

This architecture consists of a total of 1,152,131 parameters and is represented in figure 3.

Additionally, given that the segmentation of the optic disc is a simpler task than the joint disc/cup segmentation or the cup segmentation alone, a smaller architecture entitled GFI-SPP-Depth-simple was created. In here, the starting number of filters in the first depth layer is 16 instead of 32 and is then increased and decreased in the same way as the GFI-ASPP-Depth. Moreover, the transition between the encoder and decoder path consists solely in the concatenation of global average pooling performed on the original feature map, the result of a convolution block applied to the original feature map and the original feature map itself. Resulting in an architecture with 271,929 parameters.

Both networks were trained after resizing the images to 224x224 and performing data normalization. The multi-label dice loss was selected [1], together with the ADAM optimizer with a learning rate of 0.0001. Additionally, the learning rate was reduced during the learning process by a factor of 2 if the validation loss was not improved in the last 10 epochs. Furthermore, the model was trained for 200 epochs with early stopping if no improvements are verified in the validation loss after 20 epochs.

Considering GFI-ASPP-Depth in particular, the input images fed to the network consist of the ROI images with the application of CLAHE [21] and the model was trained with a batch size of 16. This model was used for joint segmentation, of the optic disc and optic cup, both with 0.5 weights in the multi-label dice loss.

In the GFI-SPP-Depth-simple case, a batch size of 16 was again used, but the input images were the center cropped full fundus images. This network was solely used for disc segmentation.

The validation approach consists of a stratified train/val/test split with 80/10/10 proportions of the original dataset. The achieved results of both networks on the test set are exposed in table 2, along with the results of other methods.

One issue when comparing these models with the literature is that the used datasets and, consequently, the validation approaches often differ quite a lot, making a direct comparison unfair in some cases. To mitigate this issue, the M-net model [1] without polar transformation (PT), was trained with the segmentation dataset used in the developed models. Afterward, the results of this model can be directly compared with GFI-ASPP-Depth and GFI-SPP-Depth-Simple.

Furthermore, the state-of-the-art results extracted from [1], compare conventional methodologies with deep learning approaches and are com-
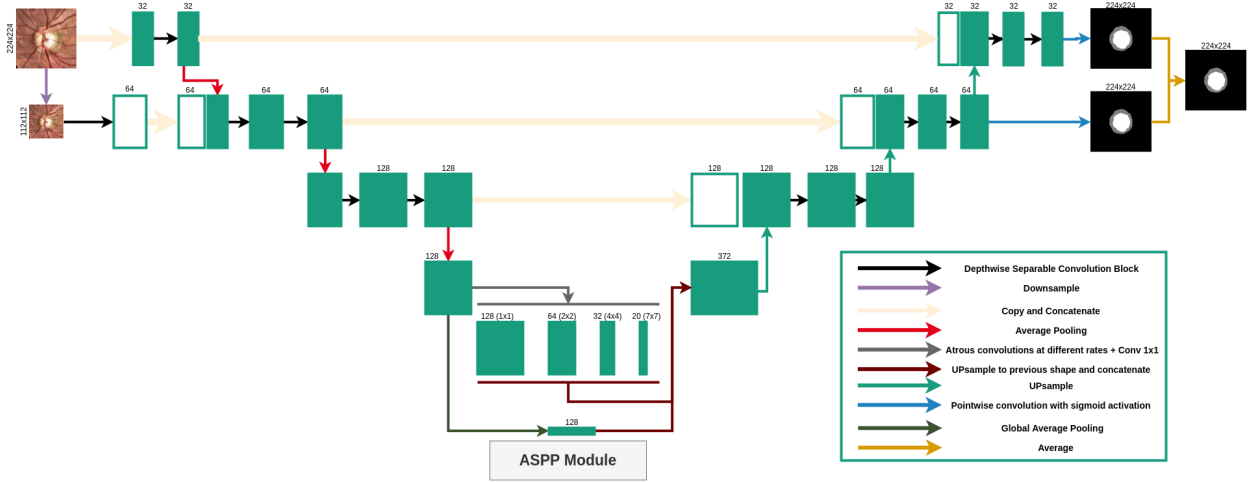
4

Figure 3: GFI-ASPP-Depth architecture diagram.

Method	IoU Disc	IoU Cup
Transcribed from [1]		
R-Bend [22]	0.871	0.605
ASM [23]	0.882	0.687
Superpixel [24]	0.898	0.736
LRR [25]	-	0.756
QDSVM [26]	0.89	-
U-net [27]	0.885	0.713
M-net [1]	0.917	0.715
M-net + PT [1]	0.929	0.77
Achieved results		
M-net (self-trained) [1]	0.87	0.70
GFI-ASPP-Depth	0.91	0.74
GFI-SPP-Depth-Simple	0.89	-

Table 2: Segmentation performance comparison with state-of-the-art. Results obtained on the test set.

parable between themselves. Hence, by comparing the obtained M-net result with the result in the referred paper and then with the other state-of-the-art methods, it is possible to establish a point of comparison between all models.

The results indicate that the developed model outperforms the most recent methods while being considerably faster and smaller. For instance, in comparison with the M-net model, the number of parameters is around 7 times smaller in GFI-ASPP-Depth and 31 times smaller GFI-ASPP-Depth-Simple.

3.4. Morphological features

The segmentation predictions of the optic disc and cup obtained by the developed models can then be used to calculate several morphological features.

In this work, the cup to disc ratio was calculated based on both the vertical length (VCDR) and the area of the cup and the disc (CDR). The enlargement of this ratio is known as 'cupping' [2] and can be classified into mild (CDR up to 0.4), moderate (CDR between 0.5 and 0.7) and severe (CDR above 0.7) Glaucoma classes [4]. Rim to disc area ratio (RDAR), was also calculated as it also provides an interpretation of the optic nerve head shape.

Furthermore, the ISNT rule states that, in normal eyes, the Neuro-Retinal Rim (NRR) width in the inferior section (I) of the ROI is greater than in the superior section (S) which in turn is greater than the area in the nasal section (N) which in turn is greater than the area in the temporal section (T);

this can be translated by the following inequation.

$$I > S > N > T \tag{1}$$

It has been declared that the compliance of this rule is useful in differentiating normal from Glaucomatous optic nerves and is not affected by race [28]. Therefore, the ISNT values and rule compliance were also computed.

3.5. Classification

To obtain a Glaucoma confidence level, a classification network, entitled GFI-Classification, was created using MobileNetV2 [16] feature extractor as a backbone.

After the MobileNetV2 block, global average pooling is performed to flat the feature maps and then two fully connected layers, interleaved by heavy dropouts, are performed, resulting in an architecture with 2,299,521 parameters.

To better assess the impact of the MobileNetV2 backbone, other backbones have also been experimented, namely, VGG16 and VGG19 [29], InceptionV3 [30] and ResNet50 [31].

All the models were trained with a batch size of 64, a learning rate of 0.00001, and resized 224x224 ROI cropped images with CLAHE transformation as input. Furthermore, callbacks to reduce the learning rate by a factor of 2 after 10 epochs and interrupt the training session after 20 epochs, if no improvements are achieved in the validation loss, were also used. With regards to the backbones, ImageNet weights were used on all networks and all the layers were trained.

Table 3 discloses the achieved results on the test set and the results of other literature methods. The proposed approach accomplishes very similar metrics to the most recent method (DL 6), that consists of a network with around 30 million parameters, and is, therefore, more suitable for integration in computational restricted devices.

3.6. CAD Pipeline

The developed segmentation and classification models along with the calculation of the morphological features were then assembled to construct an interpretable Glaucoma CAD pipeline.

The pipeline, illustrated in figure 4, starts with an acquired full fundus image; this image is then center cropped and the GFI-SPP-Depth-simple model is employed to segment the disc in step '(1)'.

This segmentation is then used to transform the full fundus image in a ROI image during step '(2)', followed by the CLAHE transformation. The CLAHE ROI image is then utilized for two different tasks that build the pipeline outputs. First, in step '(3)', GFI-ASPP-Depth is used to obtain an optic disc and cup segmentation. This powers the calculation of several morphological features, that contribute to the interpretability of the decision. Additionally, step '(5)', consists of running the images outputted from step '(2)' through the GFI-Classification network, resulting in a Glaucoma confidence level.

Finally, the decision is built, consisting of the Glaucoma confidence level and the calculated morphological features.

3.7. Mobile integration

To validate the performance of the developed pipeline in a restricted environment, it was integrated into a mobile application. This integration consisted on the creation of a background service that used TensorFlow Lite to run the developed models on a mobile environment. This service is run for a given fundus image and notifies the user when the pipeline is completed.

Afterward, the user can see the pipeline results, including the first disc segmentation, the cropped image and the joint optic disc/cup segmentation. Furthermore, the result of the morphological feature calculation, as well as the glaucoma confidence level, are also exposed in a dialog view. This visualization of the pipeline results contributes to the interpretability of the Glaucoma diagnosis, as can be seen in figure 5.

In table 4, several times and size metrics of the developed classification and segmentation models after being converted to the TensorFlow Lite format are exposed. The time metrics were obtained by performing 3 runs in isolation on the same mobile device and averaging times.

By analyzing the table, it is concluded that running the models on GPU is preferred than quantizing and running on the CPU since the times are better and no performance losses will occur. The selected models are highlighted in the table, and the pipeline can run in under two seconds per fundus image.

3.8. Interpretability

Interpretability is one of the main Achilles heels of deep learning models, since, despite having exhibited superior performances in plenty of tasks, it

Method	Acc	Sens	Spec	AUC
Transcribed from [32]				
ML 1 - Maheshwari et al. [33]	0.81-0.98	-	-	-
ML 2 - Acharya et al. [34]	0.95	-	-	-
DL 1 - Al-Bander et al. [35]	0.88	0.85	0.89	-
DL 2 - Fu et al. [36]	-	0.84	0.92	0.91
DL 3 - Li et al. [8]	-	0.95	0.92	0.98
DL 4 - Christopher et al. [37]	-	0.88	0.95	0.91
DL 5 - Mitsuhashi et al. [38]	-	-	-	0.96
DL 6 - Antón et al. [32]	0.88	0.87	0.89	0.94
Achieved results				
DL 7 - **GFI-Classification**	0.87	0.85	0.88	0.93
DL 8 - **GFI-Classification-VGG16**	0.86	0.81	0.88	0.95
DL 9 - **GFI-Classification-VGG19**	0.87	0.83	0.89	0.95
DL 10 - **GFI-Classification-InceptionV3**	0.90	0.80	0.94	0.95
DL 11 - **GFI-Classification-ResNet50**	0.88	0.83	0.90	0.93

Table 3: Classification performance comparison with state-of-the-art methods. In the methods column, *ML* and *DL*, indicate whether the method consists in a machine learning or deep learning approach, respectively. The metrics Acc, Sens and Spec stand for accuracy, sensitivity, and specificity, respectively. Results obtained on the test set.
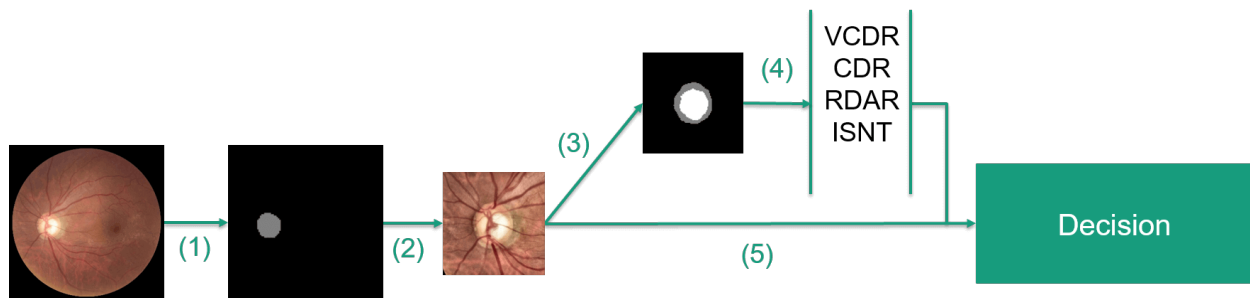

Figure 4: Proposed pipeline.

Model	CPU (ms)	GPU (ms)	Size (mb)
GFI-SPP-Depth-simple	≈ 250	≈ 215	1.1
GFI-SPP-Depth-simple_quantized	≈ 280	≈ 220	0.3
GFI-ASPP-Depth	≈ 580	≈ 450	4.6
GFI-ASPP-Depth_quantized	-	-	1.7
GFI-Classification	≈ 80	≈ 25	9.0
GFI-Classification_quantized	≈ 110	≈ 120	2.3

Table 4: Metrics of the models used in the Glaucoma assessment pipeline (Obtained on a Samsung Galaxy S8).

is difficult to make sense of these types of models when compared to machine learning approaches [39]. This mainly occurs because the feature representation is not engineered by the developer, but by the network itself, turning its translation into representative values challenging.

Also, this topic is not only important to further understand the inner workings of a model and ease the development process, but it is also essential, or even mandatory, for regulations and other bureaucracies, needed in case these methods are intended to be integrated into commercial solutions, especially in a medical context.

The proposed solution has three interpretability measures, some of them have already been partially addressed, but below a complete description is pro-
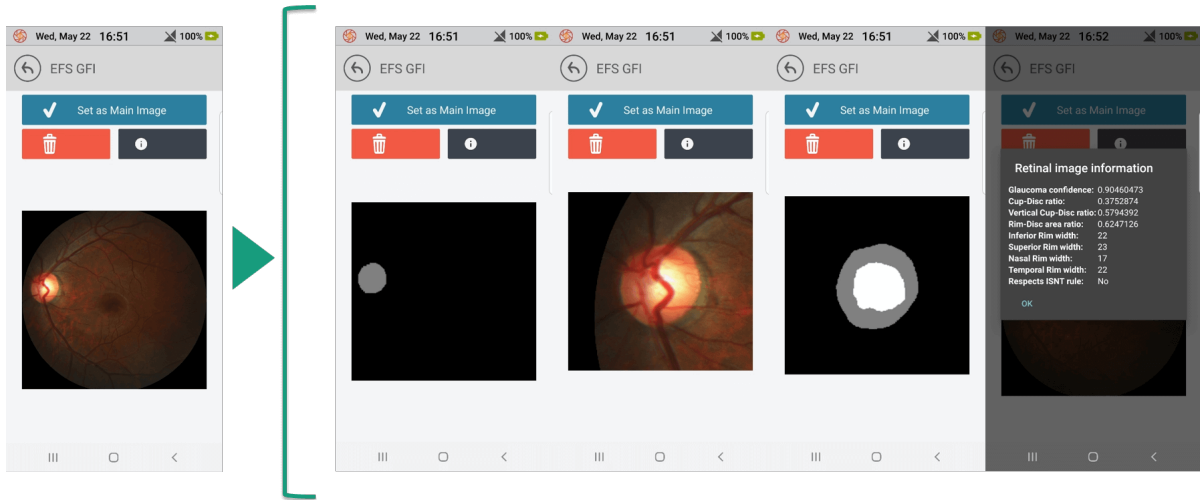
Figure 5: Pipeline integration in a mobile environment.

vided.

Pipeline intermediate results - These results are evidenced in figure 5 and were integrated into the system to allow quick confirmation of the pipeline outcome, since wrong segmentations may eventually happen. Having the possibility to visualize those results enables the user to discard them right away, if necessary.

Morphological features - These features make the model more interpretable since they have a defined and studied meaning among the ophthalmologist community. Despite not being directly used in the model that outputs the Glaucoma confidence level, the combination of the two output values turns the decision more trustworthy.

Activation Maps - The used activation maps are gradient class activation maps (Grad-CAM) [40]. These maps were calculated in the GFI-Classification network and an illustration of the results with images from both classes can be found in figure 6. In the figure, the images are overlayed with the calculated Grad-CAM maps and blue tones indicate the area was not important to the classification, while redder tones indicate a bigger influence of that region for the final decision. By analyzing the images, it is noticeable that the network is looking to the same structures that ophthalmologists inspect, such as the optic disc, cup, retinal vessels topology, and more interesting, on the left Glaucoma case, a focus on the Peripapillary Atrophy (PPA) region is visible, which is an identified

Glaucoma risk factor [41, 42, 2, 43]. All these insights contribute to increase the confidence in the network predictions. These maps are not available in the mobile pipeline integration due to Tensorflow Lite limitations, since the library is only targeted at inference. Nonetheless, they are available in a desktop environment and can be useful to make the pipeline more believable and help in regulatory processes.
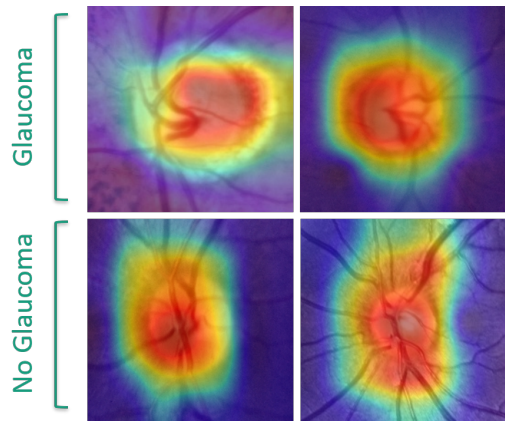

Figure 6: Grad-cam activation maps of the GFI-Classification network.

4. Conclusions

The main novelty of this work is the fact that the developed Glaucoma assessment pipeline runs

8

offline in mobile devices in seconds.

Moreover, this was achieved while having comparable, and sometimes even better, results in the segmentation and classification tasks of the developed system, contributing to the enhancement of the state-of-the-art in these topics as well.

Furthermore, the developed pipeline can be used to enable massive Glaucoma screenings in settings that were not possible before since it does not require an internet connection and runs on affordable mobile devices. This can increase the chance of detecting Glaucoma at an earlier stage and contribute to decrease both individual and economic burdens caused by the disease.

5. Acknowledgments

References

[1] H. Fu, J. Cheng, Y. Xu, D. W. K. Wong, J. Liu, X. Cao, Joint Optic Disc and Cup Segmentation Based on Multi-Label Deep Network and Polar Transformation, IEEE Transactions on Medical Imaging 37 (7) (2018) 1597–1605. arXiv:1801.00926, doi:10.1109/TMI.2018.2791488.

[2] Y. Hagiwara, J. E. W. Koh, J. H. Tan, S. V. Bhandary, A. Laude, E. J. Ciaccio, L. Tong, U. R. Acharya, Computer-aided diagnosis of glaucoma using fundus images: A review, Computer Methods and Programs in Biomedicine 165 (2018) 1–12. doi:10.1016/j.cmpb.2018.07.012.
URL https://doi.org/10.1016/j.cmpb.2018.07.012

[3] R. Thomas, K. Loibl, R. Parikh, Evaluation of a glaucoma patient, Indian Journal of Ophthalmology 59 (7) (2011) 43. doi:10.4103/0301-4738.73688.
URL http://www.ijo.in/text.asp?2011/59/7/43/73688

[4] T. Saba, S. T. F. Bokhari, M. Sharif, M. Yasmin, M. Raza, Fundus image classification methods for the detection of glaucoma: A review, Microscopy Research and Technique (2018) 1–17doi:10.1002/jemt.23094.

[5] J. Morales, Assessment of iris reflection artifacts and alignment in fundus images, Ph.D. thesis, Faculty of the Graduate School of St. Mary's University (07 2011). doi:10.13140/RG.2.2.26094.41280.

[6] B. Al-Bander, B. M. Williams, W. Al-Nuaimy, M. A. Al-Taee, H. Pratt, Y. Zheng, Dense fully convolutional segmentation of the optic disc and cup in colour fundus for glaucoma diagnosis, Symmetry 10 (4). doi:10.3390/sym10040087.

[7] A. Singh, M. K. Dutta, M. ParthaSarathi, V. Uher, R. Burget, Image processing based automatic diagnosis of glaucoma using wavelet features of segmented optic disc from fundus image, Computer Methods and Programs in Biomedicine 124 (2016) 108–120. doi:10.1016/j.cmpb.2015.10.010.
URL http://dx.doi.org/10.1016/j.cmpb.2015.10.010

[8] Z. Li, Y. He, S. Keel, W. Meng, R. T. Chang, M. He, Efficacy of a Deep Learning System for Detecting Glaucomatous Optic Neuropathy Based on Color Fundus Photographs, Ophthalmology 125 (8) (2018) 1199–1206. doi:10.1016/j.ophtha.2018.01.023.
URL https://doi.org/10.1016/j.ophtha.2018.01.023

[9] Z. Zhang, F. S. Yin, J. Liu, W. K. Wong, N. M. Tan, B. H. Lee, J. Cheng, T. Y. Wong, ORIGA-light: An online retinal fundus image database for glaucoma analysis and research, 2010 Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBC'10 (2010) 3065–3068doi:10.1109/IEMBS.2010.5626137.

[10] J. Sivaswamy, S. R. Krishnadas, G. Datt Joshi, M. Jain, A. U. Syed Tabish, Drishti-GS: Retinal image dataset for optic nerve head(ONH) segmentation, 2014 IEEE 11th International Symposium on Biomedical Imaging (ISBI) (2014) 53–56doi:10.1109/ISBI.2014.6867807.
URL http://ieeexplore.ieee.org/document/6867807/

[11] F. Fumero, S. Alayon, J. L. Sanchez, J. Sigut, M. Gonzalez-Hernandez, RIM-ONE: An open retinal image database for optic nerve evaluation, IEEE Symposium on Computer-Based Medical Systems (2011) 1–6doi:10.1109/CBMS.2011.5999143.

[12] F. Fumero, J. Sigut, S. Alayon, M. González-Hernández, M. G. de la Rosa, Interactive Tool and Database for Optic Disc and Cup Segmentation of Stereo and Monocular Retinal Fundus Images, Short Papers Proceedings - WSCG 2015, , Pilsen, Czech Republic, (2015) 91–97.
URL http://medimrg.webs.ull.es/

[13] ichallenge, http://ichallenge.baidu.com/, (Accessed on 07/12/2019).

[14] A. Almazroa, W. Sun, S. Alodhayb, K. Raahemifar, V. Lakshminarayanan, Optic disc segmentation for glaucoma screening system using fundus images, Clinical Ophthalmology 11 (2017) 2017–2029. doi:10.2147/OPTH.S140061.

[15] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, H. Adam, MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications, CoRRarXiv:1704.04861, doi:10.1016/S1507-1367(10)60022-3.
URL http://arxiv.org/abs/1704.04861

[16] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, L.-C. Chen, MobileNetV2: Inverted Residuals and Linear Bottlenecks, CoRRarXiv:1801.04381, doi:10.1134/S0001434607010294.
URL http://arxiv.org/abs/1801.04381

[17] H. Zhao, J. Shi, X. Qi, X. Wang, J. Jia, Pyramid scene parsing network, Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017 2017-Janua (2017) 6230–6239. arXiv:arXiv:1612.01105v2, doi:10.1109/CVPR.2017.660.

[18] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy,

A. L. Yuille, DeepLab: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully Connected CRFs., IEEE transactions on pattern analysis and machine intelligence 40 (4) (2016) 834–848. doi:10.1109/TPAMI.2017.2699184.
URL http://www.ncbi.nlm.nih.gov/pubmed/28463186

[19] L.-C. Chen, G. Papandreou, F. Schroff, H. Adam, Rethinking Atrous Convolution for Semantic Image Segmentation, CoRRarXiv:1706.05587.
URL http://arxiv.org/abs/1706.05587

[20] L. C. Chen, Y. Zhu, G. Papandreou, F. Schroff, H. Adam, Encoder-decoder with atrous separable convolution for semantic image segmentation, Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) 11211 LNCS (2018) 833–851. arXiv:arXiv:1802.02611v3, doi:10.1007/978-3-030-01234-2_49.

[21] K. Zuiderveld, Graphics gems iv, in: P. S. Heckbert (Ed.), Graphics Gems IV, Academic Press Professional, Inc., San Diego, CA, USA, 1994, Ch. Contrast Limited Adaptive Histogram Equalization, pp. 474–485.
URL http://dl.acm.org/citation.cfm?id=180895.180940

[22] G. D. Joshi, J. Sivaswamy, S. R. Krishnadas, Optic disk and cup segmentation from monocular color retinal images for glaucoma assessment, IEEE Transactions on Medical Imaging 30 (6) (2011) 1192–1205. doi:10.1109/TMI.2011.2106509.

[23] F. Yin, J. Liu, S. H. Ong, Y. Sun, D. W. Wong, N. M. Tan, C. Cheung, M. Baskaran, T. Aung, T. Y. Wong, Model-based optic nerve head segmentation on retinal fundus images, Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBS (2011) 2626–2629doi:10.1109/IEMBS.2011.6090724.

[24] J. Cheng, J. Liu, Y. Xu, F. Yin, D. W. K. Wong, N. M. Tan, D. Tao, C. Y. Cheng, T. Aung, T. Y. Wong, Superpixel classification based optic disc and optic cup segmentation for glaucoma screening, IEEE Transactions on Medical Imaging 32 (6) (2013) 1019–1032. doi:10.1109/TMI.2013.2247770.

[25] Y. Xu, L. Duan, S. Lin, X. Chen, D. Wing, K. Wong, Optic Cup Segmentation for Glaucoma Detection Using Low-Rank Superpixel Representation, Med Image Comput Comput Assist Interv 17 (Pt 1) (2014) 788–795.

[26] J. Cheng, D. Tao, D. W. K. Wong, J. Liu, Quadratic divergence regularized SVM for optic disc segmentation, Biomedical Optics Express 8 (5) (2017) 2687. doi:10.1364/boe.8.002687.

[27] O. Ronneberger, P. Fischer, T. Brox, U-net: Convolutional networks for biomedical image segmentation, Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) 9351 (2015) 234–241. arXiv:1505.04597, doi:10.1007/978-3-319-24574-4_28.

[28] C. Sciences, The ISNT Rule and Differentiation of Normal From Glaucomatous Eyes, JAMA Ophthalmology 124 (11) (2014) 1579–1583. doi:10.1001/archopht.124.11.1579.
URL https://doi.org/10.1001/archopht.124.11.1579

[29] K. Simonyan, A. Zisserman, VERY DEEP CONVOLUTIONAL NETWORKS FOR LARGE-SCALE IMAGE RECOGNITION, arXiv 1409.1556arXiv:1409.1556v6.
URL http://www.robots.ox.ac.uk/

[30] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, Z. Wojna, Rethinking the Inception Architecture for Computer Vision, CoRR abs/1512.00567. arXiv:1512.00567, doi:10.1109/CVPR.2016.308.
URL http://arxiv.org/abs/1512.00567

[31] K. He, X. Zhang, S. Ren, J. Sun, Deep Residual Learning for Image Recognition, CoRRarXiv:1512.03385, doi:10.1109/CVPR.2016.90.
URL http://arxiv.org/abs/1512.03385

[32] A. Antón, G. Fatti, M. J. Ledesma-Carbayo, A. Herranz, B. Liefers, A. Santos, J. J. Gómez-Valverde, C. I. Sánchez, Automatic glaucoma classification using color fundus images based on convolutional neural networks and transfer learning, Biomedical Optics Expressdoi:10.1364/boe.10.000892.

[33] S. Maheshwari, R. B. Pachori, U. R. Acharya, Automated Diagnosis of Glaucoma Using Empirical Wavelet Transform and Correntropy Features Extracted from Fundus Images, IEEE Journal of Biomedical and Health Informatics 21 (3) (2017) 803–813. doi:10.1109/JBHI.2016.2544961.

[34] U. R. Acharya, S. Bhat, J. E. Koh, S. V. Bhandary, H. Adeli, A novel algorithm to detect glaucoma risk using texton and local configuration pattern features extracted from fundus images, Computers in Biology and Medicine 88 (May) (2017) 72–83. doi:10.1016/j.compbiomed.2017.06.022.
URL http://dx.doi.org/10.1016/j.compbiomed.2017.06.022

[35] B. Al-Bander, W. Al-Nuaimy, M. A. Al-Taee, Y. Zheng, Automated glaucoma diagnosis using deep learning approach, 2017 14th International Multi-Conference on Systems, Signals and Devices, SSD 2017 2017-Janua (2017) 207–210. doi:10.1109/SSD.2017.8166974.

[36] H. Fu, J. Cheng, Y. Xu, C. Zhang, D. W. K. Wong, J. Liu, X. Cao, Disc-aware Ensemble Network for Glaucoma Screening from Fundus Image, CoRR (2018) 1–9arXiv:1805.07549, doi:10.1109/TMI.2018.2837012.
URL http://arxiv.org/abs/1805.07549{%}0Ahttp://dx.doi.org/10.1109/TMI.2018.2837012

[37] M. Christopher, A. Belghith, C. Bowd, J. A. Proudfoot, M. H. Goldbaum, R. N. Weinreb, C. A. Girkin, J. M. Liebmann, L. M. Zangwill, Performance of Deep Learning Architectures and Transfer Learning for Detecting Glaucomatous Optic Neuropathy in Fundus Photographs, Scientific Reports 8 (1) (2018) 16685. doi:10.1038/s41598-018-35044-9.
URL http://www.nature.com/articles/s41598-018-35044-9

[38] K. Mitsuhashi, Y. Fujino, M. Tanito, N. Shibata, H. Murata, R. Asaoka, M. Matsuura, Development of a deep residual learning algorithm to screen for glaucoma from fundus photography, Scientific Reports 8 (1) (2018) 1–9. doi:10.1038/s41598-018-33013-w.
URL http://dx.doi.org/10.1038/s41598-018-33013-w

[39] Q. Zhang, S.-C. Zhu, Visual Interpretability for Deep Learning: a Survey, CoRR 19 (1423305) (2018) 27–39. arXiv:1802.00614.
URL http://arxiv.org/abs/1802.00614

[40] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, D. Batra, Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization,

Proceedings of the IEEE International Conference on Computer Vision 2017-Octob (2017) 618–626. `arXiv: arXiv:1610.02391v3, doi:10.1109/ICCV.2017.74`.

[41] R. A. Shruti Gorasia, A Review Paper on Detection of Glaucoma using Retinal Fundus Images, International Journal for Research in Applied Science & Engineering Technology 4 (I) (2016) 166–170.

[42] C. Muramatsu, Y. Hatanaka, A. Sawada, T. Yamamoto, H. Fujita, Computerized detection of peripapillary chorioretinal atrophy by texture analysis, Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBS (2011) 5947–5950`doi:10.1109/IEMBS.2011. 6091470`.

[43] J. B. Jonas, Clinical implications of peripapillary atrophy in glaucoma, Current Opinion in Ophthalmology 16 (2) (2005) 84–88. `doi:10.1097/01.icu.0000156135. 20570.30`.

11