FACULDADE DE ENGENHARIA DA UNIVERSIDADE DO PORTO

Generative XAI in Computer-Aided Detection of Glaucoma Risk

Pedro António Ferreira Cardoso Videira Lopes

Mestrado Integrado em Engenharia Informática e Computação

Supervisor: Jaime dos Santos Cardoso

Second Supervisor: Filipe Cruz Gomes Soares

July 19, 2021

Generative XAI in Computer-Aided Detection of Glaucoma Risk

Pedro António Ferreira Cardoso Videira Lopes

Mestrado Integrado em Engenharia Informática e Computação

Approved in oral examination by the committee:

Chair: Prof. Jorge Alves da Silva
External Examiner: Prof. José Rouco Maseda
Supervisor: Dr. Filipe Cruz Gomes Soares

July 19, 2021

Abstract

Glaucoma is currently the leading cause of irreversible blindness across the globe. Moreover, since there is not much awareness about its risk factors, neither prevention and screening strategies until severe consequences are experienced, most individuals with Glaucoma remain undiagnosed across the entire world. For the last decade, experts developed a few approaches to tackle this problem, which utilize Machine Learning systems such as Deep Neural Networks. Some of the developed models show significant success in interpreting fundoscopic images and detecting the presence of Glaucoma. However, these models are usually models that do not provide a transparent overview of the reasoning behind their predictions, which is essential for a system to be implemented in a real-world scenario. Explainable AI (XAI) is a very recent field where researchers aim to create more interpretable and explainable Machine Learning models. The proposed solution will use state of the art techniques from the XAI field to extract the critical features on a model's prediction and better understand the current Computer Aided Detection (CADx) pipeline prediction model.

Deep Learning models applied to Glaucoma also suffer from the lack of available data, which is scarce and not very diverse. This is due to the absence of Glaucoma screening and privacy issues when aiming to make a dataset public. As a result, models must be prepared to handle all kinds of data, and it might be hard to cover most of the real case scenarios if the training data is incomplete. Our solution to this problem will be to apply Deep Learning Generative approaches to create new data while controlling their features. Besides improving the current CADx pipeline performance, this generated data will support the XAI techniques by providing data for case-based reasoning techniques, which allow clinicians to compare current cases without exposing older patients' clinical conditions.

Keywords: Glaucoma CADx, Deep Learning, Generative Modelling, Explainable AI

Resumo

Atualmente, o Glaucoma é a principal causa de cegueira irreversível a nível global. Devido à falta de sensibilidade sobre os fatores de risco desta doença, e à inexistência de proatividade relativa à saúde oftálmica, a maioria dos indivíduos com Glaucoma permanece por diagnosticar. Durante a última década, especialistas desenvolveram várias abordagens com o objetivo de solucionar este problema, que utilizam técnicas de Machine Learning como as redes neuronais profundas. Alguns destes modelos demonstraram bastante sucesso na interpretação de imagens fundoscópicas da retina e na deteção do Glaucoma. Estes modelos são muitas vezes modelos que não fornecem uma visão transparente sobre o seu "raciocínio" por detrás das previsões, um aspeto essencial na implementação de um sistema num cenário real. O campo de Explainable AI (XAI) é uma das mais recentes áreas com o objetivo de criar modelos mais interpretáveis e mais explicáveis. A solução proposta irá utilizar técnicas do estado da arte do campo de XAI para extrair of fatores mais relevantes das previsões de um certo modelo e para explicar o modelo de Diagnóstico Auxiliado por Computador (CADx) atual.

Um dos outros problemas surge devido aos dados disponíveis, que são escassos e pouco diversos devido à falta de rastreio de Glaucoma e às questões de privacidade relativas à publicação de bases de dados. Os modelos necessitam de conseguir lidar com todo o tipo de dados, e no caso dos dados de treino disponíves estarem incompletos, poderá fazer com que utilizar estes modelos em situações reais seja impossível. A solução proposta para ultrapassar este problema é utilizar abordagens de generação sintética de dados para criar novos dados, controlando as suas características. Estes dados serão usados para melhorar o sistema de CADx atual. Para além disso, também podem ser usados para suportar as técnicas de XAI adotadas, através de dados para técnicas de Case-based Reasoning, que permitem aos clinicos fazer comparações com casos atuais sem comprometer informações de casos anteriores.

Keywords:Glaucoma CADx, Deep Learning, Generative Modelling, Explainable AI

Acknowledgements

This dissertation would not be possible without the help of particular individuals and organizations, and I would like to start by giving my thanks to each and every one of them.

Firstly, I would like to acknowledge the two institutions that collaborated to create this opportunity. First, to Faculdade de Engenharia da Universidade do Porto (FEUP), I would like to thank for taking me in 5 years ago as a freshman student. It provided me until now with an academic environment that allows me to acquire a great deal of knowledge while sharing this same knowledge with other fellow students. Second, to Fraunhofer AICOS, I would like to thank for the opportunity to work with them in a welcoming environment, which granted me all the necessary tools and support to complete this project throughout the tough times everyone has had to face this past year.

To my university supervisor, Professor Jaime Cardoso, I would like to thank for all the help given when shaping work paths and discussing solutions and for being available when I reached out for help. To my supervisor from Fraunhofer AICOS, Senior Researcher Filipe Soares, I am very grateful for all the support that helped me get familiarized with this new area, for the discussions that pushed this work forward and for helping me focus on the end goal of this project.

I owe an enormous debt of gratitude to my closest friends and girlfriend, who had to listen to my complaints and frustrations throughout this work patiently. Because of their care and support, I was able to keep pushing forward towards the finish line, even when things were not going according to plan.

Last but not least, I would like to give a special thanks to all my family members, my parents, my brother, my grandparents and my uncles, who were always there from the beginning.

Pedro Lopes

"What you aim at determines what you see."

Jordan B. Peterson

Contents

List of Figures

List of Tables

Abbreviations

AE	Autoencoder
AEE	Adversarial Autoencoder
AI	Artificial Intelligence
AMD	Age-related Macular Degeneration
ASPP	Atrous Spatial Pyramid Pooling
AUC	Area Under the Curve
BM	Boltzmann Machine
CAD	Computer-aided Diagnosis
CAD(x)	Computer-aided Detection
CBR	Case-Based Reasoning
CDR	Cup to Disc Ration
CGAN	Conditional Generative Adversarial Network
CSLO	Confocal Scanning Laser Ophthalmoscopy
CNN	Convolutional Neural Network
CW	Concept Whitening
DBM	Deep Boltzmann Machine
DBN	Deep Belief Network
DCGAN	Deep Convolutional Generative Adversarial Network
DDLS	Disc Damage Likelihood Scale
DL	Deep Learning
DNN	Deep Neural Network
DR	Diabetic Retinopathy
GAN	Generative Adversarial Network
GMM	Gaussian Mixture Model
GradCAM	Gradient-weighted Class Activation Mapping
GRI	Glaucoma Risk Index
IOP	Intraocular Pressure
ISNT	Inferior, Superior, Nasal, Temporal
HMM	Hidden Markov Model
KL	Kullback–Leibler Divergence
KNN	K-Nearest Neighbors
KPI	Key Performance Indicator
OC	Optic Cup
OCT	Optical Coherence Tomography
OD	Optic Disc
ONH	Optic Nerve Head
MAE	Mean Absolute Error
ML	Machine Learning
MSE	Mean Squared Error
NB	Naïve Bayes
NF	Normalizing Flows
NN	Neural Network
NRR	Neuroretinal Rim

PACG	Primary Angle-Closure Glaucoma
POAG	Primary Open-Angle Glaucoma
PPA	Peripapillary Atrophy
RBM	Restricted Boltzmann Machine
RF	Random Forest
RNFL	Retinal Fiber Layer
ROC	Receiver Operating Characteristic curve
ROI	Region Of Interest
SLP	Scanning Laser Polarimetry
SmoothCAM	Smooth Class Activation Mapping
SVM	Support Vector Machine
TAMI	Transparent Artificial Medical Intelligence
VAE	Variational Autoencoder
XAI	Explainable Artificial Intelligence

Chapter 1

Introduction

1.1 Context

Glaucoma is a group of chronic eye diseases [56] that has become the leading cause of irreversible blindness across the globe [5]. Despite several Glaucoma variations, all of them can be characterised by loss of retinal ganglion cells, retinal nerve fibre layer (RNFL) thinning, and optic disc cupping. Moreover, intraocular pressure (IOP) is considered the major risk factor caused by the natural flow of aqueous humour inside the human eye. When this pressure increases to abnormal levels, it can damage the optic nerve head (ONH). Glaucoma is also known as the "silent thief of sight" due to mainly being asymptomatic until later stages. Several studies have tried to calculate and predict Glaucoma prevalence throughout the years. It is a common statement that the tendency is for the number of people affected by this disease to increase. Nevertheless, Glaucoma screening is not a common practice due to its low cost-effectiveness and the inexistence of a reliable and accessible strategy. For that reason, the majority of patients remain undiagnosed. This is a major concern in the healthcare community because Glaucoma can result in very severe consequences. Still, it is also possible to slow the disease's progression if treatment is applied at an early stage.

In a clinical environment, Glaucoma diagnosis is also a difficult task. As stated previously, Glaucoma progress remains hidden from both patients and clinicians for a long time. Most of the clinical procedures are focused on two eye structures, the optic nerve head and the retinal nerve fibre layer, such as tonometry (measure the eye's inner pressure), ophthalmoscopy (examination of the shape and colour of the optic nerve), etc. Retinal imaging technologies are another vital tool that allows clinicians to study the patient's retina. The most accessible and cost-effective technique created until today is fundus imaging, which essentially involves taking a 2D-photograph of the retina. From this image, clinicians can identify several morphological features, such as the optic disc and cup, and infer others relevant for the Glaucoma diagnosis, such as the Cup to Disc Ratio (CDR) or the ISNT rule.

Several researchers have presented CAD systems for Glaucoma detection based on machine learning techniques in the past decades. With the growth of the Deep Learning (DL) field, the state of the art approaches are mainly based on robust and resource-demanding algorithms that require a large amount of data to work correctly. Nevertheless, these approaches present satisfactory results and even discuss important aspects such as the computational costs of using such networks [79].

This dissertation work is also linked with a bigger project, TAMI (Transparent Artificial Medical Intelligence), focused on overcoming the lack of transparency and interpretability of AI models, not only for application in Glaucoma but also for other medical concerns and even other fields.

1.2 Motivation

Despite the enormous successes in the DL field, and more precisely in the CAD systems for Glaucoma detection, there are still huge challenges that need to be overcome in order to deploy these systems to a realistic scenario.

On the one hand, most of the concerns regarding DL models are related to their performance and their metric evaluation methods. Consequently, most approaches propose "black-box" type models that do not provide a transparent overview of the prediction's reasoning. As a result, researchers cannot understand and explain to others the reasons behind a models decision, making the task of correcting the model more challenging and less clear. On the other hand, it is also important to explain the machines' decision; otherwise, it would be very hard to regulate their usage. Moreover, end-users must trust the systems they use to make decisions. Namely, in a critical scenario like the medical field, where the clinicians need to reach a diagnosis, the system must provide clinically meaningful explanations that support its decision.

On the other hand, data obtained in the medical field can be scarce and not very diverse. Specifically for the Glaucoma case, since there is no screening strategy, this problem is even more prevalent. As stated previously, DL models require a considerable amount of data in order to generalize correctly. Moreover, it is essential to have a balanced dataset since imbalanced ones bring challenges to the learning process. A balanced dataset means having an equilibrium between the several possible scenarios and complete, which means providing enough cases to cover almost if not even all of the possibilities. Privacy issues are another data related drawback. Since models and researchers must deal with data from actual patients, there is also the risk of compromising the patients' privacy and exposing their clinical conditions.

1.3 Objectives

This dissertation aims to develop an XAI component that can be applied to CAD systems for Glaucoma detection to provide explainable model decisions to an expert from the healthcare domain. Besides, deep generative modelling will be used to obtain synthesized data that the XAI component can use to enhance the generated explanations.

1.4 Document Structure

This document is divided into the following Chapters: Chapter 1 describes the context of this works, as well as its motivations, objectives and document structure. Chapter 2 provides a background overview of Glaucoma and its characteristics, screening and clinical diagnosis, and also retinal imaging technologies relevant in the Glaucoma context. Chapter 3 describes the Glaucoma CAD systems structure, highlights several approaches, and goes over the available datasets, evaluation metrics, limitations and challenges of these systems, and systems applied to real-world scenarios. Chapter 4 is a literature review on Generative Modeling, with a focus on state-of-the-art Deep Generative Modelling approaches and on Semantic Image Editing. Chapter 5 is a literature review on Explainability and Interpretability in Machine Learning; it starts by giving an overview of important concepts of Explainable AI (XAI) and then gives a deeper notion of Interpretability techniques relevant for this dissertation's work, like Case-based Reasoning. Chapter 6 provides the problem definition and delineates the proposed solution, taking into account all the knowledge gathered in the previous chapters. Chapter 9 presents the conclusions for this monograph.

Chapter 2

Background: Glaucoma

This chapter focuses on the Glaucoma disease. Section 2.1 provides a description of the disease. Section 2.2 describes the screening strategies current status and Section 2.3 goes over the clinical diagnosis. At last, Section 2.4 describes the two main retinal imaging techniques, Fundus Imaging and Optical Coherence Tomography.

2.1 Disease Description

Glaucoma refers to a group of chronic eye diseases that can have different causes, risk factors, demographics, symptoms, duration, treatment, and prognosis [56]. Moreover, it has become the leading cause of irreversible blindness across the globe [5]. All types of Glaucoma can be characterised by loss of retinal ganglion cells, retinal nerve fibre layer (RNFL) thinning, and optic disc cupping. Intraocular pressure (IOP) is considered the primary modifiable risk factor since lowering its value usually slows Glaucoma progression or could even stop it. The natural flow of aqueous humour that occurs inside the human eye is the cause of such pressure. In abnormal cases, this substance's outflow facility is negatively affected, leading to an increase in IOP. There are still other risk factors that have shown to be relevant in the development and progression of Glaucoma: older age, ethnic background, positive family history for Glaucoma, stage of the disease and high myopia. Figure 2.1 shows the varying severity spectrum of Glaucoma. For most of that spectrum, and in most patients, no pain or relevant symptoms occur, which means the disease remains unnoticed most of the time. Only when patients start to lose their central vision ability do they seek medical assistance. However, when such symptoms manifest, Glaucoma is already at a late stage where irreversible damage has already occurred. Thus, Glaucoma is also known as the "silent thief of sight" [124].

The most common type of Glaucoma is primary open-angle Glaucoma (POAG), also existing others such as primary angle-closure Glaucoma (PACG). These two are usually the target of research in Glaucoma prevalence studies. An overview of all types of Glaucoma is presented below

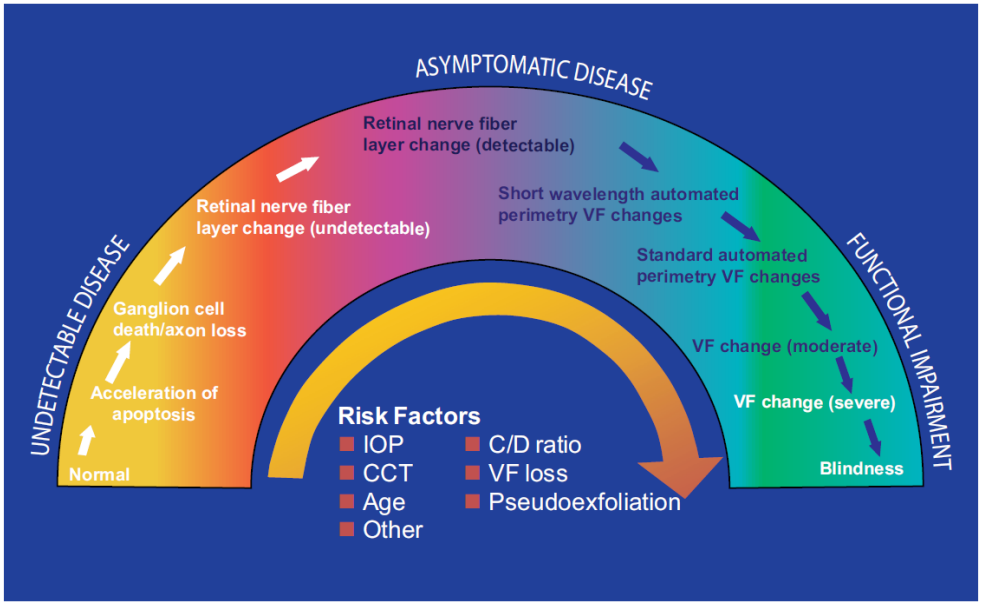


Figure 2.1: The Glaucoma severity spectrum [124].
CCT - central corneal thickness; **C/D** - cup-to-disc ratio; **IOP** - intraocular pressure; **VF** - visual field.

[108].

- **Primary open-angle Glaucoma (POAG)**: It is the most common type of Glaucoma. Its symptoms are only noticeable when optical nerve head (ONH) damage has become irreversible. This is the result of a rise in IOP due to the slow clogging of the drainage canal. With the disease's progression, blind spots start forming from the outer part of the vision field to its centre.

- **Primary angle-closure Glaucoma (PACG)**: Although being less common than the previous type, it is known for being very sudden. In this type, there is a sudden blockage of the drainage canals, leading to a rapid increase in IOP, which can cause irreversible blindness in just two days.

- **Normal tension Glaucoma**: Also known as low-tension Glaucoma, the leading cause of blindness in this type is not the increase in IOP. Although not yet proven, experts believe that in a normal range of pressure, these eyes are more susceptible to damage due to poor blood flow to the optic nerve. The IOP in these cases must be kept at even lower values.

- **Congenital Glaucoma**: This type is common amongst infants or babies, making it known as children Glaucoma. On the one hand, primary congenital Glaucoma results from incomplete or abnormal development of the eye's drainage canal. On the other hand, secondary congenital Glaucoma is caused by disorders in the eye or body.

- **Secondary Glaucoma**: This type describes Glaucoma conditions (two types below this one) that derive from other diseases.

- **Pigmentary Glaucoma**: In this type, pigment granules usually present in the back of the iris enter the aqueous humour that flows inside the eye. These flow towards the eye's drainage canal and slowly clog them, leading to an increase of IOP.
- **Neovascular Glaucoma**: the abnormal formation of blood vessels on the iris and over the drainage canals is the main cause of neovascular Glaucoma. It is usually associated with other diseases (e.g. diabetes), and the vessels block the fluid from draining correctly, causing an increase in IOP.

In 2010, Glaucoma was the cause of blindness in 2.1 million individuals and resulted in visual impairment in other 4.2 million. Glaucoma is more prevalent in high-income regions with a relatively old population than areas with a younger population. In 2013, the estimated prevalence of Glaucoma (POAG and PACG) in people aged 40-80 years old was 3.54%, and this value could increase by 74% to 111.8 million in 2040. From 1990 to 2010, estimates state that the number of individuals affected by Glaucoma increased by approximately 3.1 million people [13].

2.2 Screening

Across the entire globe, most patients (50-90%) with Glaucoma remain undiagnosed, due to the disease's characteristics and because no screening strategy has proven to be efficient enough until now. Suppose we tried to screen the entire population for Glaucoma. In that case, experts state that the number of false-positive diagnoses would be too high, due to the relatively low prevalence of Glaucoma (3.54% in individuals aged 40-80 years old as of 2013 [13]) and the insufficiently precise diagnostic methods. Nevertheless, there have been several attempts to identify a viable screening strategy for Glaucoma [56]. Burr et al. and colleagues [15] assessed the clinical screening for open-angle Glaucoma in the UK and its cost-effectiveness, concluding that general population screening at any age is not cost-effective. Furthermore, they also discovered that selective screening groups with higher prevalence (taking into account the risk factors) obtained better results and could be a more reliable approach. Another recent approach uses opportunistic case finding. In India [98], experts are attempting to integrate Glaucoma screening in an already existing cataract screening programme. With the results, they will calculate the costs of adding the new screening component to the current pipeline.

2.3 Clinical Diagnosis

Glaucoma diagnosis is a very challenging task. As stated before, for most of the severity spectrum, chronic forms of Glaucoma remain painless and measurable visual field defects do not develop at early stages. The patient is unaware of the disease's progress and only seek medical help when Glaucoma is already on a late stage.

Nevertheless, clinicians can use several procedures in a clinical environment to aid the Glaucoma diagnosis. Most of them are focused on two structures: the optic nerve head and the retinal

nerve fibre layer. Moreover, it might be necessary to examine the patient on several occasions to evaluate certain features, since healthy eye features can vary from patient to patient. Below there is a list of the several exams[41] used to help detect or diagnose Glaucoma on a patient:

- Tonometry: Measure the inner pressure of the eye;
- Ophthalmoscopy: Examine the shape and colour of the optic nerve;
- Perimetry: Examine the complete visual field of the patient;
- Gonioscopy: Classify the iridocorneal angle or the anatomical angle formed between the eye's cornea and iris;
- Pachymetry: Measure cornea thickness;

2.4 Retinal Imaging Technologies

Typically, the clinical examinations referred to previously are only used when there is already a suspected Glaucoma case. In addition to those, RNFL loss and OD changes can be detected using four modalities: confocal scanning laser ophthalmoscopy (CSLO), optical coherence tomography (OCT), scanning laser polarimetry (SLP) and fundus imaging. Besides carrying some disadvantages, the first three examinations are costly and depend on the subjective evaluation of qualified experts who manually inspect the individual retinal images. Fundus imaging is a technique that uses more economical and portable equipment, a fundus camera, resulting in a more sustainable method [50].

2.4.1 Fundus Imaging

Fundus imaging is one of the techniques used in retinal imaging, where the images are photographs of the eye's interior surface opposite to the lens. The first useful photographic images of the retina were obtained in 1891 by the German ophthalmologist Gerloff, and in 1910, Gullstrand developed the fundus camera. This idea maintains its popularity in retinal fundus imaging until today, not only for its safety (which was a very relevant feature at the time of this invention due to the prevalence of infectious diseases) but mainly for its cost-effectiveness at capturing retinal abnormalities. When used in the Glaucoma detection, it enables experts to make an earlier detection and settings where more expensive equipment is unusable. Furthermore, fundus images can also be used to identify other eye conditions such as age-related macular degeneration (AMD) or diabetic retinopathy (DR)[50].

Retinal fundus images have several features, that may vary from individual to individual, presented in Figure 2.2. Nevertheless, all individuals have the same structures, which can be identified in this type of imaging [120].

- **Optic Disc**: a central round-like yellowish part and the entry point for vessels. It is also known as the blind spot.

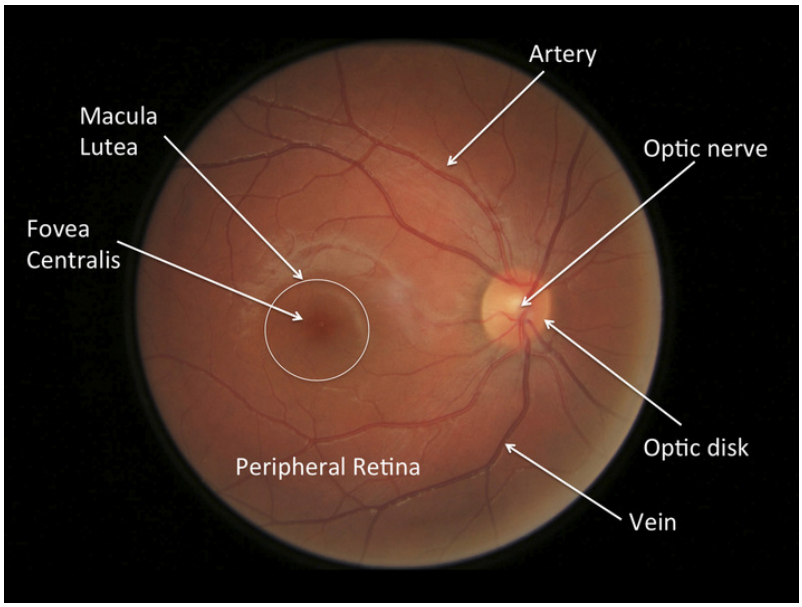


Figure 2.2: Eye anatomy with a few highlighted morphological feature.
Source[1]

- **Optic Cup**: located inside the optic disc, it is a bright central depression with variable size;
- **Macula**: Darkly pigmented area in the centre of the retina, which experts believe absorbs ultraviolet rays and excessive blue light;
- **Fovea**: Slightly concave and small area in the centre of the retina, where there are no vessels. The darkest area of the retina (dark-red or red-brown colour) and its cells provide the central vision for the human eye;
- **Retinal** Vessels: arteries and veins that carry blood throughout the eye;
- **Exudates**: Bright scattered patch like portions of the retina, formed after the leakage of vessels;

2.4.2 Morphological Features for Glaucoma CAD in Fundus Imaging

By using the previously referred features, it is possible to infer others, useful for Glaucoma detection. A description of the main ones can be found below.

Cup to Disc Ratio (CDR)

This is the most commonly used feature in Glaucoma detection across several pieces of research. CDR is the ratio between the optic cup and optic disc (illustrated in Figure 2.3 and can be calculated across the horizontal length, the vertical length or area. This metric allows to classify Glaucoma into mild (CDR up to 0.4), moderate (CDR between 0.5 and 0.7) and severe (CDR above 0.7)[120].

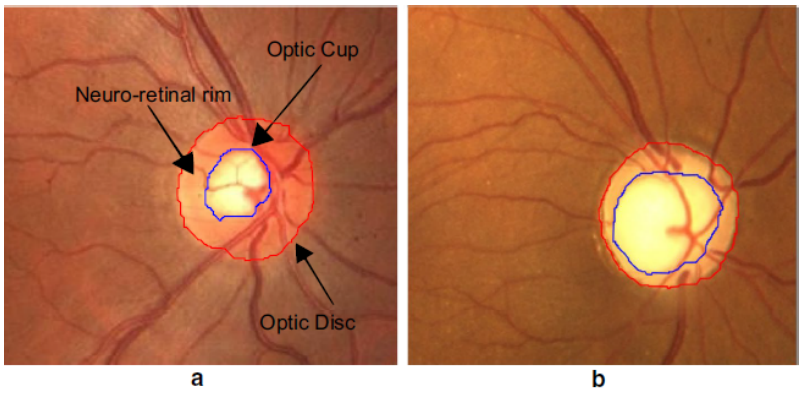


Figure 2.3: Digital fundus images cropped around optic disc. [34].
a Main structures of a healthy optic disc and **b** Glaucomatous optic disc.

ISNT rule

After identifying the optic disc, it is possible to measure the disc rim thickness in four directions, as presented in Figure 2.4: Inferior (**I**), Superior (**S**), Nasal (**N**) and Temporal (**T**). These measures should follow Formula 2.1. Although it cannot be used to diagnose Glaucoma immediately, it can be used to identify suspicious cases, since this rule is affected in most of the Glaucoma cases so far [101].

$$I > S > N > T \tag{2.1}$$

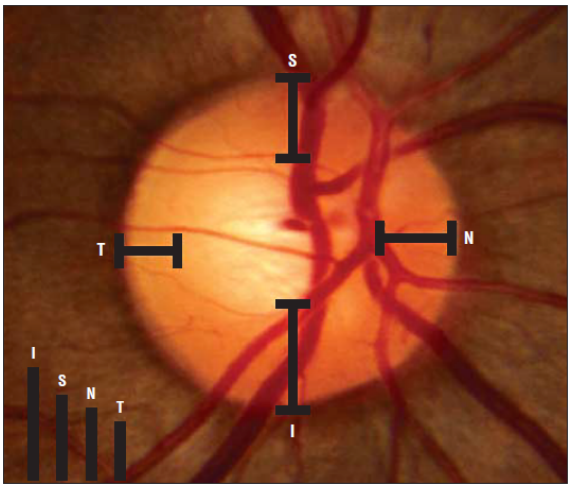


Figure 2.4: Clinical assessment of the ISNT rule for a normal optic nerve [101].

Neuroretinal Rim (NRR)

NRR (Figure 2.3) is the region between the edge of the optic disc and the edge of the optic cup. Like the CDR, the ratio between two pairs of the *ISNT* quadrants, the temporal and nasal quadrants and the superior and inferior quadrants, can be an indicator for Glaucoma.

[12].

Disk Damage Likelihood Scale (DDLS)

DDLS is the scale that calculates disc damage likelihood, giving the experts an idea of the severity of the disease. It is calculated using the formula below, where $MinRIM_{width}$ is the minimum width of the rim, and DD is the disc diameter [120]. Equation 2.2 shows the formula.

$$DLLS = \frac{MinRIM_width}{DD} \tag{2.2}$$

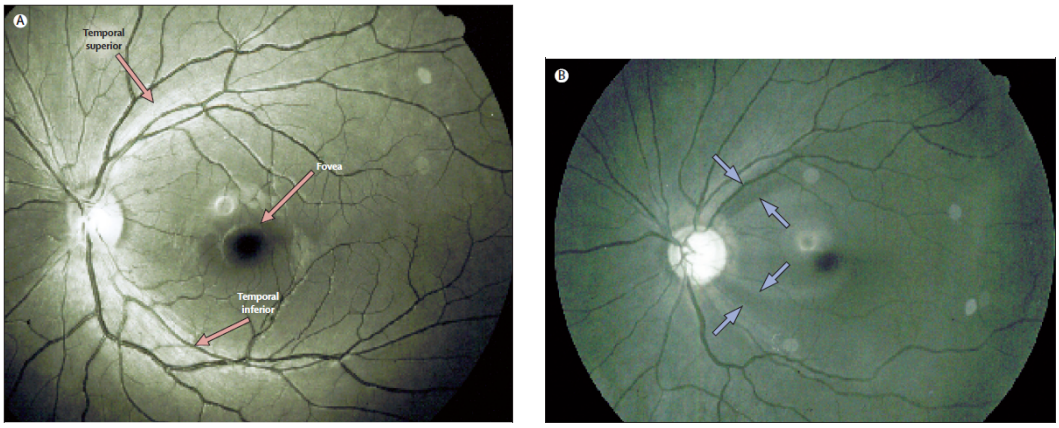
Glaucoma Risk Index (GRI)

Bock et al. [12] proposed this feature as a novel probabilistic index, that combines several components obtained from fundus images to get a single value. Experts can then use this number to distinguish a Glaucoma case from a healthy one: if the range of GRI is (8.68 ± 1.67) eye is considered normal and if the range is (4.84 ± 2.08), the eye is considered abnormal. Equation 2.3 is the original formula, but other works have modified it to fit other features. The variables PC1 to PC5 are the main components calculated using Principal Component Analysis (PCA) [120].

$$GRI = 6.8375 - 1.1325 \times (PC_1) + 1.65 \times (PC_2) + 2.7225 \times (PC_3) + 0.675 \times (PC_4) + 0.6650 \times (PC_5) \tag{2.3}$$

Retinal Nerve Fiber Layer (RNFL)

The RNFL is a part of the retina located outside the ONH, illustrated in Figure 2.5. It can be distinguished by an area with a particular texture, similar to a stripped whitish pattern. In normal cases, the RNFL is clearly visible and evenly distributed along the retina. Glaucoma reduces this layer's thickness, which leads to the loss of RNFL and consequent appearance of defects in the retinal fundus image [104].



(A) **(B)**

Figure 2.5: Retinal fundus images of two different eyes [56].
(A) The photograph shows a healthy retinal nerve fibre layer. **(B)** The photograph shows the retinal nerve fibre layer of an eye with Glaucomatous optic-nerve damage, with localised retinal nerve fibre layer defects (light blue arrows), in addition to a diffuse diminution of the retinal nerve fibre layer.

Peripapillary atrophy (PPA)

As it can be observed in Figure 2.6, PPA appears as a crescent-shaped part of the eye, composed of an alpha-zone and a beta-zone. These zones are outside the optic disc border, being the beta-zone closest to the disc. These zones tend to grow in size in abnormal cases[108], and large beta-zone can be considered as a clue of glaucoma [114].

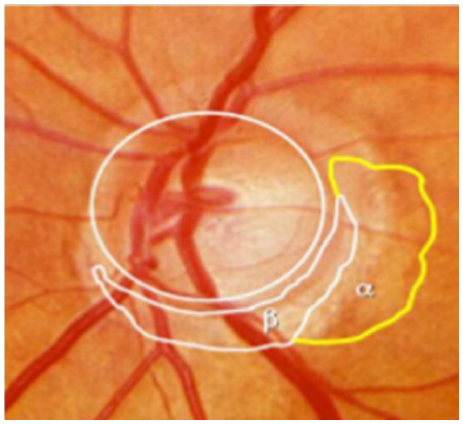


Figure 2.6: PPA With Alpha-Zone And Beta-Zone On The Right Eye [108].

Optic Nerve Notching

Optic Nerve Notching [127] is a focal loss of the neural rim width associated with a change in the rim curvature. Contrary to OD cupping, which is due to an overall OC enlargement, notching is the result of focal OC enlargement, and is mostly visible on the Inferior and Superior sections of the retina. In Figure 2.7 it is possible to observe a slight focal notching in the inferior area of the NRR.

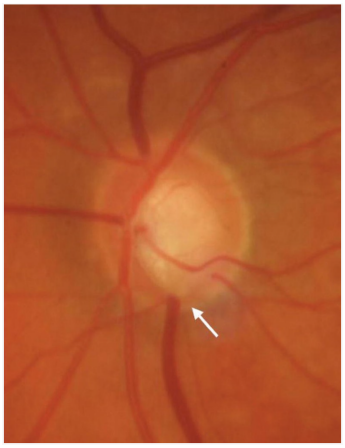


Figure 2.7: Fundus photograph demonstrating focal notching (white arrow) of the optic nerve at the inferior margin of the neuroretinal rim [127].

Optic Disc Hemorrhage

Optic Disc Hemorrhages [127] are flame-shaped or splinter-shaped hemorrhages in the RNFL at the NRR level, or close to the OD margin. Although not specific to Glaucoma, it sill is an indicator that show signs of lesion, and thus might have been caused by this disease. An example can be observed in Figure 2.8.

Figure 2.8: Fundus photograph demonstrating superior disc hemorrhage of the optic nerve (white arrow) [127].

2.4.3 Optical Coherence Tomography (OCT)

OCT is a technique which collects optical backscattering signal for cross-sectional and volumetric imaging of the biological tissues. At the cost of being more complex and expensive, it allows clinicians to assess with more detail Glaucoma-related anatomy (e.g. the anterior chamber angle closure) and structural damage (e.g. reduction of RNFL thickness).

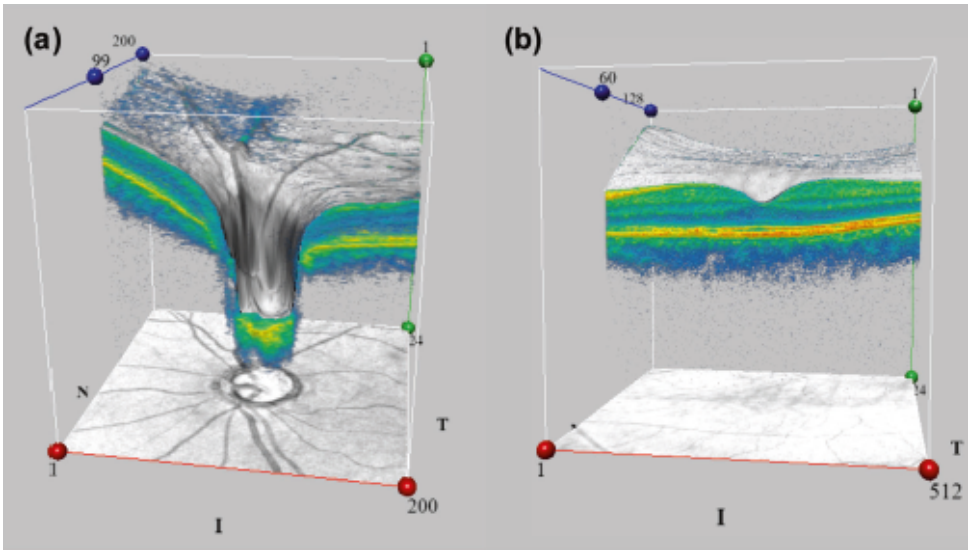


Figure 2.9: **(a)** an example of OCT volumetric optic disc scan as well as corresponding en face fundus image generated by linescanning ophthalmoscopy; **(b)** an example of OCT volumetric macula scan as well as corresponding en face fundus image [95].

There are mainly two types of OCT examinations useful for Glaucoma assessment. Posterior segment OCT is the most common modality for Glaucoma detection since it is the best-suited one to identify the most prevalent type of Glaucoma, POAG. Compared to fundus imaging, this technique enables a top view of the retina and the ONH, while capturing a more in-depth 3D view of the morphological features, and offering quantitative and topographical measurements. In this case, the traditional OCT report contains a key parameters table, a thickness and a deviation map of RNFL and its respective profiles, and specific quadrants and clock hours for Glaucoma detection. On the other hand, the anterior segment OCT is a less commonly used modality, more focused on detecting a less prevalent type of Glaucoma, PACG. Despite its lower prevalence, PACG still represents half of all Glaucoma blindness worldwide and is probably considered the most visually destructive form of Glaucoma. Moreover, this Glaucoma type is also preventable to some extent if diagnosed in the early stages. This technique allows clinicians to obtain cross-sectional images of the anterior segment of the eye and also a few measurements regarding certain biometric parameters: angle opening distance (AOD); anterior chamber area (ACA), depth (ACD) and width (ACW); scleral spur angle (SSA); rabecular iris space area (TISA); information about lens (lens thickness and lens vault), iris (iris area and pupillary diameter) and cornea (central corneal thickness and white-to-white). Due to its popularity, posterior segment OCT results are the most

widely used in DL models based on OCT imaging techniques. There are four categories of DL models with different input: Glaucoma classification based on traditionally measured thickness, thickness maps, deviation maps, and en face images; Glaucoma classification from segmentation-free OCT B-scans; Glaucoma classification from segmentation-free OCT volumetric scans; and "Machine-to-Machine" approach for OCT measurements (i.e., RNFL thickness) prediction from fundus photographs. An example of each of these approaches can be observed in Figure 2.10. In all categories, the existing DL models can use the OCT and its data as a tool to enhance Glaucoma assessment with efficiency and accuracy. The fourth category also shows fundus photographs potential since it is possible to calculate OCT associated measurements without conducting an OCT examination. Fundus imaging might substitute OCT in situations where the necessary equipment is not available or insufficient clinical expertise.

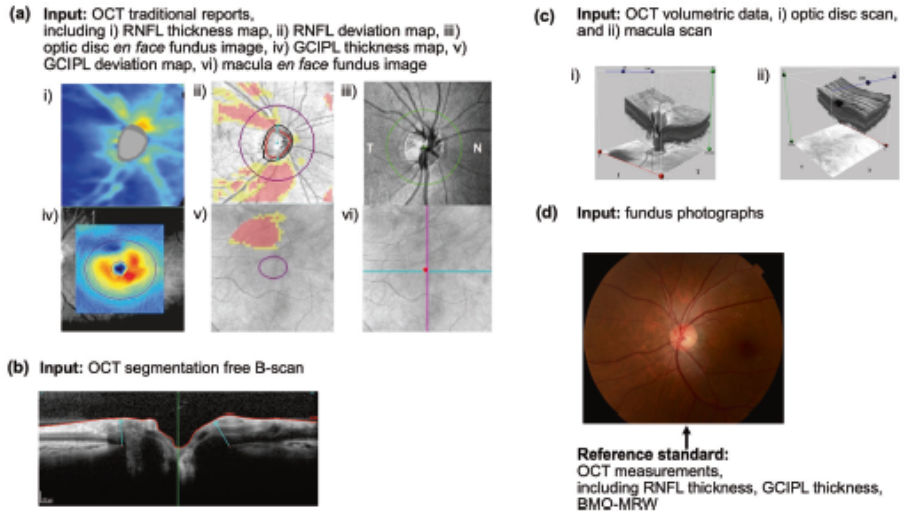


Figure 2.10: The four categories of DL models with different input [95].

2.5 Summary

This chapter gave an overview of Glaucoma and its characteristics, risk factors, prevalence and current screening and diagnosis workflows. Even though the disease is well-known amongst the medical community, it is still a major cause of blindness worldwide.

Furthermore, retinal imaging was also discussed, and there are at least two techniques which can capture features that are relevant in clinical Glaucoma diagnosis. One the one hand, there is the more economical and accessible fundus imaging, that only requires a fundus camera and a lower amount of expertise to obtain a retinal fundus imaging. Although some morphological features such as the CDR and the ISNT rule can be assessed in these photographs, clinicians do not solely rely on this technique to create the final diagnosis. As for the OCT, it is a more extensive

technique that collects much more data about the patient's eye by default. Nevertheless, it requires more expensive equipment and also more expertise.

Chapter 3

Background: Glaucoma CAD Systems

This chapter contains the background research regarding Glaucoma CAD systems and highlights state of the art techniques. After a small overview on Section 3.1, Sections 3.2,3.3 and 3.4 go over the different CAD systems' tasks, which are respectively Pre-Processing, Segmentation and Classification. Section 3.5 lists the most widely known retinal imaging databases. Section 3.6 lists the evaluation metrics proposed in several methods. Section 3.7 goes over the limitations and challenges identified by the previous literature works and Section 3.8 describes Glaucoma CAD systems that were launched on real-world scenarios or developed with that intent.

3.1 Overview

Glaucoma screening strategies are almost nonexistent, and clinical diagnosis is an expensive and complicated task. In recent years, there have been several attempts to create automated tools that make both these practices more accessible, more efficient, and more cost-effective. From the previous chapter, we know that fundus photography has shown to be very efficient at capturing retinal features. In the current clinical practice, this technique is complementary to others referred previously (2.3), since together they give clinicians an overall view of the patient's eye condition [87]. However, it is believed that the information acquired by fundus imaging still has the potential to be exploited and used to relieve the burden from clinicians and make Glaucoma detection more effective.

Earlier approaches for CAD systems were based mostly on traditional techniques, that followed a specific workflow. This workflow is represented in Figure 3.1 and consists of the following steps: input data, pre-processing, segmentation, feature extraction, feature selection and classification [50].

Most of these techniques are surveyed in [51], [9] and [120]. These methods have the major drawback of dealing with hand-crafted features, which likely do not capture the variability

Figure 3.1: CAD system workflow.

of the disease's characteristics, even in relatively small datasets. On the other hand, deep learning techniques have received a lot of attention from researchers in many fields, including retinal imaging and eye disease's detection. These can automatically find patterns within the data, obtaining relevant data representations without the need for applying any manual feature extraction techniques[87].

Convolutional Neural Networks (CNN) is the most widely implemented form of deep learning across most fields and has proven to be very useful in retinal images. CNN learn to minimise a loss function, an objective that scores the quality of results. Although the learning process is automatic, losses must be effectively and carefully designed to have an efficient model.

3.2 Pre-Processing

Independently of the type of workflow used to detect Glaucoma, it is well-known that medical images, specifically fundus images, contain noise and artefacts that harm the model's performance. For both normal and abnormal cases, minor details in relevant parts of the image can significantly impact the final prediction. For that reason, the pre-processing step is essential to remove or attenuate noise and artefacts of a single image. Moreover, it is also vital to consider inter-image variability, since images can sometimes be obtained under different conditions (different fundus cameras, for example). Each case is different from the other. Some techniques have been used to make the input data more homogeneous, giving the model a more precise input data, where essential features are enhanced [125].

Non-uniform illumination is a recurrent problem in fundus images. Normalization and standardization of RGB values, conversion from RGB to HSV values [106], illumination correction algorithms [125], image contrast enhancement techniques such as CLAHE [116] are some of the technique used to address this issue. Some approaches ([75], [85],[110]) also remove blood vessels since they represent noisy pixels for segmentation tasks. More refined techniques have been used in more recent deep learning approaches to ease the optic disc and cup segmentation. Fu et al. [42] applied a polar transformation to obtain a pixel-wise representation of the fundus images, which keeps flexibility in terms of data augmentation while adding spatial constraint for layer-based segmentation and balancing the cup proportion. Yin et al. [131] enhanced important features by applying Multiscale Detail Manipulation to change certain light values and applied dehazing to the images, which revealed certain hidden features cause by a cloudy camera or cataracts. Kang et al. [58] used pixel quantification to reduce the model's sensitivity to colour. Images obtained from different cameras usually come with a different colour scheme due to camera properties.

Images are sometimes resized, because image size has an enormous impact on the computational time used to process it or because images used as input come from different datasets. If the image is smaller, the model can more easily process it. However, if the image is too small, there might be too much loss of details necessary to detect Glaucoma. In almost every approach [87], images are cropped to the Region Of Interest (ROI), the region of the fundus image that clinicians consider to contain the most relevant features for Glaucoma detection. It is shown that this dramatically improves the model's performance in almost all cases. One of the drawbacks of this pre-processing technique is removing information from the input data, restricting the model from learning alternative features [128]. Due to the low amount of data publicly available and to reduce the probability of overfitting significantly, most approaches use data augmentation techniques, such as rotations and reflections, removing the model's sensitivity to slight changes in the position of retinal features.

3.3 Segmentation

Segmentation is a crucial step in the CAD system workflow since it allows researchers to represent certain features efficiently. The majority of approaches chooses to segment the optic disc (OD) and the optic cup (OC) due to their relevance in detecting suspicious Glaucoma cases. Several methods have been developed and approach this problem in different ways ([9], [120]). OD segmentation is based on the "ground truth" obtained from ophthalmologists and usually consists of two different steps: localisation and segmentation. Mitra et al. [83] proposed a methodology to localise the OD that uses a CNN to create a bounding box that encloses the OD. Other approaches utilise intensity values to identify the ROI since the OD represents a retina region with intense brightness.

Shantayia et al. [106] proposed two different approaches. For the OC, the green plane is extracted and converted to a grayscale image, where the contrast between the OC and other regions of the image is better. From that new image, a brightness threshold is set to obtain a binary image of the OC. For the OD, both the green plane and the V-plane are used. This combination enables a more accurate distinction of the OD from the rest of the image. Finally, the empty spaces that cross the OC and OD areas are filled since they are blood vessels' location. Singh et al. [110] approach only segments the OD, since the ROI and relevant features can be inferred from it. After detecting its location, the OD is segmented from the image and the blood vessels removed. Wavelet feature extraction is applied to the segmented optic disc image, capturing features later used by the classifier. Both evolutionary and discriminatory feature selection are evaluated to understand which method would improve the classifier's performance and accuracy.

On more recent methods, the joint segmentation of the OD and OC has shown to be very useful, improving the segmentation component's performance without harming the segmentation result. Zhao et al. [139] start by using both intensity information and blood vessels to localise the OD centre, cropping the ROI based on it. After a few pre-processing steps to improve image quality (image enhancement, blood vessel extraction and confidence calculation of the sliding window), both OD and OC are segmented using a U-shape convolutional architecture (U-Net). Chakravarty

et al. [18] also proposed an approach where a U-net is used to obtain the OD and OC segmentation, achieving a dice coefficient of 0.92 for OD segmentation and 0.84 for OC segmentation. Similarly, Martins et al. [79] utilise a U-shaped network to build two different networks: one for the joint OD/OC segmentation and one that only executes the OD segmentation. Both approaches were compared to both performance and model complexity penalty when segmenting the OC. Although the network performing the joint segmentation obtained a better IoU value of 0.91, the other network achieved a comparable value, 0.89, with less than one-fourth of the parameters. Fu et al. [42] proposed an M-Net Architecture constituted by four main parts: multiscale layer, U-shape CNN (U-net like architecture), side-output layer and a multi-label loss function. Firstly, the OD is localised, and polar transformation is used to obtain a new representation based on the previously detected disc centre. The image is then processed by the M-Net, producing a multi-label prediction map for the disc and cup regions. Finally, an inverse polar transformation operation is applied to reconstruct the segmentation result into the Cartesian coordinate.

In 2019, the first edition of the "REFUGE Challenge" competition was held to develop an evaluation framework that would ease comparison between different models and encourage innovation. Teams were given two tasks: OD/OC segmentation and Glaucoma classification. In the end, they presented several new approaches, some of them with state of the art performance [87]. Kang et al. [58] made use of an existing deep learning model for image segmentation, DeepLab v3+, which takes advantage of atrous spatial pyramid pooling (ASPP) to segment objects at multiple scales, with filters at multiple sampling rates and effective fields-of-views. This model's key feature is a simple yet effective decoder module that can refine the segmentation results, mainly along object boundaries. After obtaining a segmentation probability map from the model, it is converted to a binary image using a threshold method, where the component with the largest area is the optic disc. Liu and Fang et al. [73] presented an approach based on the already referenced U-Net like architecture, adding squeeze-and-excitation blocks that recalibrate channel-wise features responses to improve the model's performance at a low computational cost [53]. Yin et al. [131] used a framework that localises and segments the ROI simultaneously. Wang et al. [126] work, which corresponds to team CUHKMED, proposes a segmentation method that minimises the performance loss when the models need to deal with inconsistent input data, namely images from different datasets. This is achieved by applying an Output Space Domain Adaptation, which forces the network to learn the target image feature while knowing the current domain's segmentation mask. For both the current and target domains, the mask structure must be equivalent.

Table 3.1 shows the performance state-of-the-art OD/OC segmentation models.

Method	Iou Disc	IoU Cup
R-Bend[57]	0.8710	0.6050
ASM[130]	0.8520	0.6870
Superpixel[26]	0.8980	0.7360
LRR[129]	-	0.7560
QDSVM[27]	0.8900	-
U-net[97]	0.8850	0.7130
M-net[42]	0.9170	0.7440

Table 3.1: Performance comparison of state-of-the-art methods trained with the ORIGA dataset (Adapted from [42]).

3.3.1 Beyond Optic Disc and Optic Cup segmentation

One of the difficulties in OD segmentation is the presence of peripapillary atrophy (PPA). Due to its similar brightness and colour to the OD and also being located right outside the OD boundary, it is not unusual for some segmentation models to incorrectly consider the PPA as part of the OD region. Besides, PPA a risk indicator of Glaucoma, and manual annotation is a tedious, time-consuming and subjective task. For that reason, it is vital to develop a method to identify this feature in fundus images. Muramatsu et al. [84] presented a work that explored the detection of moderate to severe PPA β-type, which is the most relevant type for the already stated reason. It was possible to identify at least part of the PPA in some cases using texture analysis. However, more investigation would be needed to improve the model's sensitivity to mild and severe PPA and detect its boundaries more precisely. Cheng et al. [24] explored the PPA problem a bit deeper, presenting three different postprocessing PPA filters, each one with a specific function. Lu et al. [75] also proposed a method for removing the PPA from the OD segmentation, by subtracting an OD segmentation from an OD-plus-PPA segmentation and applying a multiseed region growing method to fix any incorrect segmentation in the boundary of both regions. Cheng et al. [25] presented a biologically inspired feature (BIF), which mimics the cortex's visual perception process to identify the PPA automatically. A threshold-based segmentation localises the focal region from where this feature will be extracted. Then the problem becomes a classification problem to determine the presence of PPA or not in that same region. The proposed approach achieved over 90% accuracy on PPA detection. More recently, Chai et al. [17] divide the PPA segmentation task into a two-part segmentation, the PPA-disc area and the Disc area. Since PPA can have irregular and non-uniform shapes, as we can see from Figure 3.2, it is more efficient to segment the PPA and Disc jointly, and then subtract from it the disc area to get only the PPA. A multi-task fully convolutional network is used for the segmentation task, achieving an average precision of 0.8929, above other state-of-the-art approaches.

Although less present in literature, the retinal nerve fibre layer (RNFL) is another risk factor of Glaucoma, and some approaches have tried to predict Glaucoma using this feature. Septiarini et al. [104] proposed an automated detection of RNFL based on the texture feature of this region. This proposal's pillar uses a co-occurrence matrix derived from small areas (patches) outside the ONH,

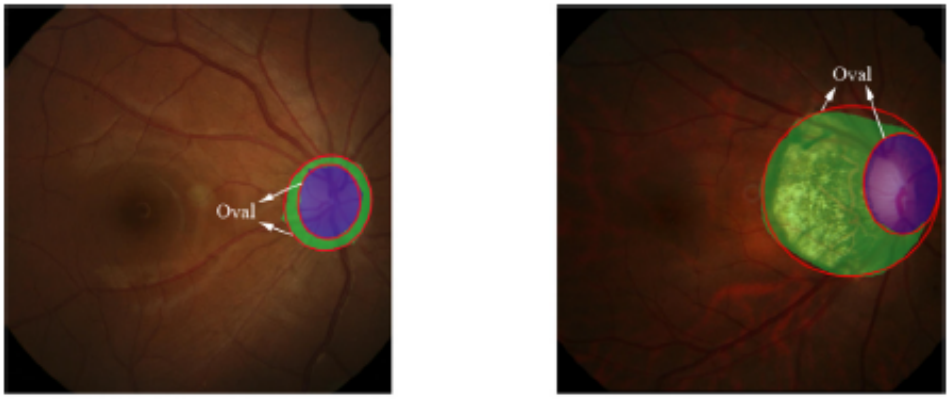


Figure 3.2: Retinal images and their PPA and Disc areas [17].

which shows RNFL loss. In the first stage, feature values are obtained from several images. In the second stage, with the images divided by sectors, these features are tested to detect the presence of RNFL. In [85], the proposed CAD system uses a polar representation of fundus images to identify RNFL defects resulting from RNFL loss. The first stages consist of pre-processing the image by correcting illumination and removing blood vessels before converting it to a polar representation. Then, RNFL candidate defects are detected by Hough transformation as dark straight vertical lines. False Positives are eliminated from these candidates by using knowledge-based rules.

Notching is another not studied morphological feature that can also be a Glaucoma indicator. Sivaswamy et al. [112] proposed a method for automatically detecting notching from the OD and cup segmentation, based on evaluating the rim thickness on the inferior and superior sections of the ONH.

3.4 Classification

Direct analysis of morphological features (e.g. CDR) is the simplest form of classification present in literature. In [106] and [58], a threshold value for the vertical CDR is defined, allowing the model to classify an image as a normal or abnormal case.

Other approaches use the most traditional machine learning classifiers, giving them the features extracted and selected in the previous stages. Singh et al. [110] tested five different classifiers: Random Forest (RF), Naïve Bayes (NB), k-nearest neighbours (k-NN), Artificial Neural Network (ANN) and Support Vector Machine (SVM). The experiments conducted consisted of a combination of these classifiers with two feature selection methods. Every experiment obtained an accuracy of over 85 %. RF and ANN showed better accuracy for the evolutionary feature selection method (94.7%), while SVM and k-NN showed better results for principal component analysis (PCA) selection method (94.7%). Maheshwari et al. [78] used a variant of a traditional classifier, Least Squares Support Vector Machine (LS-SVM), a method already applied in previous works in Glaucoma detection in fundus images. They obtained high accuracy values for the private dataset (98.33% and 96.67% using three-fold and ten-fold cross-validation). They were also able to get a sensitivity of 100%, which means the model did not predict any false negatives. Zhao et al. [139] extracted 25 features related to the OD, OC and NRR, and after removing redundant features using correlation analysis, used them as input to RF and SVM classifiers. Only SVM had relevant results in the context of Glaucoma detection, obtaining 95.5% specificity and an AUC of 83.4%.

In more recent years, deep learning methods have seen a tremendous increase in popularity and research. There are already some works that obtain state of the art or even better results than the more classical approaches. One of these methods' advantages is that they remove the need for hand-crafted features and can more easily capture all features present in the dataset. Martins et al. [79] created a classification network with MobileNetV2 as a feature extractor backbone, followed by a global average pooling layer, and two fully connected layers, interleaved by heavy dropouts. This architecture was able to obtain results similar to other state-of-the-art approaches, but with a lower amount of parameters when compared to the most recent one. Xiangyu Chen et al. [128] propose a CNN for Glaucoma detection, with a simple workflow: ROI extraction, dropout and data augmentation, CNN with a soft-max classifier for Glaucoma prediction. Raghavendra et al. [94] claim to have developed the first automated CNN architecture for Glaucoma CAD in digital fundus images, presenting a robust model with state of the art performance. The model was able to obtain 98.13% accuracy and could efficiently detect the class (normal or Glaucoma) of an unknown image. Abbas et al. [2] also implemented a CNN model to extract features from fundus images and classify them. Moreover, the workflow also had an extra component, responsible for optimising deep features through a supervised deep-belief network (DBN) deep-learning algorithm.

In [30], [47] and [34], several pre-trained CNN architectures were fine-tuned to the Glaucoma classification problem. The work shows results for two versions of each model: the native version and another version based on transfer learning. Although each work presented slightly

Method	**Datasets**	**Accuracy**	**Sensitivity**	**Specificity**	**AUC**
ML-1[12]	Private (336-/239+)	0.8800	-	-	0.8700
ML-2[65]	Private(30-/30+)	0.9167	-	-	-
ML-3[78]	Private(30-/30+) RIM-ONE(255-/250+)	Private: 0.9833 RIM-ONE: 0.8132	-	-	-
ML-4[4]	Private(132-/559+)	0.9570	-	-	-
DL-1[128]	ORIGA(482-/168+) SCES(1676-/46+)	-	-	-	0.8310-0.8700
DL-2[7]	RIM-ONE(255-/250+)	0.8820	0.8500	0.8980	-
DL-3[43]	ORIGA(482-/168+) SCES(1676-/46+)	-	0.8478	0.8380	0.9860
DL-4[71]	Private (48116)	-	0.9560	0.9200	0.9860
DL-5[30]	Private (9189-/5633+)	-	0.8800	0.9500	0.9100
DL-6[107]	Private (1768-/1364+)	-	-	-	0.9650

Table 3.2: Performance comparison of state-of-the-art Glaucoma classification methods (Adapted from [47].

different performance results, all concluded that models from other problems show competitive performance when fine-tuned, even if the training data domain is different from the original one.

Although metrics are essential in assessing the effectiveness and usefulness of a method, it is also crucial to consider the implementation environment and the end-user. The model's complexity and interpretability are two critical aspects that need to be taken into account. In the context of Glaucoma assessment, when implementing a model in a realistic environment, such as a health institution, the number of computational resources available can limit the model's performance and its usefulness. Moreover, for clinicians to make use of that model in a practical context, a user-friendly and straightforward interface must be provided, as well as explanations that support the model's predictions.

Table 3.2 shows the methods considered relevant for the context of this work and their performance metrics in classification.

3.5 Datasets

Table 3.3 gives an overview of the fundus images datasets referenced in several proposed deep learning models. Overall, there is a low amount of publicly available data compared to the amount necessary to train a deep learning model for a realistic situation.

Table 3.3: Fundus Image Dataset Information.

Dataset Name	Images	Usage	Availability
ACHIKO-K[141]	258 manually annotated images, 114 Glaucoma, 144 Normal	Glaucoma detection	Unavailable
ACRIMA[34]	705 fundus images (396 Glaucomatous and 309 normal images)	Glaucoma Detection	Available Online
CHASE[1]	28 images	Blood vessel segmentation	Available Online
DRIONS-DB	110 images, 23.1% Chronic Glaucoma and 76.9% Eye Hypertension	Glaucoma Detection	Available Online
DRISHTI-GS[112]	101 images	Glaucoma Detection	Available Online
DRIVE[2]	40 images, 33 normal and 7 mild DR	Vessel Segmentation	Unavailable
Esperanza	1446 color fundus images	Glaucoma Detection	Unavailable
HRF[3]	45 images,15 images each of healthy, DR, Glaucomatous patients	Glaucoma Detection	Available Online
ORIGA-light[136]	650 retinal images	Glaucoma Detection	Available Online
iChallenge-GON[4]	1200 annotated images	Glaucoma Detection	Available Online
iChallenge-PM[5]	800 annotated images	PPA and Myopia Labels	Available Online
RIGA[8]	760 retinal fundus images	Glaucoma Detection	Available Online
RIM-ONE[44]	783 images	OD segmentation	Unavailable
SCORM	1584 retina images	PPA and Myopia Detection	Unavailable
SEED	235 images, 43 Glaucoma and 192 normal	Glaucoma	Unavailable
STARE[6]	400 images, blood vessel annotation on 40 images	Blood vessel segmentation	Available Online

Source: [103], [47] and [87]

Below, we present a more detailed description of the datasets used throughout this work:

iChallenge-GON[7] This dataset was made available through the REFUGE challenge, an online competition organized for the MICCAI 2018 conference. The dataset contains 1200 colour fundus photographs, split into three equally sized subsets for training, validation and testing. Each of these subsets has the same Glaucoma presence percentage. Annotations for disc,

cup and fovea were provided, as well as a Glaucoma label. Since this is a competition, only the train and validation subsets were made publicly available, resulting in 800 images.

ORIGA[136] This dataset was obtained during a population-based study in Singapore (Singapore Malay Eye Study - SiMES). It consists of 650 retinal fundus images, each one with several annotations: eye side, CDR, ISNT rule, RNFL, Notch, Disc Haemorrhage, PPA, Glaucoma label and others. Despite being a public dataset, the dataset is supposed to be obtained through a request to the authors. Nevertheless, the dataset was obtained through a previous work [79].

RIM-ONE[44] This is an open retinal fundus image dataset consisting of three different releases:

RIM-ONE r1 Published in 2011, this release is composed of 169 ROI cropped fundus images, each one with the respective optic disc boundary annotation. These are classified into four Glaucoma labels (none, early, moderate and deep).

RIM-ONE r2 The second release is composed of 455 ROI cropped fundus images, and also their respective optic disc boundary and a binary Glaucoma label.

RIM-ONE r3 The third version of the dataset consists of 159 stereo retinal fundus images, with optic disc and optic cup annotations and a binary Glaucoma label. These stereo images contain two different photographs of the same eye, taken from slightly different angles, which allow the experts to create more accurate annotations.

ACRIMA[34] This dataset is composed of 705 ROI cropped fundus images, 396 Glaucomatous and 309 healthy ones. Most of the images are centred in the optic disc and were annotated with binary Glaucoma labels.

RIGA[8] This dataset contains 750 retinal fundus images, obtained from three different sources: Messidor dataset (460 images), Bin Rushed Ophthalmic centre (195 images) and Magrabi Eye centre (95 images). Six ophthalmologists manually annotated the dataset with the optic disc and cup boundaries. Unlike most of the publicly available datasets, it does not contain any Glaucoma labels.

iChallenge-PM[8] Similarly to the **iChallenge-GON** dataset, this one was also made available through an online competition called PALM-iChallenge. The dataset comprises 1200 retinal fundus images from pathological and non-pathological myopia subjects, annotated with optic disc boundary, fovea location and lesions boundaries. Each image is labelled with the degree of myopia: normal image, high myopia or pathological myopia. Although the subject of the competition is myopia, one of the diseases' resultant lesion is also common to Glaucoma: the PPA lesion. For that reason, this dataset is relevant for this work since it is, to the best of our knowledge, the only PPA annotated dataset available. Since the dataset was released for a competition, we only have access to the training subset (400 images and respective annotations).

3.6 Evaluation Metrics

This section describes the existing evaluation metrics used by researchers to benchmark their models, and estimating the performance for a given task. For the classification task, AUC and ROC ([42], [34], [47], [104], [70], [18]) are used to understand the model's capability of distinguishing between the output classes. Both sensitivity and specificity ([42], [34], [47], [104], [70], [18]) are used as a complementary metric to previous ones in cases of binary output classes. Accuracy ([42], [34], [47], [104], [70], [18]) is a classical evaluation metric for ML models, but can lead to a biased evaluation if the dataset is highly imbalanced (as happens in the majority of literature). For that reason, balanced accuracy ([42]) is used to overcome this issue, which averages over sensitivity and specificity. For the OD/OC segmentation part, most approaches use overlap metrics in order to understand the difference between the estimated structures and the ground truth, such as the Intersection-over-Union ([42]) and the Dice Index ([126], [73], [58], [18]). Fu et al. [42] also used pixel-wise sensitivity and specificity metrics.The CDR prediction can also be evaluated by calculating the error associated using the mean absolute error (MAE) ([73], [18]). Thakur et al. ([120]) also list other metrics that are not commonly used in the most recent literature.

3.7 Limitations and Challenges

Despite the many advancements towards making Glaucoma CAD systems efficient and effective, there are still some unsolved limitations and challenges. When training a new model, the first problem researchers find is the low amount of publicly available data, namely retinal fundus images. It is challenging to acquire clinical data since there is no screening strategy for early Glaucoma detection. Although some state of the art deep learning approaches can deal with a small dataset, this does not mean their behaviour in both segmentation and classification would be the same on large-scale datasets [120]. Having access to a large, well-labelled and balanced dataset would have a high impact on the performance and generalisation capability of the proposed models [82], increase data diversity and reduce model bias (due to, e.g. ethnicity/race, diseases severity, imaging protocol variances) [95]. Recent literature publications show that it is possible to use more sophisticated data augmentation methods to attenuate this limitation, such as transfer learning techniques [47], digitally generating artificial lesions inserted into normal images, inserting real lesions to other locations of normal or abnormal images, and generate synthetic data through generative adversarial networks (GANs) [95]. Nevertheless, even if data is balanced with regards to Glaucoma, there are few studies related to the impact of other morphological and pathological conditions (e.g., pathological or high myopia associated changes) in the current state of the art approaches. Ground truth is another influential factor for DL models since it established by experts' professional but subjective opinion. Certain borderline cases might have different diagnoses depending on the experts' experience [82].

Another issue is directly related to the features learned by the models, which are somewhat dependant on the input data. Due to the amount of extracted features and their complexity, not all

are considered for classification, since they would significantly impact both accuracy and performance. Some models are also limited at the start of the workflow when they only receive fundus images' ROI, leaving out features that could be relevant for Glaucoma detection [120]. When developing a Glaucoma CAD system, its deployment should be considered since a single system can have different practical applications, such as screening, triage, diagnosis or prognosis. Prospective studies should also be part of the process, since they provide analysis on the cost-effectiveness, efficiency and accuracy of the DL system in the clinical workflow during development, and ensure model refinement and quality assurance after deployment. Moreover, patients' data privacy and security and ethical and legal issues, are primary concerns for both the development and deployment of CAD systems.

OD and OC segmentation is the most widely used approach for Glaucoma detection, and researches face a few difficulties. Here, the immediate challenge is the invisible boundary between the two structures, which becomes even more challenging to identify when the image has low contrast. Moreover, fundus images also contain other morphological features, that even if not considered necessary for the classification task, can difficult the OD and OC segmentation 3.3.1. For example, the PPA has a boundary with the OD, which results in some segmentation models considering the PPA as part of the OD. The presence of blood vessels can also lead to more noisy segmentation results since they overlap with other structures. Nevertheless, these structures should not be ignored and seen as only barriers for the segmentation task because they can also be Glaucoma indicators. Consequently, several works suggest that further investigation should be conducted to correctly segment them and understand their value for the Glaucoma classification task.

Even though many proposed approaches state to have state-of-the-art performances, standardised Key Performance Indicators (KPIs) for measuring and comparing models are still mostly nonexistent [103]. In the past few years, some competitions have tried to create a unified framework that allows experts to compare models directly and better understand how and why they perform differently [87]. However, a lot of work is still needed.

Finally, models are also becoming more complex and challenging to understand. Interpretability is a crucial aspect of a DL model's implementation in a clinical scenario, not only for researchers but also for clinicians. For a clinical scenario, patients, technicians and clinicians must be familiar with DL-based clinical decision support systems. That way, they will be able to understand them better and accept them in the workflow more readily. Above all, a CAD system must be a support tool that helps clinicians reach a final diagnosis by removing certain burdens while providing them with credible and reliable explanations. For a model to be useful in a realistic scenario, it must be understandable by the people that will interact with it, making the field of explainable AI a very relevant one to be explored [122].

3.8 Glaucoma CAD systems on a Real-World Scenario

The ultimate objective of Glaucoma CAD systems is to be implemented in a realistic scenario, where they can serve as a tool to help clinicians diagnosis process. Some implementations of such systems have already been deployed in recent years and are available as end-to-end solutions, either for screening or clinical environments, while others are still undergoing further studies.

Zhao et al. [139] claim to have implemented the first App specially designed for Glaucoma screening, which can be installed on a smartphone and has shown good detection and classification accuracy in experiments. When the user uploads a retinal fundus image from the device, the App returns feedback in 4 parts: CDR analysis, NNR analysis, Glaucoma risk prediction and doctor's diagnosis display. The last one requires a professional doctor's interpretation. There is a DL model behind the interface that processes the uploaded image and returns its respective feedback. Firstly, after image enhancement and blood vessel extraction is applied, the OD is localised, and the fundus image is cropped to obtain the ROI. From the resulting image, both the OD and OC are segmented by a U-Net network enhanced with concatenating path (CP) and fusion loss function (FL), trained with the ORIGA dataset. Once the segmentation is complete, both CDR and ISNT related parameters are obtained, by calculating several morphological features such as the vertical OD and OC diameter or the ISNT-regions area and thickness. After applying feature selection methods, the selected features are used as the classifier's input, an SVM with 10-fold cross-validation. This classifier outputs the Glaucoma confidence level.

More recently, Martins et al. [79] presented another Glaucoma assessment pipeline, focused on space and time complexities. The dataset used results from merging different publicly available datasets (Origa, Drishti-GS, iChallenge, RIM-ONE r3, and RIGA) and applying augmentation techniques to reduce overfitting (e.g. blur and contrast normalisation). The segmentation task is performed by two U-shaped networks, based in the MobileNet architectures. One executes joint segmentation of OD and OC (GFI-SPP- Depth) and the other only segments the OD (GFI-SPP-Depth-simple). From GFI-SPP- Depth network segmentation, several morphological features related to the CDR and the ISNT are calculated. After the classification stage, these are shown to the user, contributing to the classification decision's interpretability. The GFI-SPP-Depth-simple network segmentation results are used as input to the classification network (GFI-C), created using MobileNetV2 feature extractor as a backbone, obtaining a Glaucoma confidence level. Contrary to the previous commercial solution, this one presents three interpretability measures to the end-user: intermediate pipeline results, morphological features and model activation maps. This system runs offline in mobile devices and achieved comparable or better results in both segmentation and classification tasks.

As for commercial solutions, Retinalize is a screening software that aids experts conduct eye diseases screening, one of them being Glaucoma. The algorithm behind the system detects signs of eye diseases through fundus imaging analysis, and can also be used as a clinical decision support system. The RetinaLize Glaucoma system[9] was introduced in May 2018 and the company aimed

[9] RetinaLize Glaucoma software web page

to make eye-screening accessible for the general public. This Glaucoma application asses the level of haemoglobin in the optic disc to measure the Optic Nerve Head (ONH) damage and calculates a Glaucoma risk index.

Eyenuk was founded in 2010 and studied since then developing a system that can be used for autonomous detection of several eye diseases. The launched screening system has been extensively validated for diabetic retinopathy, and the Glaucoma application is supposed to be launched in 2021/2022. The Glaucoma software should enable screening, grading, and reporting for Glaucoma directly at the point-of-care without the need for a human expert to grade the images. In a video[10] for the Glaucoma 360 event from the Glaucoma Research Foundation, Dr Kaushal Solanki, the CEO and founder of Eyenuk, lists three ways AI can help healthcare providers: error-checking, which means verifying the regular work done by experts for possible errors; autonomous, which automates certain work routines to allow scaling; and superpower, which enable otherwise impossible scenarios. The company's Glaucoma software framework is divided into these categories: EyeScreen, a component aimed at error-checking Glaucoma diagnosis; EyeArt, a component for autonomous eye screening; and EyeMark; a component that executes abnormality analysis and longitudinal monitoring to produce biomarkers for Glaucoma progression that would be otherwise unachievable by human experts. Moreover, it is also possible to see the six Glaucomatous signs that the software uses to detect a possible Glaucoma case.

3.9 Summary

This chapter describes several Glaucoma CAD systems, their limitations and challenges, the existing commercial solutions, and the known datasets. This literature review concludes that Glaucoma CAD systems are continually evolving, since new methods are proposed, and older ones are improved every year. Despite the approaches' diversity, it is safe to say that DL methods are up-and-coming and can bring Glaucoma CAD systems one step closer to a real scenario implementation. Nevertheless, these methods still have many limitations and challenges that must be addressed, such as the available clinical data or a model's interpretability.

[10] Video Link

Chapter 4

Literature Review: Generative Modelling

4.1 Generative Modelling

Discriminative models have been dominant in the Machine Learning field due to their ability to map a high dimensional input to a class label. On the other hand, generative models are less popular for two main reasons. Firstly, there is the difficulty of approximating many intractable probabilistic computations that arise in maximum likelihood estimation and related strategies. Secondly, there is the difficulty of leveraging the benefits of piecewise linear units in the generative context [48]. In essence, these models learn a probability distribution that resembles the original distribution of a data collection. Both discriminative and generative models use different strategies to perform the same task, calculating the target variable's conditional probability. Mathematically speaking, considering variables X and Y as the independent and target variables respectively, generative models estimate the distribution given by $P(X|Y)$ and $P(Y)$, and are then able to calculate $P(Y|X)$ using Bayes' rule shown in Equation 4.1. In some cases, this strategy is more effective because directly estimating the $P(Y|X)$ can be difficult [46].

$$P(Y|X) = \frac{P(X|Y) * P(Y)}{P(X)} \tag{4.1}$$

The following subsections start by describing the more typical generative models: Gaussian Mixture Models (GMM), Hidden Markov Models (HMM) and Boltzmann Machines (BM). Then they dive into deep generative models which are more powerful and thus very relevant for this work: Generative Adversarial Networks (GANs), Variational Autoencoders (VAEs) and Normalizing Flows.

4.1.1 Traditional Generative Models

4.1.1.1 Gaussian Mixture Models

In probability theory, a Gaussian distribution, also known as a normal distribution, is a continuous probability distribution for a real-valued random variable [46]. It is defined by a mean μ and a standard deviation σ, and is graphically shaped by a bell curve. It is possible to combine several distributions and create a mixture of N Gaussians by adding a parameter weight π to each of them, such that the sum of their weights is equal to 1. Equation 4.2 shows how to calculate the likelihood of observing x in a cluster i, given that δ are cluster i parameters, π_i is its weight, and $b(x|u_k, \rho_k)$ is its Gaussian density. Each distribution can be considered a cluster of data, and each of the weight's magnitudes represents the prior probability of finding that same cluster when considering all the data. Figure 4.1 shows an example of a Gaussian mixture.

$$P(x|\delta) = \sum_{k=1}^{N} \rho_i b(x|u_k, \rho_k) \tag{4.2}$$

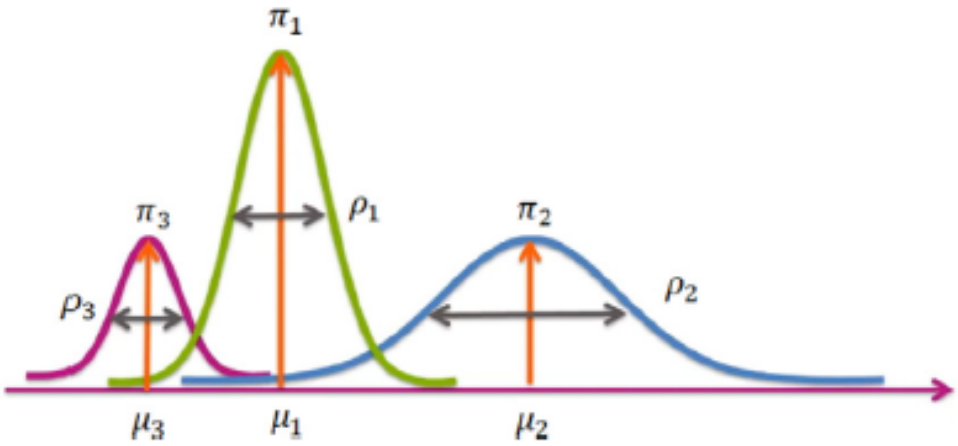


Figure 4.1: Mixture of three Gaussians [46].

Where π signifies the weight associated to the Gaussian and hence also the probability of the data belonging to the ith cluster or Gaussian, μ specifies the position of the Gaussian with the mean, ρ signifies the 'spread' of the Gaussian over the overall distribution by the variance.

GMMs are considered a generalization of K-means clustering algorithm. In a 2D space, the latter can only detect circular-shaped clusters (which are hyper-spheres in a higher-dimensional space), while GMMs can find oblong-shaped clusters. Nevertheless, it is more accurate to call GMMs density estimation algorithms, since they learn a formula in the shape of a distribution that allows new data generation. GMMs have been used for language identification systems, such as speech recognition [90] and accent recognition [123].

4.1.1.2 Hidden Markov Models

Hidden Markov Models (HMM) are statistical models widely used to model a system which is assumed to be a Markov process with unobservable ("hidden") states. These models generate sequences of states named Markov chains, where each state-transition has a corresponding probability and is dependant on the transition function of the state of origin. HMMs are a possible strategy to solve linear problems that involve time series or sequences and have similar traits to probabilistic non-deterministic finite automata. They also describe a probabilistic distribution over a non-finite number of possible sequences.

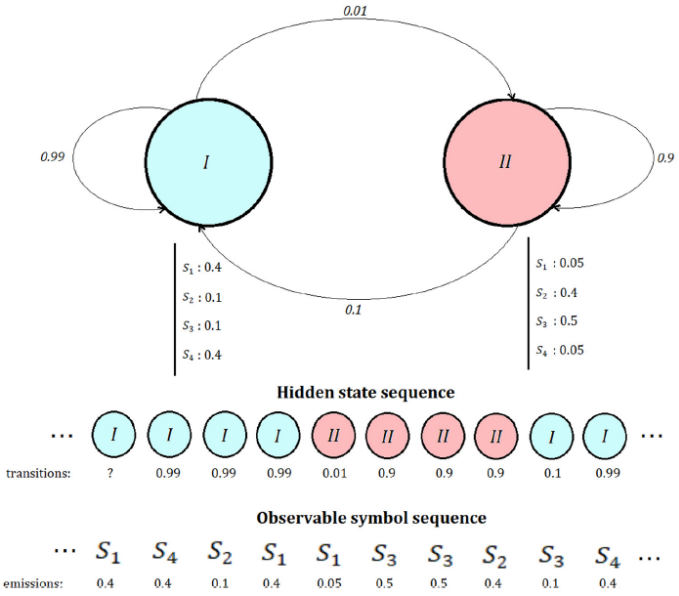


Figure 4.2: Simplified HMM with no initial and final states for the sake of simplicity [46].
Let there be a set of symbols defined by S = S1, S2, S3, S4. The two states that generate the Markov chain are labelled as I and II. State I generates sequences comprising S1 and S4 more frequently, whereas state II generates sequences comprising S2 and S3 more frequently (each state's symbol emission probabilities are stated below the respective state). All the state-transitions are implemented through arrows with their corresponding probabilities. Finally, the probability of the observable symbol sequence is the product of state-transition and symbol emission probabilities.

By observing Figure 4.2, we can see a probabilistic automata where each state as a certain probability of jumping to the next state depending on the residue or symbol emitted. Although we can see the final sequence, it is impossible to determine the specific Markov Chain that leads to it, hence the name, Hidden Markov Chains. HMMs have been used in several fields, as for example speech recognition [66], optical character recognition [6] and biological sequence modelling [39].

4.1.1.3 Boltzmann Machines

Boltzmann Machines (BMs) are undirected networks composed of many nodes linked together via weighted connections. They represent a class of unsupervised neural networks that generate data to form a system closely resembling the original one, usually a probability distribution. Nodes are

divided into hidden and visible ones, where the latter is used as the network's input and output, as we can observe in Figure 4.3. By feeding the visible nodes, hidden nodes are fed depending on their connections' weight throughout several iterations, which end up feeding back the visible nodes. A Markov chain is generated at the visible nodes layer, making each iteration a single Monte Carlo Markov Chain walk.

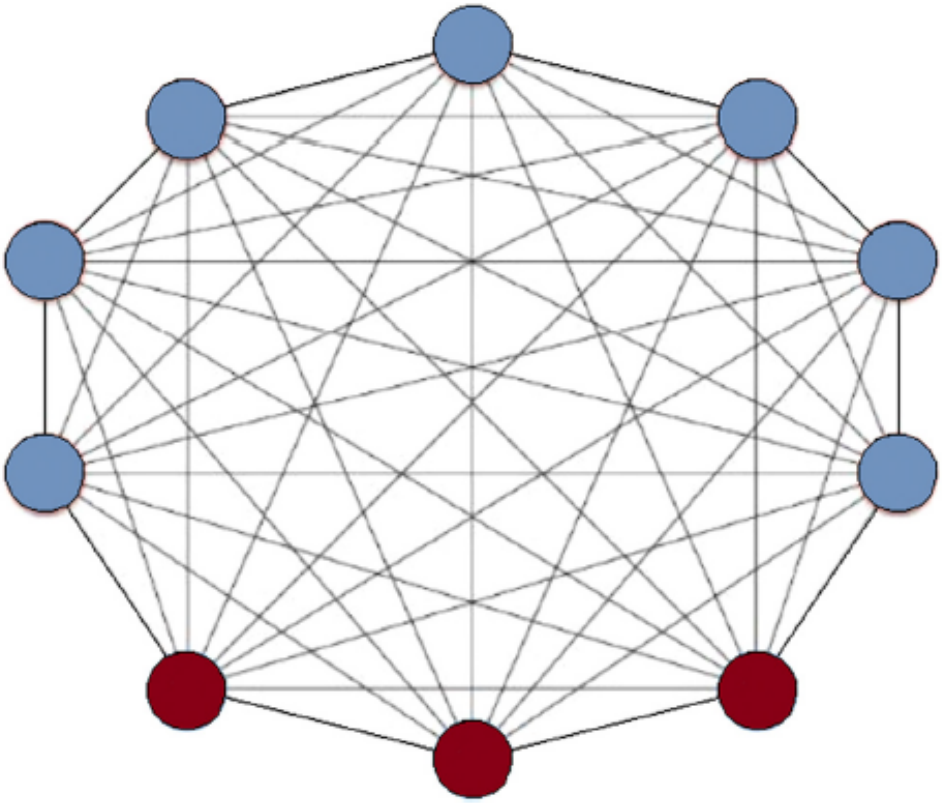


Figure 4.3: The Boltzmann machine where blue–grey nodes are hidden and maroon nodes are visible [46].

The most basic BM is simple but hard to work with due to the difficulty of sampling a network where all nodes are connected. For that reason, several BMs variations were proposed. The first one is Restricted Boltzmann Machines (RBMs) which do not allow visible-visible and hidden-hidden connections, reducing the network's complexity. RBMs were applied to collaborative filtering in the field of recommendation systems [99][45] and facial recognition [119]. Deep Belief Networks (DBNs) are also an extension of RBMs since they are a stack of several RBMs. However, this approach brings a few training problems, one of them being the "explaining away "[1]phenomenon. Some of DBNs applications are in breast cancer classification [3] and voice activity detection [135]. Deep Boltzmann Machines (DBMs) are networks where not connections are undirected, capturing hidden complex underlying features in the data such as speech and object recognition. Contrary to DBNs, these models use an approximate inference procedure that

accelerates learning and has a top-down feedback structure that allows them to deal well with ambiguous inputs. DBMs have shown success in state-of-the-art 3D model recognition [67], face modelling [28], etc.

4.1.2 Deep Generative Models

4.1.2.1 Generative Adversarial Networks

Generative Adversarial Networks (GANs) [48] are among the most well-known approaches for Deep Generative Modelling, being widely used in many fields of study. This method introduces an innovative internal adversarial training mechanism, composed of two neural networks, a discriminator and a generator, that compete in a minimax game. The generator learns how to create synthetic images that are as realistic as possible from a data distribution. The discriminator learns the distinction between a real image and a synthetic one. The generator's goal is to output synthetic images that trick the discriminator into considering them as authentic images. In contrast, the discriminator works as a classifier that outputs an image's probability of being real or fake. A great practical example of this architecture is the following. Consider the generator as a counterfeiter, whose purpose is to create fake make, and the discriminator is the police, which must distinguish legitimate money from counterfeit money. The counterfeit must make money that is as similar as possible to genuine money to succeed so that the police cannot correctly identify the fake money. Figure 4.4 shows the most basic GAN architecture, and we can observe that the discriminator's output is fed back into both models.

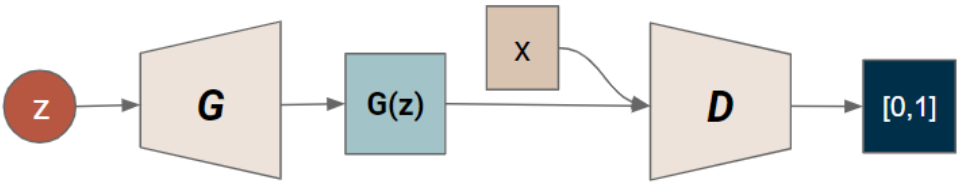


Figure 4.4: Generative Adversarial Networks architecture.

The value function for both players is shown in Equation 4.3, where x represents the real data, z and $p_z(z)$ denote the random noise input and its distribution respectively, $\mathbb{E}$ represents the expectation, $G(z)$ is the generator's output data, $D(x)$ is the probability of the discriminator considering x as real data, and $D(G(z))$ is the probability that the discriminator identifies the synthetically generated data. Both the discriminator and the generator are trained simultaneously, and the former tries to maximize the function, while the latter tries to minimize it. Once $D(G(z)) =$

[1] "Explaining away" occurs when one of the causes of an effect explains the effect entirely, which in turn reduces the probability of other reasons [46].

0.5, the discriminator cannot differentiate both distributions, and the model achieves the desired global optimum solution.

$$min_G max_D V(D,G) = \mathbb{E}_{x \sim p_{data}(x)}[logD(x)] + \mathbb{E}_{z \sim p_z(z)}[log(1 - D(G(z)))] \tag{4.3}$$

Despite their versatility in several applications and proved successes, GANs still present a few limitations that are not completely surpassed:

- **Mode Collapse**: GANs need to produce a wide variety of outputs. However, if the generator produces an especially plausible output or group of outputs, it might start to produce only that output. As a consequence, if at the same time, the discriminator gets stuck on a local minimum and is not able to find the best strategy, the generator will keep generating the same kind of output. Both conditions result in mode collapse, that is, a generator that rotates through a small group of output and a discriminator that is unable to get out of that "trap".

- **Nash Equilibrium**: In game theory, Nash Equilibrium refers to a solution of a non-cooperative game involving two or more players, where none have an incentive to change their strategy given what other players are doing. Although the original GANs definition stated that the generator and discriminator are competing until they reach a local minimum, they compete until the Nash Equilibrium is achieved. The Nash Equilibrium can coincide with a minimum, but it is not guaranteed that it always happens. For that reason, and since GANs are trained with Gradient Descent, which is designed to find a local minimum, the model might fail in convergence.

- **Model Evaluation**: Although there have been several proposals regarding metrics, GANs are challenging to evaluate due to their complexity. Moreover, since there is a large diversity of GANs applied to very different tasks, it is difficult for researches to find universal evaluation metrics.

Since Goodfellow et al. [48] proposal, several GANs derived models, and improvement techniques were published to solve the limitations of the original model. Pan et al. [88] published a survey with the recent progress on GANs and proposed three categories to distinguish different architectures. Nevertheless, researchers combine various aspects of these variants into a single network, to remove some limitations that a single variant might have.

- **Convolution Based GANs**: make use of Convolutional Neural Networks (CNN) to structure both the generator and the discriminator, having better performance in image feature extraction when compared to the original GANs that adopted Multi-Layer Perception (MLP) instead. Radford et al. [93] proposed a Deep Convolutional Generative Adversarial Network (DCGAN) that replaces the typical fully connected layers of the generator with deconvolution layers to increase performance in image generation tasks. Other examples are BigGAN [14], StackGAN [133] and InfoGAN [22].

- **Condition Based GANs**: introduce a conditional variable c, which could be additional labels, text, or other relevant data, that condition the generation process in both generator and discriminator. This helps solve the Mode Collapse problem described before since it gives some control to the researcher on the network input, usually a single random noise vector. StyleGAN [59] is an example of this architecture.
- **Autoencoder Based GANs**: merge two different generative modelling technique into a joint architecture, to maintain the advantages of both and remove their limitations. BiGAN [37] is an example of this architecture.

4.1.2.2 Variational Autoencoders

Before diving into Variational Autoencoders (VAE), it is vital to understand how Autoencoders (AE) work [46]. As we can observe in Figure 4.5, The basic architecture is comprised of 3 main components: an encoder; a middle layer **z**, known as bottleneck layer; and a decoder. The input flows through the encoder, transforming it into a lower dimensionality latent representation given by **Z**. The decoder uses that encoded representation to re-regenerate the original input. The Equation 4.4 represents the mapping function for encoding, where b is the bias, and W is the vector of weights. The reconstruction error is the distance between the original and synthetic data and is used as the loss value for improving the network by using backpropagation to adjust the weights. This results in the encoder having to condensate enough relevant information in the lower dimensionality representation, to improve the decoder capability of reconstructing the data. Autoencoders have several uses and are mainly used for compression tasks. They could also be used in supervised classification situations since the decoder can be replaced by a classifier that utilizes the encoded features extracted on **Z**.

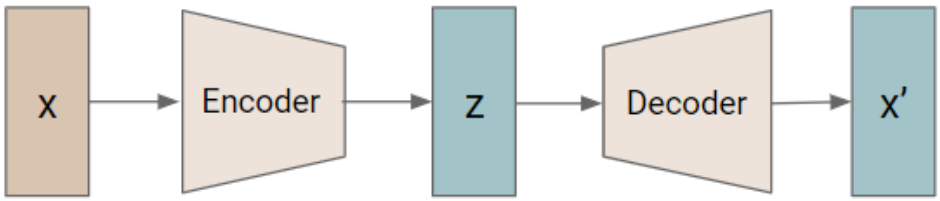


Figure 4.5: Autoencoder architecture.

$$Z = f(WX + b) \tag{4.4}$$

VAE [62] follow the same process as AE, but instead of mapping the input to a fixed vector, they map it to a distribution, as we can see in Figure 4.6. This means that the bottleneck **Z** layer is replaced by two vectors, one representing the mean μ and the other representing the standard deviation σ of the distribution. In this case, the decoder starts with a sampled vector layer that

samples from the previous bottleneck one. Due to this modification, it is no longer possible to use backpropagation due to the new sampling layer. The reparameterization trick solves this issue by adding a new parameter ε, which allows us to calculate the sampled vector layer without blocking backpropagation. This layer is given by Equation 4.5, where $\varepsilon \sim Normal(0,1)$.

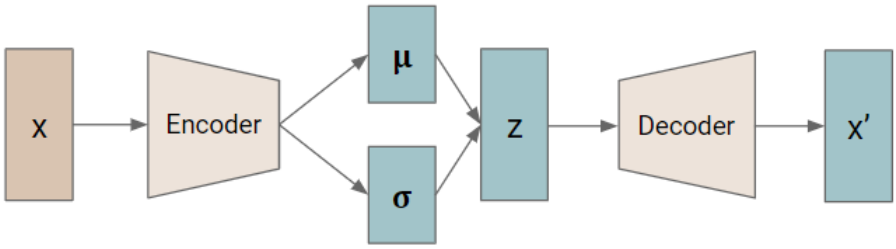


Figure 4.6: Variational Autoencoder architecture.

$$\mathbf{Z} = \mu + \sigma \odot \varepsilon \tag{4.5}$$

It is also necessary to update the original loss function, which results in Equation 4.6. The first term is still the reconstruction loss, which guarantees that the encoder outputs enough information to the bottleneck layer, allowing the decoder to reconstruct the original data correctly. The second term is the regularization loss, given by the KL divergence, ensuring that the generated distribution does not deviate too much from the Gaussian distribution.

$$\Lambda(\theta,\phi;x,z) = \mathbb{E}_{q_\phi(z|x)}[log p_\theta(x|z)] - D_{KL}(q_\phi(z|x)||p(z)) \tag{4.6}$$

Disentangled Variational Autoencoders [52] are one of the classes of VAE. The basic idea for disentanglement is to have independent neurons, each one learning a different feature. These models introduce a new adjustable hyperparameter β to the loss function that influences the latent channel capacity and independence constraints with reconstruction accuracy. In other words, this would make the model use a specific latent variable only if it benefits the training. Otherwise, the latent variable would remain equal to the initial distribution. Moreover, it is possible to evaluate how manual changes on latent variables are reflected in the network output by adopting a disentanglement strategy. This aspect can be investigated from an interpretability perspective.

4.1.2.3 Normalizing Flows

Normalizing Flows (NF) [96] are a technique used in ML that builds complex probability distributions from simple ones. They have been applied in generative modelling since they have appropriate properties for this scenario, as will be described below. As we can see from Figure 4.7, these models start with a simple probability distribution, for example, a Gaussian, which

flows through a sequence of invertible and differentiable transformation to create a more complex one.

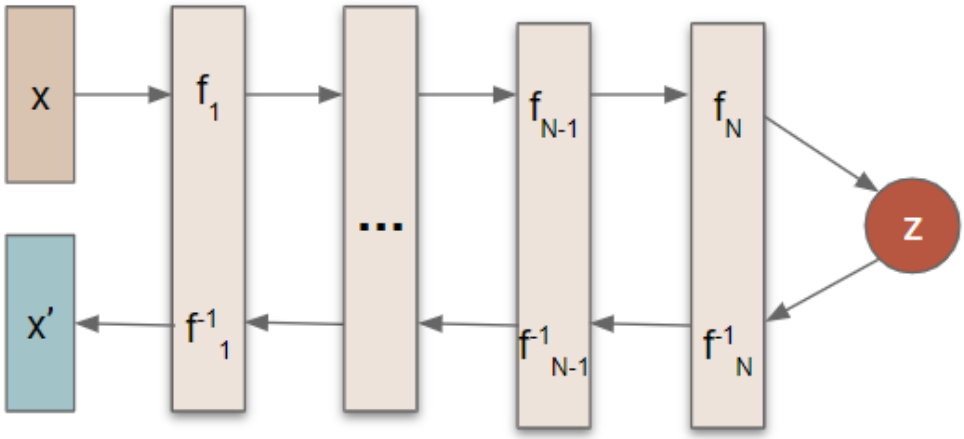


Figure 4.7: Normalizing Flow architecture.

From a formal perspective, consider a continuous random variable z that follows a distribution $p_\theta(z)$ according to a Gaussian distribution ($N(0,1)$), and a function f such that a new random variable $x = f(z)$. Then, $f_\theta(z)$ represents a sequence of N invertible (bijective) transformations, like the one in Equation 4.7, which make the overall transformation also invertible. It is possible to compute the probability density function of the random variable x according to Equation 4.8, using the change of variables formula, where the second term is the magnitude of the Jacobian of f^{-1}. The equation can be simplified to Equation 4.9 by replacing the invertible function by z, and it is easier to understand that the equation maps x to its inverse z, evaluating the magnitude of z over its distribution and multiplying it by some scalar magnitude. The magnitude of the Jacobian shows how the distribution expands and contracts along with the transformations.

$$f = f_1 \circ ... \circ f_{N-1} \circ f_N \tag{4.7}$$

$$p_\theta(x) = p_\theta(f^{-1}(x))|det(\frac{\delta f^{-1}(x)}{\delta x})| \tag{4.8}$$

$$p_\theta(x) = p_\theta(z)|det(\frac{\delta z}{\delta x})| \tag{4.9}$$

In the generative modelling context, the function f^{-1}, which will be referred to as g, is considered a generator, since it moves from a simple base density p_θ to a final complicated one. To generate a new data point x', one can sample a value from the base distribution of z, and apply the generator: $x = g(z)$. Contrary to the generative direction, we have the normalizing direction, which

moves from a complex distribution to a simpler one through the function f, both represented in Figure 4.7.

In general, NF should at least satisfy the following three conditions to ensure they are practical to work with:

- **Be invertible**: This allows researchers to use both the normalizing direction to compute the likelihood and the generative direction to sample and generate data;
- **Be expressive**: This ensures that the model can transform the simple distribution and approximate it as much as possible to the original one;
- **Be computationally efficient**: This is important for both computing f and g, but also crucial when calculating the determinant of the Jacobian.

The two most widely used NF architectures are Coupling and Autoregressive flows, which have their popularity due to their architecture, allowing invertible non-linear transformations. An overview of these and other variations is given in the list below [63]:

- **Elementwise Flows**: apply non-linear elementwise transformations, which means that each element in the flow is independently processed. They do not take into account possible correlations between elements;
- **Linear Flows**: apply linear transformations to a combination of variables, but have due to that linear restriction, they have limited representational power;
- **Planar and Radial Flows**: apply non-linear transformations but are not widely used in practice, since their inverses are hard to compute;
- **Coupling Flows and Autoregressive Flows**: use coupling functions has buildings blocks and have high expressive power;
- **Residual Flows**: use invertible residual networks that try to discretize the continuous dynamical system;
- **Infinitesimal Flows**: contrary to the residual flows, these flows try to learn the continuous dynamical system in two ways: infinitesimal, which comes from ordinary differential equations (ODE) and continuous, which comes from stochastic differential equations (SDE).

Glow [61] is a recent work that proposes a new flow built on the NICE [35] and RealNVP [36] flows. The models adopt a multi-scale architecture. Each step of flow consists of an *actnorm* layer, which increases performance for large images; an invertible 1 x 1 convolution layer; and a coupling layer. The authors concluded that these last two aspects contributed to a faster model convergence and a lower negative log-likelihood during the evaluation. Compared to its precursors, Glow is also stated to be the first likelihood-based model that can efficiently generate high-resolution natural images, such as human faces.

Lugmayr et al. [76] applies flows to the super-resolution problem and proposes SRFlow, a network capable of accurately learning the distribution of realistic high-resolution images from low-resolution ones. Besides presenting the state-of-the-art super-resolution quality, the model is can also be used for image denoising and restoration. The authors also developed techniques for image manipulation and editing and evaluated the approach with perceptual and reconstruction-based metrics.

4.1.2.4 Generative Modelling Techniques Comparison

From the previous sections, one can conclude that generative modelling has very diverse techniques, each with specific advantages and disadvantages, that make them more suitable for particular situations and individual goals. On a higher level, both traditional and deep generative modelling techniques are useful on different occasions. Although one might be tempted to use DL approaches' latest fashion, that is not always the best and most efficient solutions. Both categories should remain relevant depending on the context where they are used. Nevertheless, deep generative modelling techniques are the ones that have a more significant potential of showing greater results when used in contexts of complex input data, since they have the mechanisms and computational power needed to deal with higher dimensionality feature spaces (such as images), and also higher amounts of data. Although more powerful, the technique complexity brings issues regarding the understanding of those same models. In a realistic situation, it is not enough to provide useful quantitative and qualitative evaluation metrics; it is also necessary to give the reasoning behind model predictions. This is crucial when models are applied in critical contexts, such as the healthcare field.

Nevertheless, deep generative modelling approaches seem to hold the most potential in the context of retinal fundus images, since it involves high-resolution retinal images, that contain complex and delicate structures, essential for the models' performance. From this premise, the three previously presented deep generative modelling techniques are further compared below.

Starting with the most widely researched of the three techniques, GAN bring several advantages to the table. Adversarial training applied in GAN is very useful because it means the network can model the underlying distribution of plausible images only from training data without manually interacting with complex parameters. Moreover, GAN are the technique that generally produces the best quality images, being less blurry when compared to others. GAN also have probably the most considerable amount of published literature work, which resulted in many improvements compared to their initial version. With all the new variations, it is possible to generate images with even higher quality and have more stable training. However, GAN are prone to suffer from the *mode collapse* in certain situations and relying on the discovery of the Nash Equilibrium to reach convergence is harder than minimizing a typical objective function. Isola et al. [55] also showed that for image-to-image tasks, the generator ignores the random vector given as network input, which signifies that GAN mappings are deterministic. Moreover, explainability and interpretability are not of this technique's strengths, and density can only be estimated implicitly.

On the other hand, VAE allows explicit density estimation, and the latent representation can be precisely controlled to fit a specific context. Furthermore, this technique can achieve a high value for the data likelihood and for that value to be very similar to the true posterior distribution. Compared to GAN, VAE have the likelihood lower bound, which can be used as a measurable objective during the model training. Despite all the pros, VAE are not able to produce images with the level of quality of GAN images, and even with a high likelihood value, images are not guaranteed to be realistic. Another limitation imposed by VAE is the posterior distribution modelling. It is limited from the beginning to some specific distribution, which might not be similar to the true data distribution.

Normalizing Flows are the most recent technique applied to generative modelling from all three and have received little attention than the previous methods. Nevertheless, it has shown remarkable results, some even more promising than state-of-the-art GAN or VAE [76]. If we consider the case of conditional generation, Normalizing Flows are more stable than their equivalent CGAN. Zhu et al. [140] proposed CycleGAN having to carefully tune the eight loss function terms and balance the generator and discriminant. Normalizing flows only have a single network and a single loss, simplifying the hyperparameter tuning and the training. Furthermore, as shown in [76], the output in flows is usually more consistent with the input than in GANs, due mainly to the later unsupervised loss that encourages image hallucination. Flows are also the only technique from the three that allow explicit tractable density since a bijective function defines each transformation. This is very relevant to understand how the network models the distribution.

As for evaluating generated images, independently of the technique used, visual inspection is one of the immediate evaluation techniques used. Even if a human observes a sharp image and considers it as visually "realistic", it does not mean the model does a good job generating realistic images in the training data context. For example, let's consider a model trained on a dataset of trees and houses. If the model ends up only generating images with trees and with no houses, but with high quality and that "look" realistic, the human evaluator would not be able to identify that issue if he was not aware of the training data. For that reason, it is essential to invest not only in higher-quality generated images but also in new ways of unbiasedly evaluating models.

Finally, combining these techniques to eliminate one or more limitations of using a single approach is possible. BiGAN [37] are an example where VAE were combined with GAN. This work also shows that it is possible to improve VAE using Normalizing Flows [118].

4.1.2.5 Deep Generative Modelling for Retinal Imaging

CAD systems in the healthcare field have progressed a lot in the past decade, mainly due to their potential to help clinicians detect or diagnose diseases. Medical image interpretation is one of these systems' uses, which is very relevant in Ophthalmology. In this area, state-of-the-art techniques revolve around DL architectures, which have achieved acceptable performance levels in several tasks, such as retinal imaging segmentation and eye diseases classification. However, these models require large, diverse and high-quality datasets that help training them and play an essential role in the approaches' validation. Specific pathologies have a relatively low prevalence,

reflecting in the datasets as a class imbalance. Moreover, since we are dealing with real patient data, legal and privacy limitations need to be considered when designing such models.

Generative Modelling, and more specifically Deep Generative Modelling, is one of the most recent and innovative solutions that has shown success in several fields, including in retinal imaging in Ophthalmology. Due to their popularity, GANs and their derivations are the most common architecture in retinal imaging synthesis literature. Still, a few approaches also include VAE modules that improve overall performance. Despite being a challenging task due to the complexity of the eye's anatomy, these approaches have provided valuable and consistent results in image synthesis [11, 31, 132, 138], segmentation [105, 115, 77, 111, 49, 137] and super-resolution [77] tasks.

Costa et al. [32] approach involved using pairs of real vessel networks and their corresponding retinal fundus images to train a model so that it could learn how to generate new data from a given vessel network. The model employed a GAN, which combined the adversarial loss with a global L1 to produce sharper results and was trained with 614 pairs of images from the MESSIDOR dataset. They used a general U-net architecture to segment the vessel networks from those images, trained with images from the DRIVE dataset. Image quality was evaluated using Q_v score, focused on the contrast around vessel pixels, and Image Structure Clustering (ISC) metrics, focused on a global evaluation. Costa et al. [31] proposed a follow-up work that removed the model's dependence on the vessel network availability. This was achieved by implementing an Adversarial Autoencoder (AEE), which would learn a distribution representing the vessel networks by sampling it into a multi-variate Gaussian distribution. This allowed creating an end-to-end system composed of an AAE and the previously created GAN that would generate the vessel networks and use them to generate a retinal image. Both models were trained jointly. To evaluate the synthetic images, besides using metrics to look into their quality, they were also used to prepare the AEE for the segmentation task. If trained with only synthetic images, the model showed a slight decrease in performance relative to a real image trained. When trained with both natural and artificial images, the model's performance decreased considerably.

Guibas et al. [49] also proposed a two-stage pipeline, which generated retinal networks from noise using a DCGAN architecture, and then created colour fundus images using a Conditional GAN (cGAN). The first GAN was trained using the DRIVE dataset, while the second one was trained using images from the MESSIDOR dataset. A U-net segmentation network was trained with that same data and evaluated synthetic images reliability using F1 score on images from the DRIVE dataset. Variability between the original images and synthetic ones was assessed using a Kullback-Leibler (KL) divergence score.

Beers at al. [11] applied Progressive Growing GAN (PGGAN) to the retinal imaging synthesis task, more specifically images associated with retinopathy of prematurity. Initial training resulted in low-resolution images (4 x 4 pixels) that progressed into 512 x 512 pixel images. The network employed the Wasserstein loss. Segmentation maps were also used as network input, and they enhanced the final images detail level. Vessels quality was evaluated using a segmentation technique trained on reading images. The image variability was also assessed through a network

that encoded the synthetic images to predict a latent vector for each image, enabling latent space evaluation while interpolating between images.

Zhao et al. [138] proposed a network called Tub-sGAN that can synthesize several realistic-looking retinal images from the same vessel network. This model can learn from a minimal set of images, 10-20; hence the authors trained it with 20 DRIVE images, 10 STARE images and 22 HRF images, resized to fit the network and improve performance. The generator is built using an encoder-decoder strategy paired with U-net style skip connections, which inherently introduces a noise code and allows the model to retain the main vessels structure. The authors also added image style transfer to the network by adding another training input, which conditions the resultant image to a "particular style". Consequently, the model's loss was based on the style, content and total variation loss. The network was extensively evaluated using Patch-based CNN baseline and DRIU baseline segmentation methods, and the authors state that 90% of the generated images are realistic. The same authors also published another work [137] that was specifically focused on generating retinal images suited for the segmentation task.

Iqbal et al. [54] proposed another GAN architecture called MI-GAN, which generated both medical images and the respective segmented masks. Similarly to Zhao et al. [138], the authors applied a style transfer variant. The models were trained using the DRIVE and STARE datasets. The generator convergence and overall training time were reduced by updating it twice as much as the discriminator.

Yu et al. [132] proposed a new preprocessing pipeline named multiple-channels-multiple-landmarks (MCML), which improves image synthesis by combining vessel network, optic disc and optic cup images. The performance was evaluated by comparing it to a single vessel mask input on the DRIVE and DRISHTI-GS datasets, implementing several Pix2Pix and Cycle-GAN architectures. The authors concluded that the Pix2Pix based model with ResU-net generator achieved superior performance compared to other GAN, and it can synthesize realistic fundus images. Moreover, the MCML preprocessing pipeline also seems promising in the context of Glaucoma CAD systems.

Diaz-Pinto et al. [33] investigated retinal image synthesis applied to Glaucoma assessment based on DCGAN. 86926 images were merged from fourteen public datasets, not all annotated for the Glaucoma classification task. All the images were cropped around the optic disc since that is the ROI most relevant for Glaucoma assessment. Besides a DCGAN, the authors also trained an SS-GAN based on recommendations from another source [29]. For both quantitive and qualitative evaluation, a new dataset was created, composed of 100 synthetic images from the DCGAN, 100 synthetic images from the SS-DCGAN, 100 images from a state-of-the-art method [31]. Real and artificial images were compared with t-SNE to evaluate the feature differences, pixel proportion of vessels, optic disc and background, and Mean-Squared-Error (MSE) comparison. The authors concluded that SS-GAN shows lower performance than the DCGAN for the image synthesis task, but they still evaluated the discriminator for the Glaucoma assessment task.

4.2 Semantic Image Editing

Semantic image synthesis and manipulation is a popular research topic in ML and Computer Vision. Recent advances in generative modelling led to the creation of power image editing tools. The image-to-image translation problem is one of the sub-fields of this topic. It consists of having a set of source images, like horses, and a set of target images, like zebras, but they are not explicitly paired with one another in the training set. The goal here is to translate one possible representation of an image to another, given sufficient training data.

Qiu et al. [91] explored the impact of semantic manipulation on Deep Neural Networks (DNN) predictions by generating "unrestricted adversarial examples". The authors proposed SemanticAdv algorithm that utilizes disentangled semantic factors to generate adversarial perturbation, that induces the learner towards "adversarial" targets. These perturbations are more controlled since semantic attributes guide them. The experiments involved testing the method in the face recognition and street-view images domains. Regarding the former, targeted attacks at real-world face verification services were performed, showing a high success rate.

Isola et al. [55] investigated the use of Conditional GAN (CGAN) to solve the image-to-image translation problem. This is a condition-based generative model, with a "U-Net" based architecture for the generator, and a convolutional PatchGAN classifier for the discriminator. The approach was evaluated in different experiments in various tasks and datasets, to test how widely applicable it would be. Results show that this is a promising approach for various image-to-image translation tasks, especially those involving highly structured graphical outputs. However, this approach has a considerable limitation: it requires paired training data between the source and target domains, which is very rare and hard to get.

Cycle Consistency appears as a solution that can enable Unpaired Image-to-Image translation techniques. It is based on the idea of using transitivity as a way to regularize structured data, and it has been used for many decades in other situations, such as visual tracking or language translation. Practically speaking, cycle consistency involves going back and forward between domains to force consistency when moving from one to another.

Zhang et al. [134] proposed a network called HarmonicGAN that learns bi-directional translations between the source and the target domains. The goal is to use similarity-consistency to have inherently consistent samples, in a similar setting to CycleGAN. The algorithm behaves harmonically along with the circularity and adversarial constraints to learn dual translation between domains, resulting in improved CycleGAN due to better transformation consistency.

Zhu et al. [140] also investigated the unpaired image-to-image translation problem using cycle consistency. Still, their proposal was not task-specific, nor demanded the input and output to lie on the same low-dimensional space. The proposed algorithm is compared with paired and unpaired image-to-image translation state-of-the-art approaches and obtains better classification performance in various applications.

Contrastive Learning is an alternative approach to Cycle Consistency, which does not rely on going back and forwards between the source and target image. Instead, it uses image patches

from the entire dataset as positive or negative patches. It applies patchwise comparisons to ensure that the patches from different images on the same location are similar to one another but different from the others.

Park et al. [89] proposed a method that uses contrastive learning to encourage two elements (corresponding patches) to map a similar point in a learned feature space. Patches are compared by comparing the different resolution of the feature maps as they are processed by the generator's encoder, using a patch noise contrastive estimation (PatchNCE) loss.

Chen et al. [23] present SimCLR is an algorithm that uses contrastive self-supervised to leverage unlabeled datasets for representation learning. Self-supervised learning is a subtype of unsupervised learning based on the idea of creating a supervised learning task automatically from unlabelled data. According to contrastive learning, SimCLR compares the differences between positive and negative pairs. The positive pairs are generated through Composition of Data Augmentation. This technique chooses the adequate traditional augmentation techniques to apply to an image, while negative pairs are the dataset's remaining images. This method outperformed other unsupervised learning methods and even reached ResNet50 supervised learning level performance when scaled up four times.

4.3 Summary

This chapter shows state-of-the-art Generative Modelling, focusing on Deep Generative Modelling approaches relevant to this dissertation's work due to their potential with image data. Semantic Image Editing techniques were also described, and there are very promising approaches that fit this dissertation work.

Chapter 5

Literature Review: Explainability and Interpretability in Machine Learning

This chapter describes the current status of Explainability and Interpretability in Machine Learning, paired with a state-of-the-art research about approaches relevant to this dissertation. Section 5.1 gives an overview of the Explainable AI (XAI) field. Section 5.2 lists the different literature taxonomies to classify Interpretability approaches, while Section 5.3 lists a few of the existing interpretability evaluation metrics. Section 5.4 describes the state-of-the-art interpretability techniques, with a highlight for Case-based reasoning approaches in the context of this world. Section 5.5 discusses the idea of a Glaucoma CAD system with Explainable Decisions.

5.1 Overview

Machine Learning (ML) is becoming more prevalent in society, not only for research purposes but also for real scenario applications. Significantly, Deep Learning (DL) methods are gaining ground due to increased computational power and available data collections. Not only do these systems show better results, but theyhave also grown in complexity. In a few fields, failure is considered critical, since it can lead to catastrophic consequences [38], such as in the healthcare industry. Despite ML systems' current success, other questions have grabbed researchers' attention, one of them being the interpretability and explainability of these systems. Questions like "who is accountable if things fail?" and "How can we explain why something went wrong?" still don't have a sure answer. For that reason, the topic of Explainable Artificial Intelligence emerged as a new field of study and has become one of the hotspots in the research community.

The first necessary step that needs to be addressed is the notion of concepts around explainable artificial intelligence. Firstly, "explainability" and "interpretability" are two core terms that do not have an agreed-upon meaning, and are used interchangeably across the literature. Nevertheless, they are tied concepts: "interpretable systems are explainable if their operations can be understood

by humans" [16]. Doshi-Velez et al. [38] defined interpretability as "the ability to explain or to present in understandable terms to a human". Miller et al. [80] defines it as "the degree to which an observer can understand the cause of a decision". One could also try and use other concepts such as "transparency" or "accountability" to define the previous ones, but then those would have to be defined in the context of ML as well. For that reason, it is relatively safe to assume that interpretability is related to the perception human have over some information and how they reason about it. Moreover, interpretability is also not a "quantifiable" metric, as common performance measures such as accuracy are. Other auxiliary criteria [38] also depend on the notion of interpretability to be evaluated:

- **Fairness/Unbiasedness**: Ensure there is no explicit or implicit discrimination against certain groups;
- **Privacy**: Ensure that the methods protect any sensitive information in the data;
- **Reliability and Robustness**: Ensure that algorithms can have a satisfiable performance with perturbation in parameters or inputs;
- **Causality**: Ensure that a certain perturbation leads to a certain output in the real system;
- **Usability**: Ensure methods provide information that aid users to accomplish a given task;
- **Trust**: Ensure systems have the confidence of human users that interact with them.

A question that could be asked is "Why interpretability?", "Where does the necessity for interpretability come from?". Doshi-Velez et al. [38] start by stating that explanations are not necessary in every scenario, for one of two reasons: (1) a system does not have significant consequences in case of unacceptable results or (2) the problem is well-studied and validated in real scenarios, and the decisions made are trusted even if the system is imperfect. On the other hand, the authors also argue that interpretability necessity comes from incompleteness in the problem formalisation, which either blocks further optimisation or evaluation of a system. One should not confuse incompleteness with uncertainty though: "the fused estimate of a missile location may be uncertain, but such uncertainty can be rigorously quantified and formally reasoned about". Explanations are an interpretation tool that allows us to understand the gaps in problem formalisation, ensuring they are visible.

CAD systems' interpretability is of major interest in the healthcare field due to a clinical diagnosis's critical nature. A system must be transparent, understandable and explainable to gain clinical experts, regulators and even patients. A new barrier was recently imposed on the typical "black-box" models: the new regulations like the European General Data Protection Regulation (GDPR), which requires a system to have re-traceable decisions. [109].

5.2 Taxonomy of Interpretability approaches

Several taxonomies have been proposed to classify interpretability methods, using different criteria [16, 109]. Not only are non of these criteria absolute, but also can lead to an overlapping or non-overlapping classification of specific methods.

5.2.1 Model-specific vs. Model-agnostic

Model-specific interpretability involves methods built for a specific model because they use particular parameters on the model. Model-agnostic interpretability is applied in a post-hoc manner. Its use is not restricted to one specific model architecture, relying on such a model's input and output.

5.2.2 Global Methods vs Local Methods

Global methods are focused on understanding the overall model's knowledge, its training and the data. On the other hand, local methods are specific for a single outcome of the model and explain a particular prediction.

5.2.3 Pre-model vs In-model vs Post-model

Pre-model interpretability techniques are only applicable to the data collection, thus being model-independent. This mode is focused on analysing the available data to understand fundamental properties that can be relevant in the future model choice.

In-model interpretability is closely related to intrinsic interpretability. This refers to models that inherently provide explanations for their decisions, without the need of an external method or tool to interpret them.

Post-model interpretability is applied after building the model, similar to the Post-Hoc methods. In this case, the methods used are external to the model and improve it by providing explanations.

5.2.4 Intrinsic vs Post-hoc

Intrinsic interpretability refers to inherently interpretable models; that is, they explain their decision by themselves. One could say that the explanations presented are a consequence of the model's learning and help answer "how a model works". On the other hand, Post-hoc interpretability involves explanations generated outside of the model, usually by a model specifically designed for that effect. These explanations result from a "replication" of the original model's behaviour.

5.3 Interpretability Evaluation

As we could see from the previous sections, there are no mathematical definitions for interpretability, neither absolute criteria to categorise its methods. Consequently, there is no uniform framework to evaluate such methods and compare them fairly to each other. Nevertheless, some works try to list useful metrics to measure and evaluate the current ML systems' "interpretability level". Doshi-Velez et al. [38] proposed a framework that splits interpretability evaluation into three distinct levels:

- **Application-grounded** evaluation: implicates conducting user experiments within a real application. Evaluating a system on-site is probably the best way to ensure it works according to expectations and provides useful input to the domain expert involved, regarding the intended task;
- **Human-grounded** evaluation: involves conducting simpler user experiments that maintain the essence of the real application. This evaluation model is handy when the target scenario entails challenging evaluation conditions. Moreover, domain expertise is not needed, which means the candidate tester population is broader than the previous point.
- **Functionally-grounded** evaluation: does not require human experiments. It is executed using a formal definition of interpretability as a proxy for explanation quality. It is most appropriate for systems that are still under development, or end-user experiments are unethical.

5.4 Interpretability Techniques

Several surveys in the literature summarise ML interpretability techniques, each using a different or several taxonomies to classify them. The approach used to highlight these techniques is also varied. Tjoa and Guan et al. [122] provide a more technical overview of each existing method. Stiglic et al. [117] use the global vs local and specific vs agnostic taxonomies to distinguish several techniques and then presents their usage in the healthcare context. Singh et al. [109] also reviews these techniques and goes even further when describing their real-world applications, giving specific examples of each one. Section 5.4.1 provides an overview of more generic interpretability techniques, and Section 5.4.2 explains case-based reasoning approaches.

5.4.1 Overview

Elshawi et al. [40] proposed four quantitative indicators for measuring the quality of explanations in various interpretability techniques, that can be used as a unified quantitative measure framework: similarity, bias detection, execution time and trust. To evaluate these indicators, six popular local model agnostic interpretability techniques were employed: LIME, Anchors, SHAP, LORE,

ILIME and MAPLE. Moreover, three other axioms were used to relate an instance to its corresponding explanation: identity, stability and separability. Definitions for these concepts can be found in the published paper and an overview of each technique. The experiments involved two types of datasets, tabular and text datasets, divided into different experiments according to the data domain. The results showed no particular technique that achieves the best performance in all the metrics across all datasets. For that reason, the authors conclude that it is essential to specify the focus of each evaluation metric and to understand its strengths and weaknesses on different scenarios.

Selvaraju et al. [102] proposed a method to localise input regions relevant to model predictions, Gradient-weighted Class Activation Mapping (Grad-CAM), which produces visual explanations. These explanations result from the combination of the localisation technique's output and high-resolution visualisations. The results were compared with Guided Backpropagation, which could also be combined with the proposed technique to improve the method. The experiments also involved testing the technique's ability the help investigate and explain classification mistakes.

Smilkov et al. [113] proposed Smooth Class Activation Mapping (SmoothGRAD), a method based on gradient interpretation that improved gradient-based sensitivity maps sharpness. This technique is beneficial in image classification systems, where sensitivity maps are regularly used to identify the image regions that were the most influential to the final classification [102]. The authors present two complementary strategies that can improve these maps: the first one is averaging maps made from small perturbation of a particular image, followed by a new training on data perturbed with random noise. The results were promising and also suggested other avenues for future research, such as investigating the reasoning behind noisy gradients or methods to create systems with smoother class score functions.

Chen et al. [23] introduced Concept Whitening (CW), a mechanism that alters a given layer to force latent space disentanglement, which is useful at the bottleneck layer of a network. This method falls under the intrinsic interpretability category and does not hurt the predictive performance of the model. The CW module can be applied to any layer in a CNN to align the latent space axes with interest concepts. This allows researchers to understand how the model gradually learns those concepts along several layers. The authors also conducted a quantitative evaluation of the resulting concept axes and compared them to other concept-based NN methods. The adoption of CW resulted in a higher value of concept purity than other posthoc methods, which means it provided better latent space disentanglement, and consequently can improve practical insights executed on the network.

Schutte et al. [100] highlighted that the popular heatmaps or sensitivity maps are a limited explanation method since they provide the location of predictive features without explaining how they contribute to it. They presented a new method that can be applied to any "black-box" model with image data, showing how a particular image can be modified to produce different predictions. Like the StyleGAN architecture, this technique identifies the optimal direction in the latent space that can create a change in the prediction, enabling more powerful explanations than the ones provided by typical heatmaps like Grad-CAM [102]. The authors developed a StyleGAN that

generates small synthetic transformations in the original images, which allows the user to observe the possible progression towards a different outcome. Besides building clinicians' trust in the model's predictions, the method can discover new relevant bio-markers and even reveal potential biases.

5.4.2 Case-based Reasoning Approaches

Case-based Reasoning (CBR) systems provide explanations from previous examples or cases using a *retrieval*, *reuse*, *revise* and *re-train* cycle [64]. The simplest implementation of this strategy starts with a query-case, which is the data entry that will be classified, which is used by the *retrieval* step to match features from other cases using an ML algorithm like k-nearest-neighbor (k-NN). These retrieved cases can be used as similar examples, which are from the same class as the query-case and have similar features, or counterfactual examples (also counterexamples), which are from a different class of the query-case but have enough distinct features not to be considered of the same class as the query-case. CBR is claimed to have "natural" transparency since its reasoning is similar to a human expert since it is frequent to use past cases to understand new ones.

Keane and Smyth et al. [60] proposed an approach focused on counterfactual cases generation, exploring the ideas of counterfactual potential and explanatory coverage of a case-base. Authors claim that counterfactual explanation is intuitively more explanatory that the popular factual one, and supports this affirmation with works from fields outside of XAI such as Psychology. The technique identifies useful candidate counterfactuals and reuses their patterns to generate even better counterfactuals adapted to the original query-case, which helps deal with challenge like conterfactual sparsity and plausibility.

More recent approaches do not limit themselves to the direct feature comparison. Prototypes are a concept that can also be used as an explanation tool. Li et al. [69] proposed an architecture that contained an autoencoder and a particular prototype layer, which stores a weight vector that serves as an encoded version of the input. The encoder is used for comparisons within the latent space while the decoder is used to evaluate the learned prototypes. The training objective encourages both prototypes and encoded inputs to be similar. Since the prototypes are learned during training, the final explanations result from the natural learning process and are therefore faithful to the network computations. Experiments showed that prototypes are very useful because they give essential insight into the network decision process, the relationship between different outcome classes, and in the learned latent space.

Ming et al. [81] also presented a model with natural explanations derived from CBR called ProSeNet, aimed explicitly at sequential data. The prediction is obtained by comparing the inputs and prototypes, enabling the model to provide interpretable representations. Similar to the previous work, ProSeNet architecture comprises three parts, the recurrent sequence encoder network, the prototype layer and two more layers (fully connected and softmax layer) to output the probabilities in the multi-class classification task. The significant difference is on the prototype interpretation since instead of using a decoder, this network has a projection step which ensures that

the prototypes are meaningful. Moreover, the network and the prototypes can be refined by domain experts if they find the need to, without being necessary to have any underlying model knowledge. The experimental evaluation consisted of four case studies, each one with a real-world sequence dataset from a different domain. These experiments confirm the prototypes' reliability quantitatively, and user experiments show that they provide understandable and accurate prototypes for predictive explanations.

Chen et al. [20] also proposed a network similar to the previous ones called prototypical part network (ProtoPNet), which finds prototypical image parts and combines evidence to reach a classification outcome. The method was tested in two domains, bird species and car model identification, not sequential type datasets as the previous network. The authors claim that this network provides a level of interpretability that surpasses other interpretable deep models and compares with other baseline models trained with the same augmented dataset of cropped bird species images to ensure fairness.

5.5 Towards Glaucoma CAD Systems with Explainable Decisions

From Chapter 3, one can conclude that Glaucoma CAD systems have progressed throughout the last decade, and can achieve remarkable performance results by applying state-of-the-art ML techniques. However, there have not been many advances concerning those same systems' explainability, which are crucial for deploying such systems in realistic scenarios. We consider that one of the ways a model can become more "explainable" is by providing explainable decisions that can be understood by clinical experts and help them in the Glaucoma diagnosis. The question that could be asked is, what are the requirements for Glaucoma CAD systems to have explainable decisions?

There are very few works that are solely focused on exploring explainability in Glaucoma CAD systems. Chang et al. [19] proposed an adversarial explanation based method to explain the reasoning behind the "black-box" model, applied to Glaucoma detection, along with critical morphological features such as Cup to Disc Ratio (CDR), disc rim narrowing (DRN) and Retinal Nerve Fiber Layer (RNFL). This was achieved by generating Adversarial Example (AE) that would remove (negative AE) or add (positive AE) pathologic features to explain the model's decision. Gradient-weight class activation mapping was also provided using GradCAM but offered low levels of explainability for normal images. On the other hand, the generated AE provided logic explanations for both pathological and normal images. The method output was evaluated by specialists from a location and rationale explainability perspective, whose reviews showed that the explanations provided were successful for the aspects mentioned above (Glaucoma, CDR and DRN). This work shows the potential of Adversarial Explanations and shows that they can be applied for Glaucoma CAD systems.

Oh et al. [86] proposed a machine learning model for Glaucoma prediction, which also provides explanations for individual predictions. Firstly, 22 clinical features from several examinations were collected from a group of patients. These feature were filtered through the chi-square

feature selection measure and a combination test, which resulted in the 5 final features that were going to used on the prediction model. The authors tested several algorithms for the model, from which XGBoost showed the best performance with regards to AUC, Sensitivity and Specificity. Furthermore, three graphical charts (gauge, radar and SHAP charts) were suggested to explain the model's predictions. These tools provide an insight on the model's prediction and help understand that each feature contributes differently towards a prediction. Authors also claim that since features are not completly independent, they cooperate with each other, creating and interaction that affects the final prediction.

As we could see from Section 3.8, a few systems that have undergone a lot of testing and have been fully launched for commercial use (or are going to do so soon). These systems also address the explainability issue and are adapted for specific uses. These cases prove that it is possible to create such systems and make them accessible to the health care industry. The Eyenuk solution seemed to be one of the most promising in eye disease CAD systems so far, and their three-part CAD(x) system shows how AI can be used to help clinicians in the healthcare industry.

A more detailed description of the scope and future work of this dissertation regarding developing a Glaucoma CAD system with explainable decisions can be found in Section 6.2.

5.6 Summary

This chapter gives an overview of the current state of Explainability and Interpretability in Machine Learning. We can conclude that the XAI field is still very "fresh" and there are many steps that need to be taken to solidify our knowledge fully. Nevertheless, a few literature works are useful for this dissertation's work, mainly Interpretability techniques such as Case-based Reasoning. Their proven success in other applications can be transferred to the Glaucoma context.

Chapter 6

Problem Definition and Proposed Solution

6.1 Problem Definition

From Chapter 2 we know that there is no current efficient strategy for Glaucoma screening and that most Glaucoma patients remain undiagnosed. Moreover, several studies show that Glaucoma has been one of the most prevalent causes of irreversible blindness and visual impairment [92, 121]. This condition was also the second individual cause that mostly contributed to visual impairment in 2020, with 3.6 million known cases. We can expect this number is, in fact, more significant due to the asymptomatic nature of the disease. Varma et al. [124] studied the Glaucoma's economic and individual burdens by reviewing literature published from 1991 to 2010, showing that Glaucoma prevalence contributes to high direct and indirect costs. As the disease progresses, the financial burden increases even more. Glaucoma will also impact patients' health-related quality of life, not only in daily physical tasks as driving, walking and reading but also in their mental health. For these reasons, it is essential to create efficient and useful techniques to aid the Glaucoma diagnosis and screening.

Section 3.7 describes the several limitations and challenges of state-of-the-art CAD systems. This work proposed a solution that tackles mainly two of those applied to the context of Glaucoma. The first and most important one is interpretability, one of the Achilles' heels of "black-box" deep learning models in several fields. The majority of literature focuses on obtaining new models with better performance than the already published ones, which left explanations in the shadows. As explained in Section 5.1, explainability has gained a lot of interest in the ML field, not just because there are new regulatory barriers imposed on ML real-world applications, but mainly because systems' end users do not have "out-of-the-box" trust over them, namely in critical decisions like disease diagnosis. The second limitation we propose to address is the scarcity and imbalance of retinal datasets directed towards Glaucoma diagnosis.

6.2 Proposed Solution

As already stated in Section 1.3, this dissertation aims to create an Explainability Module to improve Glaucoma CAD systems with methods that provide explainable decisions to the system's end-user. One could even say that the target outcome is to create a system that can be compared to a "diagnosis companion" that would provide reasoning for a particular Glaucoma prediction as another clinician would do. Adopting the taxonomy used by the Eyenuk CEO, described in Section 3.8, this system would fit more in the Autonomous AI category, with the benefit of also providing explanations for its decisions.

Moreover, the objective is to provide those explanations using retinal fundus imaging data since it is the most accessible and cost-effective technique for both Glaucoma diagnosis and screening 2.4. Nevertheless, we should not exclude the OCT technique's exploration since it might provide valuable information that can be transferred to the fundus imaging scope.

This dissertation combines two distinct fields. On the one hand, there is Explainable AI (XAI), a vast area with minimal uniform frameworks. For that reason, explainability tasks would be mainly focused on providing explainable decisions to a Glaucoma domain expert, not necessarily on having an intrinsically interpretable classification model.

Chapter 2 describes several morphological features present in fundus images and their relevance on the Glaucoma diagnosis, while Chapter 3 highlights several CAD systems that use those features for the Glaucoma classification task. Although the most explored features are the Optic Disc and Optic Cup, several works refer others such as the PPA, RNFL and Macula as relevant for Glaucoma detection and diagnosis. Section 3.3.1 describes a few of the approaches that explored the PPA and RNFL segmentation. Besides, as stated in Section 5.5, these structures are already being studied to be used as an explainability tool. Due to their importance and lack of exploration in literature, we propose investigating the potential of using these "secondary" morphological features (PPA, RNFL, Notching and Macula) as decision explanations for a Glaucoma expert.

As for Generative Modelling, the objective is to generate synthetic data to improve the quality of explanations. Therefore, we propose to explore image generation with specific morphological structures (PPA, for example) that are less prevalent in the available datasets but can be essential for the above mentioned XAI component. From the collected database information on Section 3.5, only ACHIKO-K and SCROM claimed to have the PPA annotated, but they are not publicly available. To generate such images, literature works like [55, 140] are relevant since they explored the paired and unpaired image-to-image translation problem respectively and proposed methods which showed successful results.

Chapter 4 presents three possibilities for the deep generative model that could be developed. GAN, VAE and Normalizing Flows each have their strengths and weaknesses and could even be used to create a hybrid solution. We propose to use GAN in this work. This technique is the one that usually generates images with higher quality, which is crucial when working with retinal fundus images since they are challenging images to segment and classify due to the nature of

morphological features. Besides, even though GAN have a very high literature prevalence within this topic, few works have applied this technique to retinal fundus image generation, particularly for the Glaucoma classification task [33]. Moreover, a few works developed for privacy-preserving methods in GAN, which is also relevant since fundus images contain the vessel network of the retina, which works like a fingerprint as a biometric authentication technique.

Ophthalmology experts were available during this work to both evaluate and validate the results. Although they could be asked to annotate morphological features of fundus imaging, we expect that there is be more value in obtaining their feedback in either evaluating synthetically generated images, correcting their segmentation, and validating the explainability methods proposed.

6.3 Project Plan

This dissertation work is divided into two semesters. The first semester was focused on background and literature reviews regarding Glaucoma and respective CAD systems and state-of-the-art in Generative Modelling and Explainability fields. This review continued through the first months of the second semester. Furthermore, the second semester's primary focus was the development of the XAI component, as well as its validation on existing datasets. Before implementing this component, the tasks involved data aggregation, pre-processing and augmentation, deep generative learning and morphological feature extraction. Finally, the last month of work was concentrated on writing this dissertation.

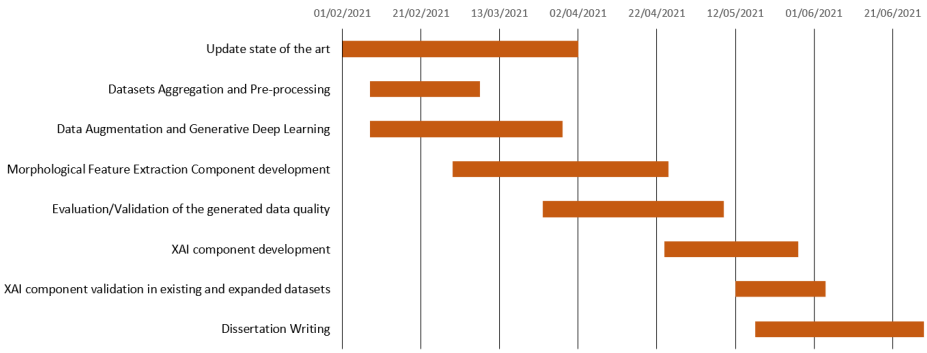


Figure 6.1: Gantt chart for Project Plan.

6.4 Summary

This chapter defines both the problem that gives this work motivation, which is related to the current state of Glaucoma CAD Systems and their lack of explainability. Moreover, the proposed solution is also described, and the project plan for the second part of the dissertation's work.

Chapter 7

Segmentation Approaches for Morphological Feature Extraction

As stated in Chapter 3, it is possible to extract morphological features from retinal images, which could be potentially used as an explainability tool on Glaucoma Risk CAD Systems. Therefore, these features must be correctly obtained from the data since their quality and correctness will influence the validity and quality of the explanations for the clinical context. The first step in this work was to explore several segmentation approaches and evaluate which one should extract the most relevant structures from retinal fundus images.

7.1 Segmentation Datasets

In general, deep learning approaches require a large amount of data to achieve successful results. This aspect is even more relevant when working with retinal fundus images because image quality varies considerably, depending on the device used to obtain them and what conditions they were taken in (for example, lighting, position). For this reason, the majority of literature only evaluate their approaches on a single dataset, which reduces the high variability between images from different datasets. The major drawback of this choice is that the model might not achieve the same performance on other datasets.

In this section, only public datasets were used. Similar to Martins et al.[79], several datasets were merged to obtain a new dataset that better represents real-world retinal fundus images. Despite resulting in a more complex and challenging dataset, the end model could have a better generalization capability.

A dataset was built using the iChallenge-GON, ORIGA, RIGA and RIM-ONE r3 datasets for the OD/OC segmentation task because all possessed OD/OC annotations. When several annotations were given for the same image, for example, the RIGA dataset, the ground truth was calculated as the region of agreement between the annotations. This dataset contained 2517 images:

396 Glaucomatous cases, 1372 healthy cases and 749 unlabeled cases (from the RIGA dataset). The RIM-ONE r3 dataset consisted of stereo images, with two side-by-side retina photographs of the same eye. Each image was split into two and considered a separate case, duplicating the dataset size. This dataset is very similar to the dataset used to train the GFI-ASPP-Depth network from Martins et al. [79], the only difference being the DRISHTI-GS dataset not being included.

7.2 Image Pre-processing

Since the datasets used in this work are collected from different datasets, it is vital to ensure that a few aspects are consistent across all images. Furthermore, there are a few pre-processing techniques that enhance the existing images, which can lead to a better and more robust system.

The first aspect that needs to be taken into consideration is the image aspect ratio. Photographs from different datasets are usually obtained from different devices, which means they have different aspect ratios. In order to normalize this situation, the first step should be to crop the images. Of course, one could also resize the images, but that would modify retinal structures' shape. Consequently, we would not be able to obtain correct values for some of the features described in Section 2.4.2, since they are dependant on widths and areas. As for the aspect ratio chosen, since most state-of-the-art architectures for deep learning use a 1:1 ratio as the input, that same value was used to crop the data. Furthermore, a second crop was performed around the ROI. As stated in Section 3.2, the ROI contains the most relevant information for Glaucoma risk assessment and allows deep learning models to obtain better results in general. This crop is done around the optic disc boundary, annotated in the majority of datasets.

Image quality enhancement techniques became popular in the Computer Vision field to improve the model's performance. Previous works that utilize data from retinal fundus images apply one or more of these techniques, independently of the exact task at hand.

Data Normalization is a common pre-processing technique used across several machine learning approaches, which consists of scaling an image's pixel values to be between 0 and 1.

The Contrast Limited Adaptive Histogram Equalization (CLAHE) is one of those techniques [142], used not only the retinal fundus images but also in other contexts. It is an improvement to a more traditional technique called Histogram Equalization, which improved the contrast of an image by stretching the image histogram to both ends of the spectrum. Although this technique yields good results for when the histogram is restricted to a particular region, the performance decreases when the histogram variability covers a broader part of the spectrum, for example, an image with both very bright and very dark areas. For this reason, adaptive histogram equalization is used by dividing the image into smaller tiles and equalizing them individually. However, applying this technique alone will also increase the noise present in a noisy image. Then, contrast limiting is applied beforehand to clip specific pixels above a certain threshold on the histogram bin. The result of this combination resulted in the CLAHE technique. An example is presented in Figure 7.1.

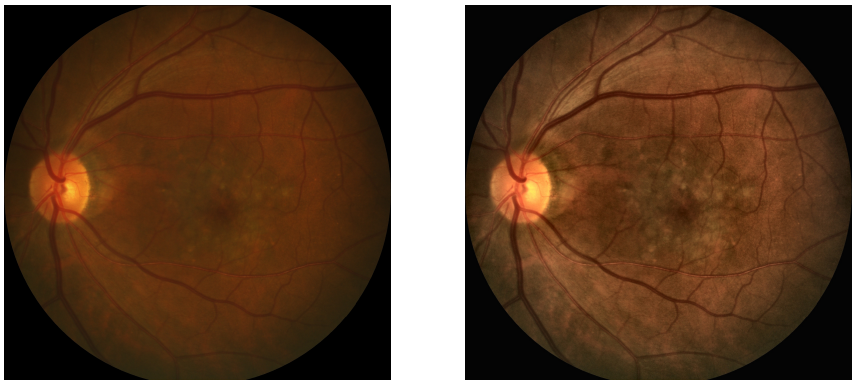


Figure 7.1: Retinal Fundus image before (left) and after (right) CLAHE technique.

Pixel quantification is a recent technique proposed by one of the REFUGE Challenge participants, which applied it in the pre-processing stage of a segmentation task. This technique aimed to reduce the colour variability between the training and validation datasets, thus improving model robustness. For an (RGB) image x where each pixel of x belongs to [0,255], the pixel quantification method can be formulated as follows:

$$x' = ceil(x/r) * r$$

r is a hyper-parameter that controls the quantification impact on the image, and x' is the output image. In general, after applying pixel quantification, pixel values that belong to $[r+1, kr]$ will share the same pixel value of k. An example is presented in Figure 7.2.

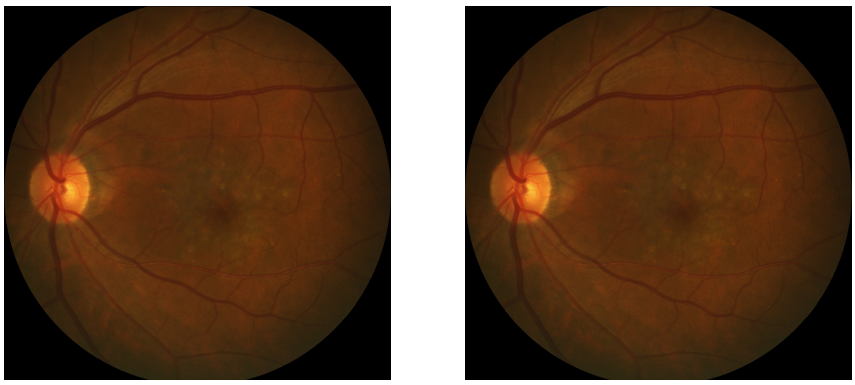


Figure 7.2: Retinal Fundus image before (left) and after (right) Pixel Quantification technique.

Data augmentation techniques can also be handy for increasing the outcome value obtained from an existing dataset. Creating an augmentation pipeline makes it possible to have a controlled creation of new images that are still representative of real data. This is even more beneficial for deep learning approaches due to their dependence on a large amount of diverse and representative data. Applying these techniques usually leads to a more robust model with a better generalization capability and thus is more useful in a real scenario.

An important aspect to consider when augmenting retinal fundus images is that the positioning,

orientation, and width/height scaling of an image are relevant for a clinical evaluation. When geometric transformations such as rotation, scaling or flipping are applied to these images, they can lead to errors in extracting certain morphological features. For example, if we rotated or flipped an image, we would be introducing errors when calculating the width of the ISNT sectors since they would no longer be in the position clinicians expect them to be. For this reason, this kind of augmentation techniques should be used with caution. This augmentation pipeline will be referred to as Traditional Augmentation from now on.

However, another type of augmentation was explored to introduce variability in the image quality without changing the image's properties mentioned above. In a real scenario, photos are taken with different devices that possess different resolutions and light conditions are not the same. By introducing this variability by changing contrast, brightness or other aspects, we are supposedly enhancing the dataset by making it more representative of real data.

A real-time augmentation pipeline was adopted from a previous work with retinal fundus images datasets [79], which does not demand more disk space for storing the augmented images. It is implemented using the imgaug library [1], and it composed of 4 steps that are applied in random order with a certain probability. The steps are the following: blur addition (Gaussian, Average or Median), contrast normalization, brightness changes and sharpness modifications through a sharpening kernel. This augmentation pipeline will be referred to as Image Quality Variation Augmentation from now on. Figure 7.3 show a few of the examples generated by this pipeline.

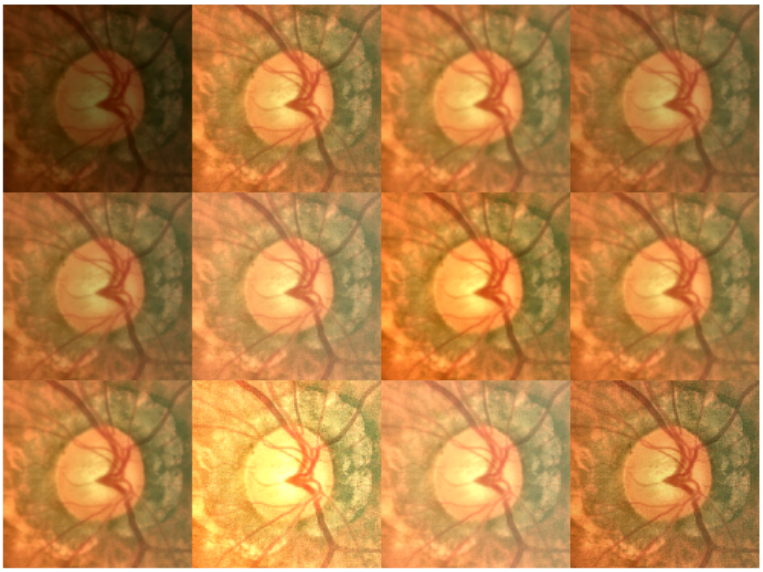


Figure 7.3: Examples of Augmented data using the Image Quality Variation Augmentation.

[1] https://imgaug.readthedocs.io/en/latest/

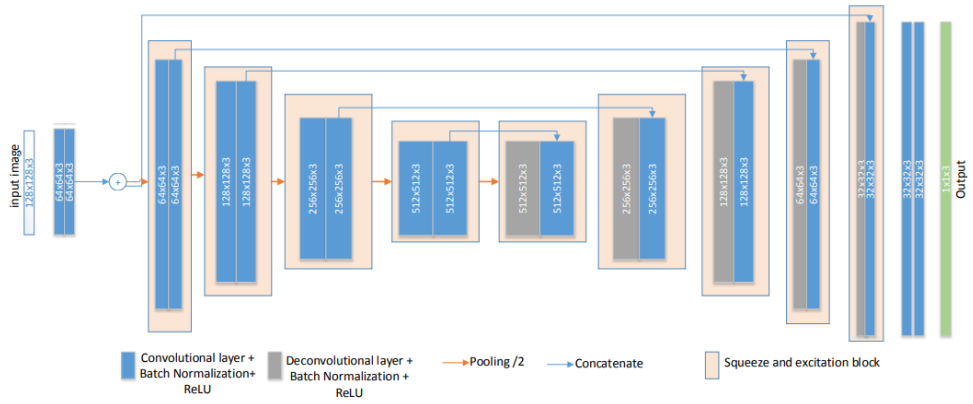


Figure 7.4: X-Unet architecture diagram. Adapted from [74].

7.3 Optic Disc and Optic Cup Segmentation

The Optic Disc and Optic Cup segmentation task can be seen as two separate tasks for each structure or a single joint segmentation task. The majority of state-of-the-art approaches opt for joint segmentation. It is reasonable to assume that since both structures are closely related to one another (Optic Disc contains in the Optic Cup), it makes sense to have the model train on both segmentations simultaneously. Moreover, it also simplifies the entire task since there is only a need to train and fine-tune a single model. For these reasons, this work explores joint segmentation.

Based on Section 3.3, two architectures were compared in this work. One of them was the GFI-ASPP-Depth network proposed by [79]. Not only are the results reported on pair with the state of the art performance, but the pre-trained model for the Joint segmentation task was also available. The architecture is based on the MobileNet architecture, which reduces time and space complexity compared to other state-of-the-art networks without compromising predictions performance.

Secondly, this work also explores the X-Unet network proposed by [74] for the REFUGE Challenge. Figure 7.4 shows this network's architecture. While the GFI-ASPP-Depth network only requires 1.152.131 parameters, the X-Unet architecture requires 13.889.506 parameters, which makes the latter more complex and computationally heavy. On the paper, the authors only reported results for the individual segmentation task of the Optic Disc and Optic Cup. For that reason, the network's performance was tested for the joint segmentation task in this work to be compared with other networks. The implementation of this network was based on the source code[2] provided by the authors. The network has a U-Net [97] like architecture, with squeeze-and-excitation blocks that recalibrate channel-wise features responses to improve the model's performance at a low computational cost. Although more complex and computationaly heavy than the previously mentioned GFI-ASPP-Depth network, its performance with retinal fundus images is promising.

Since most resources were implemented using Tensorflow, this framework was used for this part of the work.

[2] https://github.com/cswin/RLPA

7.3.1 Training

Regarding the GFI-ASPP-Depth network, the pre-trained Tensorflow Lite model and source code were available, so there was no need to train the network from scratch.

As for the X-Unet network, there is access to the source code but not to any pre-trained models. The first step was to replicate the setup described in the paper as much as possible to evaluate the network's performance. The initial dataset used was only composed of the iChallenge-GON images from the Train and Validation sets. Train, validation and test subsets were created according to a 70/15/15 split proportion from the 800 images available. The full fundus images were also cropped into the region of interest (ROI) patches around the Optic Disc, using the network (DEnet) provided by the authors for that purpose. A pre-trained model of this network is available on the source code repository.

While the authors create and store the augmented data before training, a real-time augmentation pipeline was adopted due to hardware memory limitations. Since this is a segmentation task, it is possible to use geometric transformations as augmentation techniques. They do not modify the ratio between retinal structures in very abrupt ways (for example, scaling only one dimension of an image by a significant amount). The images were resized to 128x128 pixels before being used as input on the network. The network outputs two values per pixel resulting from the final sigmoid activation layer, each representing either the Optic Disc or the Optic Cup probabilities. The ADAM optimizer was used across all experiments due to the good results widely presented across the literature. The paper authors used the Mean Absolute Error (MAE) as the loss function because they were focused on calculating the pixel-wise difference between label and prediction of a single structure (either optic cup or disc). However, the joint segmentation is a multi-label segmentation, so the cross-entropy was used as the loss function, with equal weights for both classes (Optic Disc and Optic Cup). Besides the loss, two other metrics were used to track the training session: Intersection over Union (IoU) and the Dice Coefficient.

Nested hyperparameter optimization was adopted, allowing for tuning parameters one at a time and finding the optimal value for each one individually. This strategy is more efficient than a "guess-based" strategy since it is a more systematic method while not being as time-consuming as a grid search strategy. Moreover, callbacks were used to automate certain procedures during training. Every training session had a checkpoint callback to store the last best model according to an evaluation metric, a learning rate reduction callback and an early stopping callback to halt training when the validation loss was not improving for several epochs. The training session was configured to run a maximum of 200 epochs, but almost every single one stopped training before reaching that number due to the callbacks.

As stated previously, the first experiments were performed with the iChallenge-GON dataset, without any augmentation and resized to a $128 * 128$ dimension. The learning rate was the first tuned parameter, using a starting value of 0.0001 since it is used on the original paper. Other values were explored by reducing or increasing the initial value by a factor of 10. Still, the optimal value remained 0.0001 since it obtained the best performance on the test set with a dice coefficient

of 0.8717. As for the batch size, 16 was chosen from an interval of values that varied by a factor of 2 since it presented the best performance relative to lower values. It was not possible to use higher values due to hardware limitations. The model had a good performance on the test set at this stage, with an IoU and Dice Coefficient of 0.7666 and 0.8649, respectively. Consequently, the learning rate used from now on is 0.0001, and the batch size is 16.

The second part of the experiments consisted of testing the impact of two preprocessing techniques that have been used in previous literature to enhance the original data. When applying CLAHE, the performance greatly improved, leading to an increase in IoU to 0.9404 and in Dice Coefficient to 0.9678. When using Pixel Quantification ($r = 5$), performance resulted in an IoU of 0.8909 and a Dice Coefficient of 0.9380. As a result of these experiments, the CLAHE technique was chosen as a preprocessing technique for future experiments.

Another training experiment verified that the original architecture's input, which was tripled when fed to the network, did not bring value compared to having a single input. On the test set, the triple input strategy decreased the IoU value by 0.05 and the Dice Coefficient by 0.02. For that reason, the network was simplified, and a single input was used.

At last, the complete segmentation dataset described in Section 7.1 was used to retrain the model. Images were also cropped to the same ROI image as performed by Martins et al. [79], and a CLAHE preprocessing technique was applied. The data was split across three subsets, train, validation and test, with an 80/10/10 split proportion. By minimizing the difference between the data used to train different networks, we can compare their performance more fairly.

Dataset Augmentation	**Loss**	**Iou Disc**	**IoU Cup**	**Dice Disc**	**Dice Cup**
No Augmentation	0.2002	0.9015	0.7129	0.9473	0.8230
Image Quality Variation	0.2014	0.9041	0.7318	0.9486	0.8354
Image Quality Variation + Traditional	0.2021	0.8762	0.6586	0.9330	0.7812

Table 7.1: Performance of X-Unet models, trained on datasets with different augmentation techniques.

The first model was trained using the previously mentioned dataset without any augmentation techniques. The model converged after 42 epochs, reaching an IoU of 0.8603 and a Dice Coefficient of 0.9235 on the test set. A second model was trained using the same dataset, augmented using the Image Quality Variation Augmentation pipeline. IoU had a slight increase of 0.008, and the Dice Coefficient improved by 0.004. Another model was trained with data augmented with the previous pipeline and also traditional augmentation techniques. The parameters used were a rotation up to 90º, horizontal and vertical flipping and width and height shift of 0.02. In this case, the model performance decreased on all sets by a considerable amount (test set IoU decreased by 0.04 and Dice Coefficient decreased by 0.02).

Table 7.1 provides the segmentation performance for each of the segmented structures, allowing for a more detailed analysis. Image Quality Variation Augmentation seems to enhance the dataset better, leading to better performance. As expected, the optic disc segmentation is more

accurate than the optic cup, which is an issue that can also be observed on manual clinical annotations. Since the intensity difference between the optic cup and optic disc is often shallow, it is hard to accurately draw the boundary between both structures, even for clinicians. For this reason, annotations produced by different experts can have significant differences in the optic cup boundary.

The results were compared with other state-of-the-art approaches, which include the GFI-ASPP-Depth network proposed by Martins et al. [79]. The IoU was used as the comparison metric since it was the most available one across the literature. Nevertheless, one must consider that a direct comparison between models does not lead to absolute truths. Most models are not trained with the same dataset, which could also be subject to different pre-processing techniques, leading to an unfair comparison in some cases. Table 7.2 was adapted from Fu et al. [42] and contains a performance comparison between the state-of-the-art models. As we can observe, the X-Unet network performance is on par with other successful approaches, outperforming most of them and almost reaching the performance of the top two methods: M-net and GFI-ASPP-Depth.

Method	Iou Disc	IoU Cup
Adapted from [42]		
R-Bend[57]	0.8710	0.6050
ASM[130]	0.8520	0.6870
Superpixel[26]	0.8980	0.7360
LRR[129]	-	0.7560
QDSVM[27]	0.8900	-
U-net[97]	0.8850	0.7130
M-net[42]	0.9170	0.7440
Achieved results		
GFI-ASPP-Depth[79]	**0.9100**	**0.8260**
X-Unet	0.9040	0.7310

Table 7.2: Performance comparison between X-Unet and state of the art methods. Segmentation performance comparison with state-of-the-art methods trained with the ORIGA dataset.

For a more in-depth comparison, Figure 7.5 shows the loss evolution over the epochs for both the X-Unet and the GFI-ASPP-Depth networks. In both cases, loss converges rapidly in the beginning before stabilizing, and early stopping is performed.

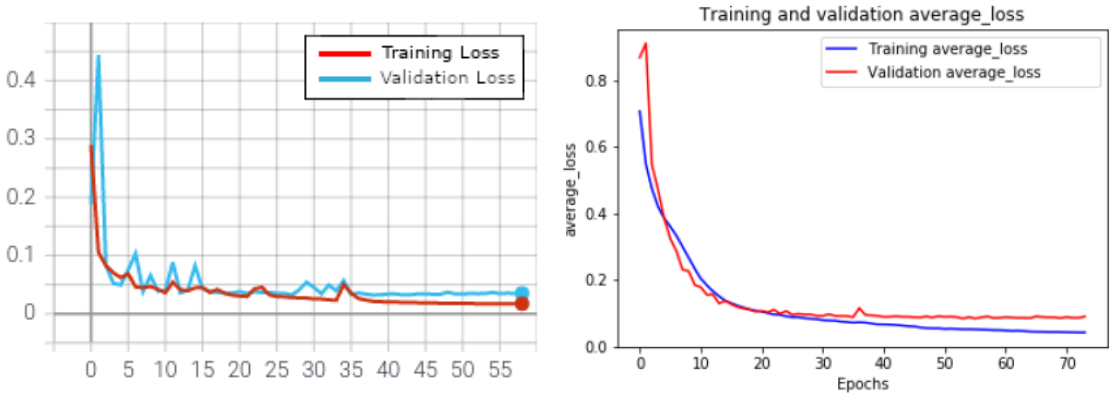


Figure 7.5: X-Unet (left) and GFI-ASPP-Depth[79] (right) training losses.

7.4 Parapapillary Atrophy (PPA) Segmentation

As Section 3.3 shows, the majority of literature work on Glaucoma Risk Detection and Retinal Fundus Imaging Segmentation is deeply focused on the ROI region, more specifically on the Optic Cup and Disc, and its respective characteristics. Although not very present, there are a few approaches focused on the PPA segmentation, which are further described on Section 3.3.1.

Two strategies were used to obtain the PPA mask. On the one hand, a network was trained using the masks containing only the PPA segmentation. On the other hand, the PPA and Disc masks were merged to obtain a single mask, and the model was trained to predict the PPA-Disc region. The PPA has an irregular and non-uniform shape that can vary significantly depending on

the lesion progression, as shown in Figure 7.6. While **(a)** and **(b)** have a circular shape, **(c)** has a semicircular shape and **(d)** an almost crescent one. Moreover, the PPA does not always have an obvious boundary since pixels around the PPA gradually change colour, as we can observe in Figure 7.7. These factors make the PPA extraction task very challenging. Chai et al. [17] proposed a new strategy to overcome this issue by transforming the complex segmentation object to be a new object with a more uniform shape, which eases the overall complexity of the task. Figure 7.8 shows two retinal images where the green area represents the PPA region, and the purple area represents the Disc region. By calculating the Union between both regions, the PPA-Disc area obtained has an almost oval shape and is easier to extract than the PPA area. Afterwards, one only needs to subtract the Disc area from the PPA-Disc area to obtain the PPA area. The Disc area is a well-known and explored task across literature, with several approaches obtaining successful results. Consequently, we can assume that for most cases, the disc segmentation error will not affect the final PPA mask greatly.

Figure 7.6: Retinal images and their PPA and Disc areas [17].

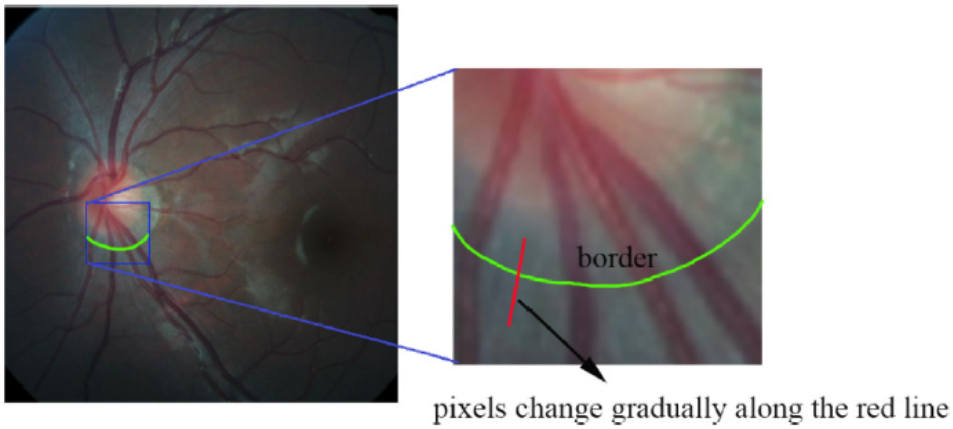


Figure 7.7: Illustration of PPA area border [17].

The data used for this task is from a single subset of the only publicly available dataset: the training set from the iChallenge-PM dataset. Moreover, not all retinal images showed the PPA lesion, and there is another lesion called "Detachment", which is also present on myopic eyes, which

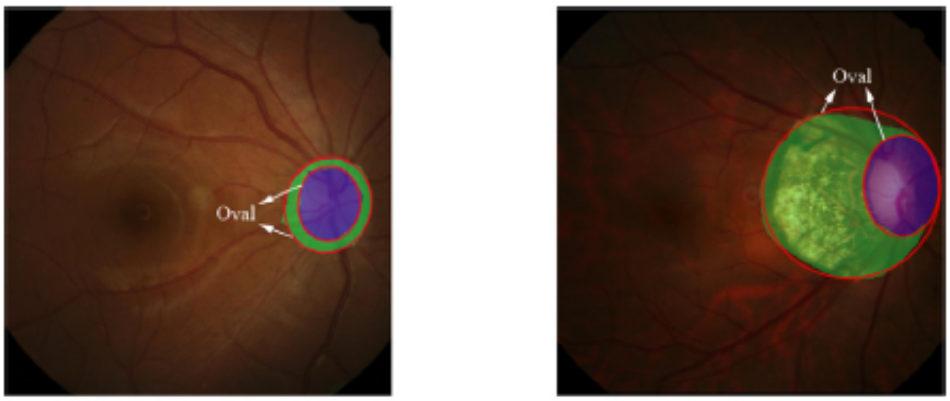


Figure 7.8: Retinal images and their PPA and Disc areas [17].

is not a symptom of Glaucoma. Depending on the myopia degree of each image (normal, high or pathological), the PPA had a different appearance. In Figure 7.9, the PPA lesion is much larger and irregular on the pathological labelled image than on both high myopia and normal images. Regarding literature work on Glaucoma and the *Terminology and Guidelines for Glaucoma*[1] published by the European Glaucoma Society [114], the PPA typical of Glaucoma always appears much similar to the one shown on high myopia and normal cases, then to the one shown on pathological images. Furthermore, it is much more beneficial to use just the retinal images' ROI for the segmentation task in the Glaucoma context. If pathological images were used, they could not be cropped to the ROI because the PPA can appear outside of that area. For these reasons, and in order to simplify the learning process during training, the pathological myopic images were not used to train and test the model. Instead, empty masks were added to the dataset as masks for the images where PPA does not happen. This addition will help the model learn when an image has PPA or not since not all Glaucomatous retinal images have PPA presence necessarily. Similarly to the Joint Segmentation task, since the PPA is located near the Optic Disc, an ROI crop was performed on the retinal images. The final dataset used during training contained 145 images, split into three subsets, training, validation and test, with an 80/10/10 proportion.

The same X-Unet architecture used for joint segmentation was also tested on the PPA segmentation task, and the same fine-tune strategy was also adopted. While joint segmentation meant that a single pixel could have multiple labels, only a single label is needed in the PPA segmentation. For that reason, the network's last activation layer was a sigmoid activation. No significant results were achieved after performing several training sessions with binary cross-entropy loss and varying the learning rate and batch size. Both IoU and Dice Coefficient values for PPA only were never higher than 0.15 and 0.2, respectively. Figure 7.10 presents an example of the segmentation, with a network trained with a learning rate of 0.0001 and a batch size of 8, which shows how inaccurate the segmentation is. From these results, three possible issues were identified. The first one is regarding the dataset. Contrary to the previous task, the PPA dataset is very small. Thus it might not contain enough information for the model to learn anything significant. The second

(a) Normal Case **(b)** High Myopia Case **(c)** Pathological Myopia Case

Figure 7.9: iChallenge-PM images

problem could be related to the PPA and background ratio of each image. The loss used does not consider how little area the PPA lesion occupies in the retinal image, giving equal importance to all "classes" (PPA and background). Finally, the PPA shape is more complex and more variable than the oval shape of the ONH, which means it demands more effort from the network's learning process.

Figure 7.10: Left: PPA ground truth, Right: Network PPA segmentation.

The first step was to change the loss function to a function that could give more weight to the PPA lesion and not so much to the background. The Focal loss [72] function addresses the class imbalance problem during training. This function is a dynamically scaled cross-entropy loss, which reduces the weight of easy examples during training and focuses on predicting the hard examples. Formally, this function adds factor $(1-p_t)\gamma$ to the standard cross-entropy formula, where p_t represents the probability for a certain class, and γ is a tunable focusing parameter. Equation 7.1 defines focal loss. In this case, the easy examples would be the background pixels, while the hard ones would be the PPA pixels.

$$FL(p_t) = -(p_t)^{\gamma} \log p_t \tag{7.1}$$

A custom focal loss was implemented and used to retrain the model. Several values for γ were

tested. The Image Quality Variation Augmentation and the Traditional Augmentation were used. Contrary to what was expected, results obtained were even worse than with cross-entropy loss, with both IoU and Dice Coefficient dropping below 0.1.

Despite the unsuccessful PPA segmentation results, it is still possible to reach a few conclusions and infer possible causes. Firstly, even if the PPA is visually similar for clinicians on both myopia and Glaucoma, it might be wrong to assume that a model trained on one domain could have similar performance on the other. Moreover, models evaluate the entire retinal image, not just the segmentation goal, which means these different domains demand different strategies and techniques due to the retina having different characteristics (colour, textures, artefacts). In order to investigate whether these differences were plausible, the optic disc was segmented from the myopia dataset. Although the model was not trained for this domain, the results could indicate if the ROI has similar characteristics to the one from Glaucoma images and if the model can segment the optic disc accurately. The optic disc segmentation network proposed by Martins et al. (GFI-ASPP-Depth-simple), trained with Glaucoma domain image was accurate, obtaining an IoU of 0.8763 and the Dice Coefficient was 0.9329 on the entire dataset.This outcome supports shows that despite having different diseases, the ROI region of retional images is still similar and does not affect the already accurate segmentation of the OD.

The dataset size is another aspect to take into consideration. Since there is a relatively low amount of images, there might not be enough data for the model to learn how to distinguish the PPA from the rest of the image. As previously stated in this section, PPA segmentation is a challenging task. In addition, certain aspects such as the lesion's unclear boundaries or proximity to other structures separate it from a simpler segmentation task.

For this reason, one of the ways clinicians could contribute to this work would be to annotate the PPA structure in one of the public datasets considered in this work. Since there was this possibility, a dataset was built combining two types of images. One the one hand, a smaller portion of images consisted in images with already available PPA annotations. On the other hand, images from datasets with Glaucoma labels were also included in a larger portion, since there is a high probability that PPA also appears on images with a positive Glaucoma label.

7.5 Fundus Image Feature Extraction

The previously described models generate segmentation masks that can be used to extract several morphological features. These features are listed in Section 2.4.2, and are not only relevant for the Glaucoma risk detection but can also be helpful as an explainability tool that increases the transparency of the system. This will be further described in Chapter 8.

Before calculating all the morphological features, the following values are calculated from the segmentation mask: the area and vertical diameter of the optic cup and optic disc; the Neuro-Retinal rim (NRR) widths in each of the four quadrants of a fundus image (Inferior, Superior, Nasal and Temporal). From these values, eight morphological features are calculated. Firstly, the CDR is calculated using the area ratio between the optic cup and optic disc. Then, the VCDR

follows the same principle but utilizes the vertical diameter of each structure. Then the Rim-to-Disc Area Ratio is obtained using the NRR widths and the disc area. The ratio between each NRR width and the longest NRR width is also stored as a morphological feature. Finally, these values are compared to each other to evaluate the ISNT rule compliance.

For this task, a feature extraction pipeline was adapted from [79], which can receive two kinds of input. When given a full fundus image, it starts by cropping the image to a 1:1 aspect ratio and then uses a pre-trained model called GFI-ASPP-Depth-simple (also proposed by [79]) to locate the disc region. The ROI region is cropped from the original image after localizing the disc region, and CLAHE preprocessing is applied. Next, the processed image is used as input for a joint segmentation model to obtain the optic disc and optic cup segmentation mask. From this mask, all of the above features are extracted. If the input is an image already cropped to the ROI region, then the first part of the pipeline can be skipped, and the image is resized and used on the joint segmentation model. A diagram of this pipeline is presented in Figure 7.11.

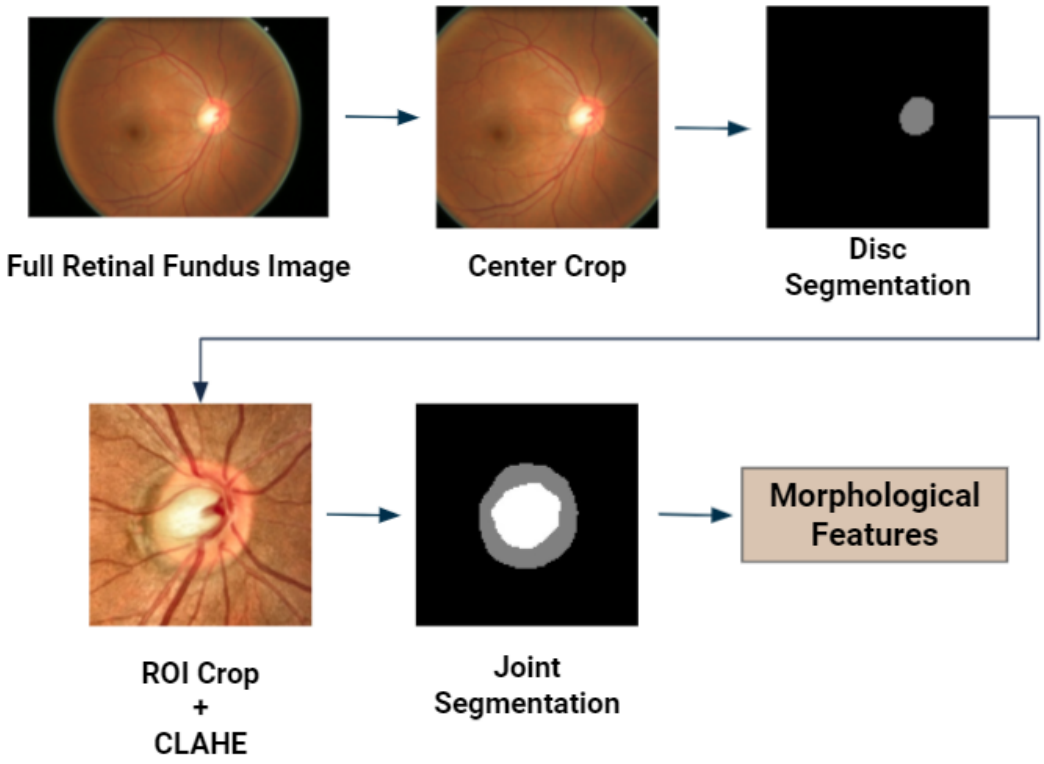
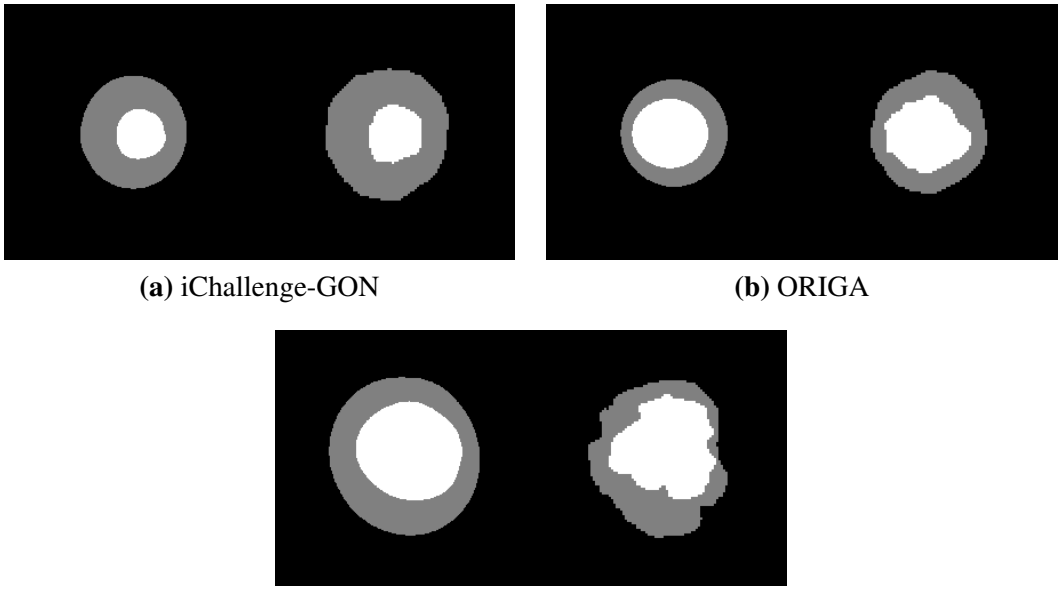


Figure 7.11: Morphological Feature Extraction pipeline.

The joint segmentation model used in the feature extraction pipeline must be as robust and accurate as possible to minimise the error on the morphological feature calculus, which is done from the segmentation masks. For this reason, the GFI-ASPP-Depth was used on the pipeline, which is the model with the best performance. To evaluate if it was possible to segment other datasets than those used in the network's training, each one was segmented, and the output was

visually analysed. Figure 7.12 shows segmentation examples from three datasets where the ground truths masks are available.



(a) iChallenge-GON **(b)** ORIGA

(c) RIM-ONE r3

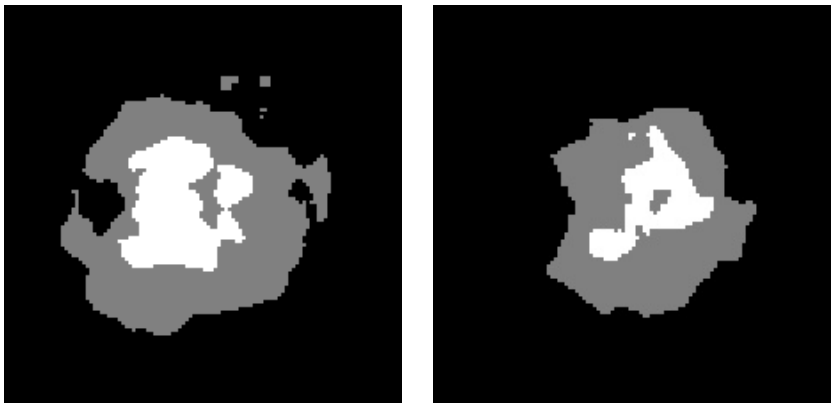
Figure 7.12: GFI-ASPP-Depth Segmentation examples with ground truth masks comparison (left side of image is ground truth and right side is the predicted mask).

The output masks for the iChallenge and ORIGA datasets are very accurate by visually comparing the output with the ground truth masks and evaluating the model's performance through metrics (see Table 7.3). However, the performance with the RIM-ONE r3 dataset decreases considerably. Despite being able to segment both structures, the mask compromises the morphological feature calculus.

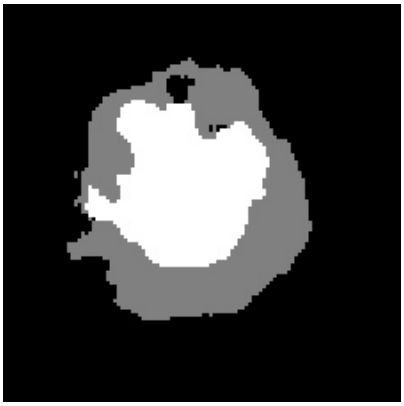
Dataset	**Disc IoU**	**Cup IoU**
iChallenge-GON	0.7703	0.7350
ORIGA	0.7952	0.7579
RIM-ONE r3	0.7831	0.4969

Table 7.3: GFI-ASPP-Depth Segmentation performance on iChallenge-GON, ORIGA and RIM-ONE r3 datasets.

Figure 7.13 shows segmentation examples for the remaining datasets. In the ACRIMA, RIM-ONE r1 and RIM-ONE r2, it is obvious that the segmentation is inaccurate due to the irregular and unnatural shape of the optic cup (white pixels) and optic disc (grey pixels). A possible explanation for this is the different zoom between the images. The ROI is much more zoomed on these three datasets, occupying more area than on the iChallenge-GON and ORIGA datasets. The network might only work on images where the ROI crop is similar to the one used on the training dataset.



(a) ACRIMA **(b)** RIM-ONE r1



(c) RIM-ONE r2

Figure 7.13: Segmentation examples with ground truth masks comparison (left is ground truth and right is predicted mask).

7.6 Summary

This chapter illustrated the work developed on segmentation of important retinal structures, which is crucial for developing and implementing Glaucoma CAD systems. On the one hand, the new Optic Disc and Optic Cup segmentation architecture achieved state-of-the-art results, proving its potential with retinal fundus image data. On the other hand, PPA segmentation was not as successful and still needs more research to create a robust segmentation model. Lastly, morphological feature extraction was studied and is a reliable way to obtain more information from retinal fundus images, which can be used in other tasks involving retinal data analysis.

Chapter 8

Model Explainability

As already stated in Chapter 5, there is a wide variety of explainability techniques in ML, some with more nuances than others, that can be divided into several categories depending on the chosen criteria.

This work aimed at improving Glaucoma Risk Detection CAD systems by enhancing their classification outcome with meaningful explanations for clinicians. In Chapter 7, the morphological features extraction pipeline was described. The clinical relevance of these features makes them suitable to be used as an explainability tool. Contrary to what is stated on the initial proposed solution (Section 6.2), the most valuable explanations we could provide to a clinician would be the ones that describe the what features the model considered important in a particular classification. For this reason, one of the techniques explored was Concept Whitening, described in Section 8.2. Nevertheless, several Post-hoc explainability techniques can provide a decent insight into the data. SHAP is a very recent and popular technique used to explain the output of any machine learning model, which was also explored in this work.

8.1 Datasets

Before applying any explainability technique, it was necessary to define what model would be used for the classification task. The only requirement for this was that the dataset much have Glaucoma labels. A pre-split dataset was used for this part, which was built using the ACRIMA, iChallenge-GON and ORIGA datasets (80/10/10 split proportion). The images were obtained by processing the datasets on the morphological feature extraction pipeline (see Section 7.5), which cropped the images to the ROI and applied the CLAHE technique.

In order to utilize fundus images morphological features as a basis for any explainability component, these features must be as close to ground truth as possible. Since the feature extraction occurs on the segmentation masks, the more accurate the segmentation model is, the more precise features will be. According to Section 7.5, the datasets where segmentation is the most accurate

are iChallenge-GON and ORIGA, making them the best candidates for the explainability task. Furthermore, this task is part of a classification task, meaning the data also must have Glaucoma labels. That is the case with both of these datasets.

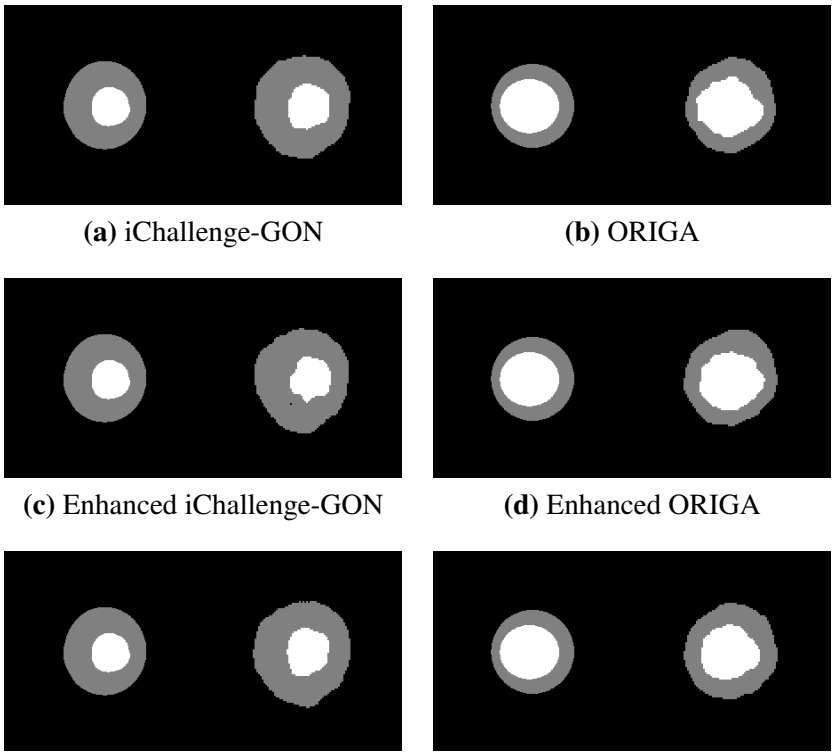
8.1.1 Generative Modelling on Retinal Fundus Imaging

One of the objectives of this work was to take advantage of Generative Modelling to enhance the quality of explanations and increase the amount of data to be used in this component. A lot of research was done about this field (see Chapter 4), and Generative Adversarial Networks (GAN) were one of the approaches used in the context of retinal imaging. In parallel with this dissertation, Fraunhofer is developing a Generative Model approach focused on evaluating the impact of generative modelling on Glaucoma CAD systems. Leonardo et al. [68] proposed a model based on CycleGAN architecture that transforms retinal images by improving/degrading their quality to augment the original data. This transformation process is part of the Unpaired Image-to-Image translation problem, which involves generating an image on domain Y from an input image on domain X. In this case, this translation occurs between domains with different levels of image quality, evaluated through the presence or absence of defects such as blurring, over/under-exposure, etc. The generated images were validated using a Retina Quality Evaluator also proposed by the authors, which showed there are tangible improvements in image quality using the proposed generative model. Furthermore, the new images were used to train a Glaucoma CAD system that presented a considerable gain in Sensitivity, Specificity and Accuracy after image data augmentation compared with state-of-the-art methods targeted at offline inference in mobile devices. These results support the statement that image quality diversity and realistic augmentation are crucial aspects that can increase the model performance on other tasks.

One of the tasks described in the proposed solution in Chapter 6 is to use a Generative Modeling approach (namely GAN) to augment the available data, with the ultimate goal of improving the value of explanations. Moreover, it is also stated that image quality is a crucial aspect when working with retinal imaging due to the level of detail of retinal structures used to calculate certain morphological features. Despite GAN experiments conducted on a preliminary stage, it was impossible to fully explore and develop a satisfiable data augmentation approach using a generative model due to task prioritization and time constraints. However, the generated data from the previously described work was available to use, even though the work had not been published yet. By taking advantage of this opportunity, it was possible to create a more extensive and robust classification dataset and evaluate the benefits this type of data could bring to this work's explainability segment. The generated data available from these experiments resulted in a dataset with an improved and a degraded version of each input image, which essentially triples the size of the input dataset. The input dataset consisted of the iChallenge-GON, ORIGA, ACRIMA, DRISHTI and RIM-ONE datasets.

Initially, the final classification dataset would consist of iChallenge-GON and ORIGA only since they are the models where the optic disc and cup are the most accurate. By adding the improved and degraded versions of each of these images from the GAN generated dataset, the

size would go from 1450 to 4350 images. The generated images segmentation was evaluated for each improved and degraded version of each dataset on the GFI-ASPP-Depth segmentation model compared with the regular version of the images. Figure 8.1 shows an example of each version of images from the iChallenge and ORIGA datasets.



(a) iChallenge-GON **(b)** ORIGA

(c) Enhanced iChallenge-GON **(d)** Enhanced ORIGA

(e) Degraded iChallenge-GON **(f)** Degraded ORIGA

Figure 8.1: GFI-ASPP-Depth Segmentation examples with ground truth masks comparison (left side of image is ground truth and right side is the predicted mask).

Dataset	**Disc IoU**	**Cup IoU**
Regular iChallenge-GON	0.7703	0.7350
Enhanced iChallenge-GON	0.7738	0.7265
Degraded iChallenge-GON	0.7867	0.7241
Regular Origa	0.7952	0.7579
Enhanced Origa	0.7911	0.7194
Degraded Origa	0.7533	0.7184

Table 8.1: GFI-ASPP-Depth Segmentation performance on iChallenge-GON and ORIGA datasets.

As it is possible to observe in Table 8.1, there is not a significant decrease in segmentation performance when using the enhanced or degraded images, which means that they can be used on the Explainability module.

Table 8.2 sumarizes the datasets used in this section.

Dataset Name	Datasets Used	Glaucoma	Non Glaucoma	Total
Pre-Split	ACRIMA, iChallenge-GON and ORIGA	644	1511	2155
Enhaced/Degraded	iChallenge-GON and ORIGA	744	3606	4350

Table 8.2: Overview of classification datasets.

8.2 Concept Whitening

Concept Whitening is a mechanism introduced by Chen et al. [23] that can alter a network's layers to allow us to understand better the computation leading to that layer. This mechanism shapes the latent space through training, imposing its axes to be aligned along with certain concepts. In this case, our concepts are the morphological features extracted from retinal fundus imaging, which are clinically relevant for Glaucoma risk detection.

After applying concept whitening a network's layer, target concepts can be extracted in several ways. In order to obtain the concepts relevant for an individual classification, the authors employ empirical receptive fields, which highlight the regions of the image relevant for target concepts. Figure 8.2 show several examples in a grid, where each row represents the most activated image for a specific concept axis, as well as the respective receptive field for each of the concepts learned by the network. As a general rule, these fields tend to be more prominent on image regions relevant to recognizing the correct concept.

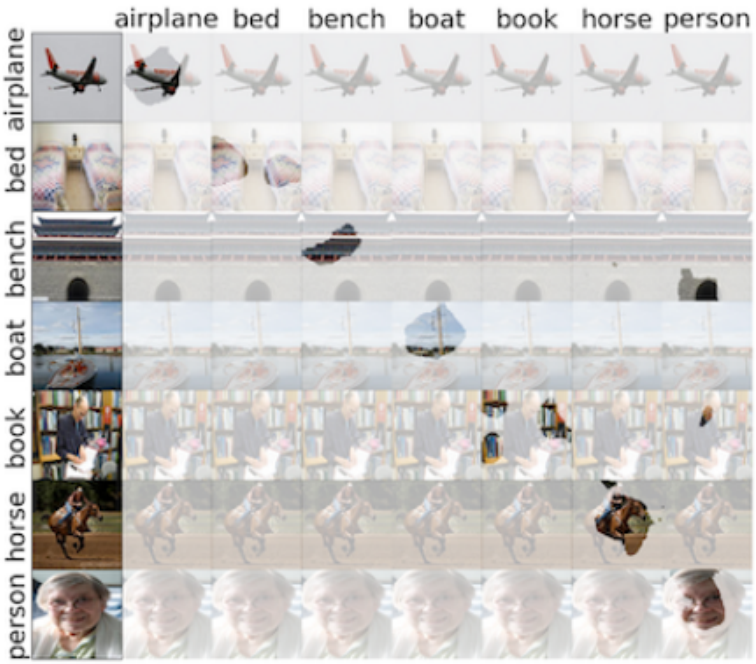


Figure 8.2: Some top activated images visualized with empirical receptive fields (highlighted regions). Adapted from [23].

Concept Whitening source code is publicly available on a GitHub repository[1], and contains a Pytorch implementation of the mechanism. For that reason, this part of the work was developed in Pytorch.

Similarly to the original paper, a classification network was used to evaluate the Concept Whitening mechanism's efficiency on the Glaucoma Risk Detection task. The ResNet architecture was used for the classification network, particularly ResNet50, due to achieving good performance in literature[30] when compared to other architectures. Nevertheless, a ResNet18 network was also trained since it is based on the same building blocks but is a simpler network that converges faster while still obtaining satisfactory results.

Regarding the data, explicit concepts that the network should learn must be explicitly provided to it. These concepts are inferred from the morphological features previously extracted. Datasets images were split into an auxiliary concept dataset, using criteria based on the risk factors described in Chapter 2. Each image was evaluated according to these criteria and copied to a particular concept folder. The criteria used are the following:

- **Cup-to-Disc** ratio is greater than **0.5**;
- **Vertical Cup-to-Disc** ratio is greater than **0.7**;
- **Rim-to-Disc Area ratio** is less than **0.5**;
- **ISNT rule** is **True**;

Initially, there was also the intention to add the "PPA presence" as another concept. However, since it was not possible to train a model with good PPA segmentation, it was decided not to include this concept. Table 8.3 shows the size of each auxiliary concept dataset.

Concept Dataset	Glaucoma	Non-Glaucoma	Total
CDR	185	32	207
VCDR	251	83	334
RDAR	196	35	231
ISNT	30	154	184

Table 8.3: Image count on each auxiliar concept dataset.

8.2.1 Training

Training is divided into two stages. The first one consists of training a baseline model, without the concept whitening layers, on the classification task only. The second one retrains the network in a transfer learning fashion, only replacing the Batch Normalization layers with Concept Whitening ones where necessary.

The first stage started with evaluating which network showed the best performance. A Resnet50 and a ResNet18 were trained with the *Pre-Split* dataset. Images were resized to 244x244 before

[1] https://github.com/zhiCHEN96/ConceptWhitening

being input in the networks, and no augmentation technique was applied. Nested hyperparameter tuning was used to find the optimal values for the parameters. The ADAM optimizer and Cross-Entropy loss were used across all experiments. Class weights were used to minimize the impact of the class imbalance on the loss since there are many more cases of Glaucoma than non-Glaucoma. The number of epochs was set to 50, with an early stopping callback is the validation loss hadn't improved for a certain amount of epochs. Networks were also initialized with ImageNet weights. In each training session, accuracy, precision, recall, area under the curve and F1 score were tracked.

Table 8.4 reports the results on the test set for the most relevant experiments. Since we are in the context of a disease diagnosis, it is vital to minimize the number of Glaucomatous cases classified as non-Glaucomatous. Recall or sensitivity is the metric that tells us the proportion of Glaucomatous cases classified as such. At the same time, the model should have good overall accuracy. Despite the ResNet50 with a learning rate of 0.01 and a batch size of 4 having the best recall value, it also presents the second-worst AUC value. ResNet18 models show a lower Recall value by 0.01 approximately but have a higher AUC value. Since we are looking for a balance between these two metrics, the ResNet18 model is more suitable. The two highlighted models are very similar, so opting for either one should not significantly impact the following experiments. Therefore, the second model, with a learning rate of 0.001 and batch size of 8 was chosen.

Model	**LR**	**BS**	**Loss**	**Accuracy**	**Precision**	**Recall**	**AUC**	**F1 Score**
ResNet18	0.01	4	0.0424	0.8711	0.8632	0.9820	0.8711	0.9188
ResNet18	0.001	8	0.0429	0.8711	0.8632	0.9820	0.8751	0.9188
ResNet50	0.01	4	0.0536	0.8667	0.8513	0.9940	0.7932	0.9171
ResNet50	0.001	8	0.0532	0.8489	0.8482	0.9701	0.8188	0.9050
ResNet50	0.001	16	0.0636	0.8578	0.8524	0.9760	0.7870	0.9106

Table 8.4: Results for ResNet18 and Resnet50 experiments on the test set with *Pre-Split* dataset. **LR** represents Learning Rate and **BS** represents Batch Size.

At a later part of this stage, the network was trained using the *Enhanced/Degraded* dataset, augmented with the Image Quality Variation pipeline from the segmentation task. Since the dataset is much larger than the previous one, the model parameters were tuned once again. The optimal learning rate found was 0.0001 and a batch size of 32. Table 8.5 presents the most relevant results on the test set and also compares the model trained with and without the enhanced and degraded versions of the original images (only with the iChallenge-GON and ORIGA datasets). The model trained with the entire dataset presented a better performance than the other one, proving that the enhanced/degraded versions of the images are helpful for the Glaucoma classification task.

After settling with a robust classification model, the work moved on to the second stage to evaluate the concept whitening mechanism. Firstly, the model weights are initialised with the pre-trained weights from previous training sessions. Then, the architecture is modified by replacing the Batch Normalisation layers with Concept Whitening ones, which have their implementation in the source code provided by the authors. Not all layers need to be replaced, and each one will

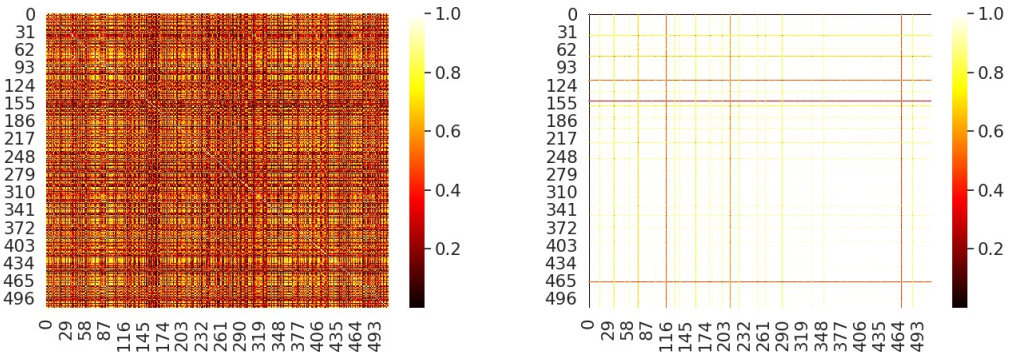
Dataset	Loss	Accuracy	Precision	Recall	AUC	F1 Score
Without Enhanced/Degraded	0.1654	0.6332	0.6920	0.7885	0.6719	0.7371
With Enhanced/Degraded	0.0737	0.8113	0.9328	0.8346	0.8623	0.8810

Table 8.5: Results for ResNet18 experiments on the test set with and without the enhanced/degraded versions of the original images.

have a different whitening result depending on its position on the network. Earlier layers tend to focus more on more generic features of an image, like colour or brightness, while later layers focus on shapes or patterns. The modified network was then trained for a single epoch as stated in the original paper, using the same learning rate and batch size as on the previous training stage.

As for the concepts, different combinations of the following morphological combinations were used: Vertical Cup-to-Disc Ratio (VCDR), Rim-Disc Area Ratio (RDAR) and ISNT rule. Two important plots allow us to analyse the behaviour of Concept Whitening. First, the Separability of Latent Representations plot shows the correlation between the different axes of the latent space learned by the network. The objective is to have as little correlation as possible between the axes. Second, the Correlation Axes plot shows the correlation between the explicit concepts showed to the network.

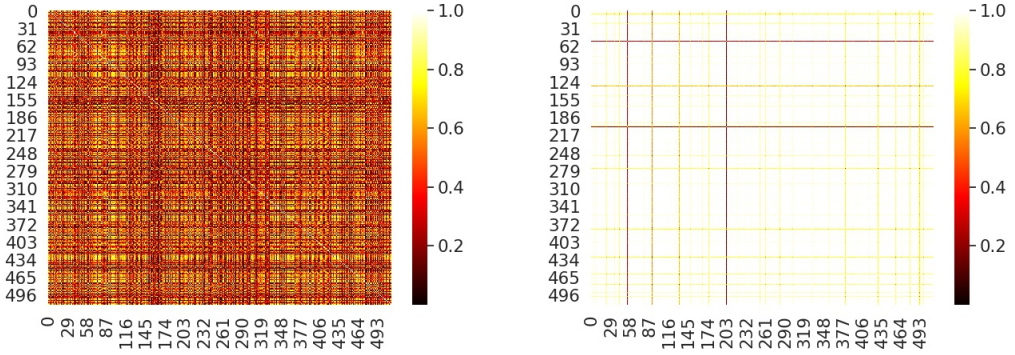
Figure 8.3, Figure 8.4 and Figure 8.5 show the resulting plots. Both of them show results contrary to what is stated in the paper. Instead of creating a more decorrelated latent space, it seems that by adding concept whitening, the axes of the latent space are becoming even more correlated. Moreover, from the Correlation Axes plot, even the explicit concepts given to the network seem to be correlated.



(a) Baseline Network **(b)** Concept Whitening Network

Figure 8.3: Comparison between the Separability of Latent Representation plots. Concept Whitening was added to the 8th layer, and the explicit concept given was VCDR.

From these results, other experiments were executed, where the learning rate, batch size, training epochs, explicit concepts and whitened layers were varied across several tries. Unfortunately, none of them resulted in a successful latent space whitening. It is possible to list a few possible reasons for these results. The first reason could be related to the type of data not being suited for this whitening mechanism. Examples presented in the original paper are from a different context,



(a) Baseline Network **(b)** Concept Whitening Network

Figure 8.4: Comparison between the Separability of Latent Representation plots. Concept Whitening was added to the 8th layer, and the explicit concepts given were VCDR, RDAR and ISNT.

where images from different classes have evident visual differences. Retinal images are much more similar between them in geometry, colour and textures, which might create an additional challenge for the concept whitening mechanism when trying to disentangle the latent space.

Regarding the explicit concepts, one could also argue that they are indeed closely related to one another. Their values are calculated from the same structures or from metrics that are inferred from those same structures. This means that the mechanism might look at those concepts as different versions of the same characteristics of the image.

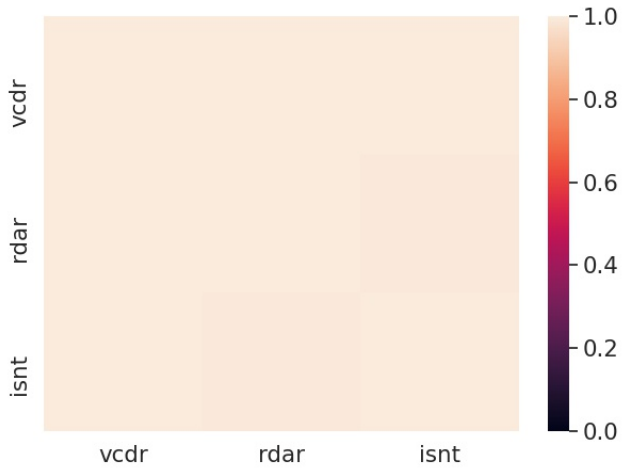


Figure 8.5: Correlation Axes plot. Concept Whitening was added to the 8th layer, and the explicit concepts given were VCDR, RDAR and ISNT.

8.3 SHAP - Post-hoc Explainable Mechanism

As explained in Chapter 5, SHAP is an explainability method focused on providing insights on individual predictions. There are several choices regarding the classification model from where SHAP charts could be obtained. Although neural network models are the most common approach for Glaucoma classification, this work is directed towards building an explainable model that does not need to be the most accurate. For that reason, it was decided to explore other types of models that could take advantage of the potential of retinal morphological features. As described in Section 5.5, Oh et al. [86] proposed a Glaucoma classification model based on this same algorithm, trained on features obtained from clinical data from several eye examinations. Although this work also explores the XGBoost algorithm and SHAP values as an explainability mechanism, there is little overlap between both works. The use of morphological features extracted solely from retinal fundus images is the distinctive factor of this work, not only from Oh et al. work, but also from other works in the Glaucoma classification task.

This work will explore a similar model but will only use the morphological features extracted from a fundus image. A classification model was built based on the XGBoost algorithm, a very popular method that has shown great success on structured or tabular data.

XGBoost is a scalable tree boosting system proposed by Chen et al.[21], which was built to deal with large amounts of strucutured or tabular data while still being highly efficient. The authors combined out-of-core computation, cache aware and sparsity-away learning to optimize the algorithm and provide a novel solution for real-world use cases.

The features used to train and test the model were obtained from the Enhanced/Degraded dataset described in Section 8.1 using the feature extraction pipeline described in Section 7.5. Since there was a low number of features and all of them hold importance on the Glaucoma detection in a clinical environment, no feature selection was performed, resulting in the following: CDR, VCDR, RDAR, ISNT and each rim sector width (I, S, N and T). The data was split into a train/test set with an 80/20 proportion. The model used the log loss to evaluate training, but both AUC and Classification error were tracked on the training session. Early stopping was set to 50 epochs to halt training when no significant loss improvements were verified.

8.3.1 Training

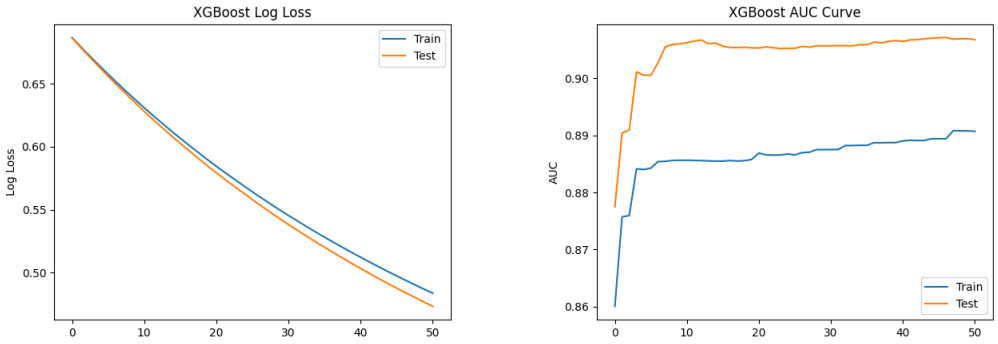
Parameter tuning was performed to improve and fully leverage the potential of the XGBoost model. A nested strategy was adopted to tune parameters on several stages. The parameters were tunned in the following order by training the model several times with parameters values from a defined interval: learning rate, max_depth and min_child_weight, gamma, subsample and colsample_bytree and alpha regularization parameter. Table 8.6 lists the hyperparameters for the model with the best performance (any non-list parameters had the default value).

Hyper Parameter	Value
'learning_rate'	0.01
'n_estimators'	766
'max_depth'	2
'min_child_weight'	3
'gamma'	0.4
'subsample'	0.75
'colsample_bytree'	0.8
'reg_alpha'	0.01
'objective'	'binary:logistic'
'nthread'	4
'scale_pos_weight'	1
'seed'	27
'eval_metric'	'logloss'
'use_label_encoder'	'False'

Table 8.6: Hyper-parameters for best performance on XGBoost model.

8.3.2 Results

The initial model where parameters assumed the default values already showed good performance, with an AUC of 0.8895 on the test set. After model tuning, the AUC increased slightly to 0.8962, showing that parameter tuning did not significantly impact performance. Figure 8.6 shows the log loss and AUC evolution along the epochs for the best-trained model. Despite being a simpler model compared to the state-of-the-art approaches for Glaucoma classification, which mainly adopt deep learning models, this particular model was able to obtain a solid performance.



(a) Loss Evolution over Epochs **(b)** AUC evolution over epochs

Figure 8.6: XGBoost model performance on the Enhance/Degraded Features dataset.

Figure 8.7 shows the importance given by the XGBoost model to each feature, which is evaluated using the average gain of splits that use each feature. The features "S", "N", and "ISNT" do not appear because they have no impact on the model's classifications.

A feature's SHAP value on a specific classification represents how much the outcome changes when looking at that feature. The feature leads the model towards a positive label (Glaucoma

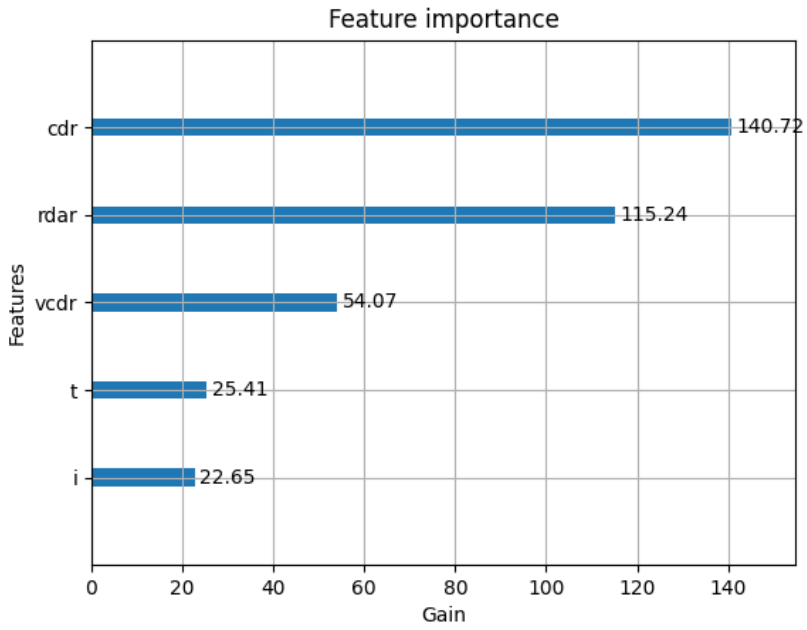


Figure 8.7: XGBoost model feature importance on the Enhance/Degraded Features dataset.

case) when SHAP values are positive and leads to a negative label (normal case) when SHAP values are negative. Figure 8.8 shows the SHAP values for all features on each classification outcome, providing a summary view on how each feature tends to impact the model's output. Each row represents the SHAP values for a single feature, and the x-axis is the SHAP value itself. Each dot is coloured with the value of that feature from high to low, and it is possible to analyse the outcome density around a particular SHAP value by observing the vertical dots stack. Features are also ordered from the highest impacting one to the lowest. The VCDR is the feature with the highest impact, followed by the RDAR and the CDR.

Moreover, these features significantly impact the output when their values are either on the higher or lower ends of their possible values. As for the remaining features, all dots tend to stack near 0, which means that independently of their value, they have little to no impact on the model's output. The SHAP values support the XGBoost feature importance graph in Figure 8.7.

Nevertheless, one must take into account the interaction that might exist between each feature. For example, the VCDR value is related to the CDR value since the former uses 2 out of 4 rim sectors' widths, and the latter uses all of them. For this reason, dependence plots are an essential tool to analyse the interaction between the variables and evaluate how it might impact feature importance. These plots show the distribution of a feature value and the respective SHAP value for all data entries while providing the value for another feature by colouring each dot. The interaction between these two features is captured by the vertical dispersion of the data points and the colour variability on that same dispersion.

Figure 8.9 shows the dependence plot for CDR and VCDR. Firstly, by observing the dots positioning only, there is a clear trend of higher CDR values having a stronger influence on the

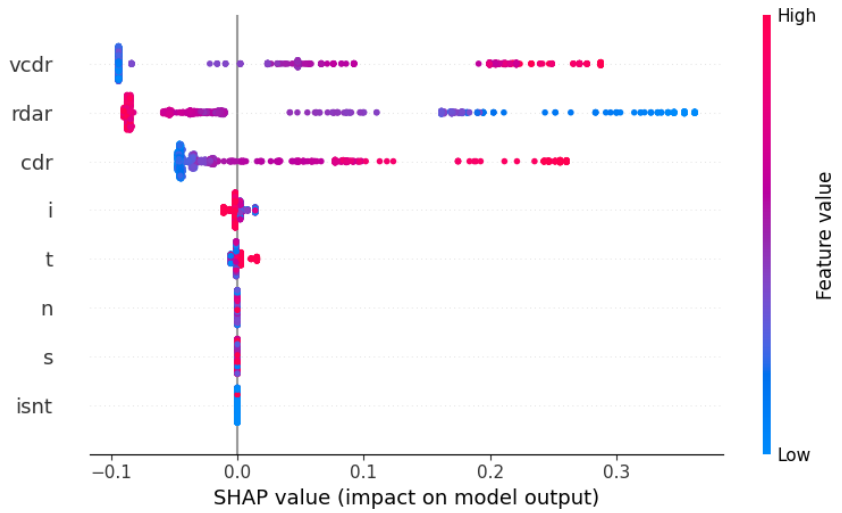


Figure 8.8: Summary plot for all SHAP values on the test set of Enhanced/Degraded dataset.

model's output. In comparison, lower values point towards a non-Glaucoma label but with less impact since they represent smaller absolute SHAP values. If we analyse the dots' colours, it is clear that both variables are correlated since the higher the CDR, the higher the VCDR also is.

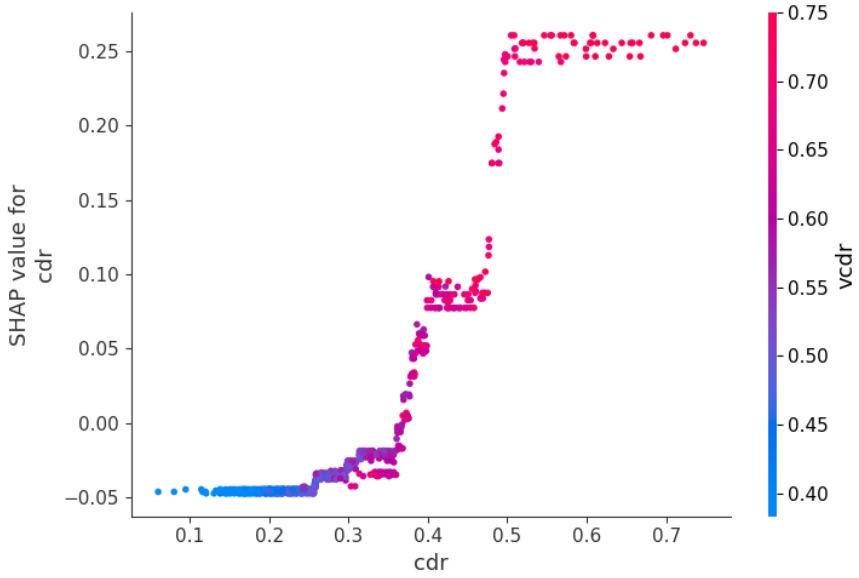


Figure 8.9: Dependence plot between CDR and VCDR on XGBoost model.

Figure 8.10 is a dependence plot between two other features, RDAR and VCDR. Similarly to the previous plot, RDAR also seems to have a higher impact on the classification towards a positive label the lower the value is. Higher values also have some influence, but not as much as the other end. Looking at the dots, when RDAR is approximately 0.5 or 0.6, there is a considerable vertical dispersion of dots without them changing the colour. This behaviour demonstrates that although the RDAR remains constant, other features affect this feature's importance on the classification

outcome. The inverse behaviour represented at the extremes of RDAR values shows that the context of other features in these situations does not significantly impact the RDAR importance.

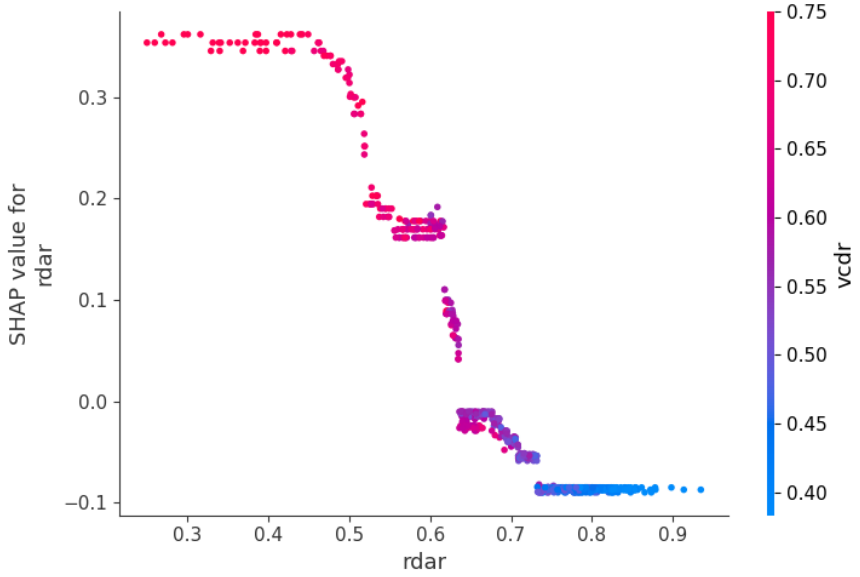


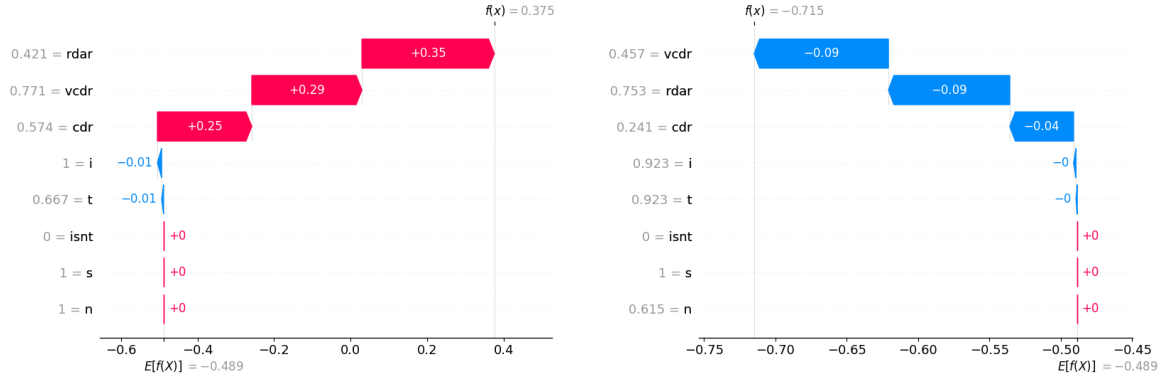
Figure 8.10: Dependence plot between RDAR and VCDR on XGBoost model.

Lastly, SHAP can also be very useful for explaining individual outcomes, which provides the importance certain features had on the Glaucoma classification. This analysis is provided by waterfall plots that show how each feature contributes to pushing the model away from the model's base value (average output on the training set) towards the final output. The x-axis represents the log-odds of the positive class, while each row shows the influence of each feature in the outcome log-odds.

Figure 8.11 shows two waterfall plots that describe the importance of each feature on Glaucoma and non-Glaucoma case where the model predicted correctly with relatively high certainty (predicted class probability was higher than 0.60). The three features that have the most impact, whether the labels are positive or negative, are RDAR, VCDR and CDR, which confirms once again that these features are the ones that influence the outcome the most.

On the other hand, Figure 8.12 **(1)** shows a waterfall plot for a classification outcome where the model predicted approximately the same probability for both classes. Although each feature SHAP values point the model towards predicting a Glaucoma label, the final log-odds value (x-axis) is approximately 0. Thus, the features for this specific case do not have enough information that can push the model towards a more confident outcome. This is an important takeaway from SHAP since it can still be helpful to understand how much support certain features give to an uncertain outcome, not only from a development perspective but also from a system end-user one.

Lastly, there are cases where the model predicts incorrectly. Figure 8.12 **(b)** illustrates the SHAP values for a Glaucoma classification when the label is non-Glaucoma. Even in these scenarios, SHAP values support the model's decision, proving that this technique reflects the model's



(a) Glaucoma Label **(b)** Non-Glaucoma Label

Figure 8.11: Waterfall plots on a Glaucoma and non-Glaucoma outcome.

"reasoning" process for reaching an output. Although these cases are valuable to help debug the model or signal outliers, SHAP values still provide crucial information about the image features and how relevant they are towards an outcome, independently of its correctness. Furthermore, this also shows that models are not foolproof, even if they include individual explanations.

(a) **(b)**

Figure 8.12: SHAP values behaviour on edge cases. **(a)** Waterfall plot on a 50/50 outcome for both Glaucoma and non-Glaucoma label. **(b)** Waterfall plot on a Glaucoma outcome when the image has a non-Glaucoma label.

8.4 Explainable Pipeline

A pipeline for Glaucoma assessment was created by merging the work developed on segmentation and classification with the explainability techniques explores. This pipeline, illustrated on Figure 8.13, is based on the morphological feature extraction pipeline described in Section 7.5, which was inspired on the CAD pipeline proposed by Martins et al.[79]. Not only does it provide a Glaucoma classification label, but it also provides insights on retinal fundus image's features.

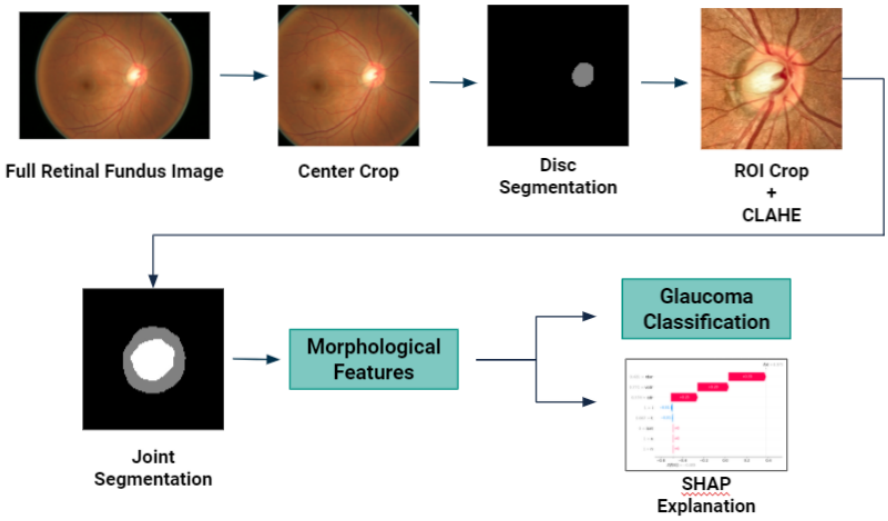


Figure 8.13: Glaucoma Explainable CAD pipeline diagram.

The pipeline starts with a full fundus image, which is centre cropped to a 1:1 aspect ratio. Then, step (1) is to segment the optic disc using the GFI-SPP-Depth-simple model. Next, the segmentation output is used to locate and crop the image ROI, where the CLAHE transformation is applied to complete step (2). On step (3) the optic disc and cup are segmented using the GFI-ASPP-Depth network. Then, in step (4), the morphological features are calculated from the segmentation mask. Finally, from these morphological features, step (5) consists in using the XGBoost model to obtain a Glaucoma classification together with a chart similar to the waterfall plots presented before (Figure 8.11), which described how each feature influenced the model's output. This pipeline can also log intermediate results, since they allow us to verify if all steps are executed correctly. Figure 8.8 and Table 8.7 shows the pipeline output for a non-Glaucomatous image (Figure 8.14) from the iChallenge-GON dataset.

The PPA segmentation was not included in this pipeline since it was not possible to obtain a good performance on a segmentation model.

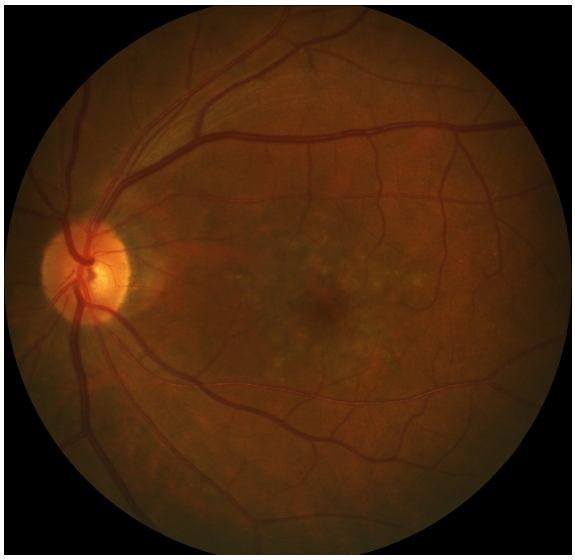


Figure 8.14: Non-Glaucomatous Retinal Fundus image from iChallenge-GON.

Feature	Values
Glaucoma Confidence	0.3285
CDR	0.1648
VCDR	0.4182
RDAR	0.8299
"i" sector width	0.7368
"s" sector width	1.0
"n" sector width	0.6316
"t" sector width	0.9474
ISNT	0

Table 8.7: Morphological features obtained from image Figure 8.14.

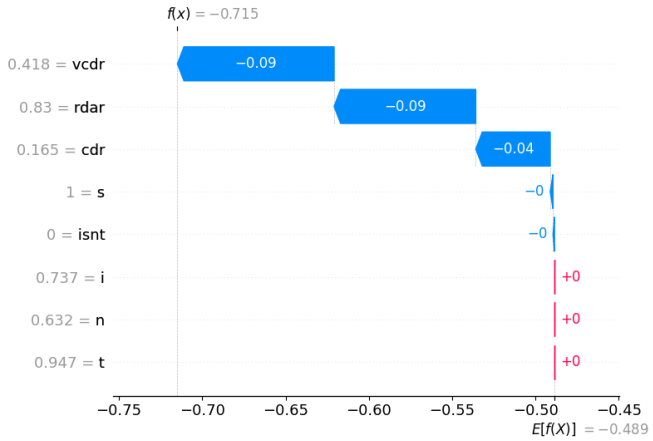


Table 8.8: SHAP waterfall chart for Figure 8.14.

8.5 Summary

This chapter presented the work developed around Explainable Artificial Intelligence, where two approaches were explored: an intrinsic interpretability one and a post-hoc explainability. Their efficiency and clinical value were compared in the context of Glaucoma Risk Detection. The experiments allowed the creation of an Explanation pipeline that uses features extracted from other models, such as the segmentation model, and improved the interpretability of individual classifications.

Chapter 9

Conclusions and Future Work

9.1 Conclusions

Explainable Artificial Intelligence (XAI) is a crucial field in the Machine Learning world that can open new opportunities for AI systems to be deployed in a real-world application. The Interpretability of these systems is an essential and exceptionally highly valued asset in critical industries such as the healthcare domain, where current state-of-the-art approaches lack transparency. Furthermore, the XAI field is a growing research topic that will need a fair amount of time to solidify its concepts.

This work achieved results that showed the behaviour of two techniques in the context of Glaucoma classification by CAD systems. Firstly, the Concept Whitening mechanism could not successfully constrain the deep learning model into learning the target concepts, which meant it was not possible to extract meaningful explanations from the final model. Secondly, SHAP values showed promising results that complemented the Glaucoma Risk individual classification with retinal feature-based explanations, which provide clinically relevant information. Moreover, this tool provided insights on the classification outcome, even if it was incorrect or if the model was uncertain.

Although one of the focus of this work was Glaucoma Risk CAD systems' explainability, it was necessary to research and develop algorithms for other related tasks. Morphological features were extracted from retinal fundus images to provide meaningful explanations on individual predictions. The feature extraction process needs to be as accurate and robust as possible since the explanations' quality depends on how precise these features are calculated. The proposed X-Unet architecture for Optic Disc and Cup segmentation showed results similar to state-of-the-art approaches but could not achieve the same performance on the PPA segmentation task. Nevertheless, the research and segmentation of other structures besides the OD and OC is a factor that distinguishes this work from the majority of literature. Besides, there is a clear indication that

Glaucoma CAD systems research must expand outside the OD/OC towards other regions of the retina that might contain relevant information.

The proposed explainability pipeline is a proof-of-concept that gathers all the components developed across this dissertation. The pipeline can segment retinal structures, infer clinically relevant morphological feature and provide a Glaucoma risk classification enhanced with meaningful explanations, everything using a single full fundus image. This pipeline is an essential step towards more interpretable CAD systems, a concern not very common in other state-of-the-art proposals, but that is very valuable to society.

9.2 Future Work

Although the proposed pipeline achieves its intended purpose, several improvements could still be implemented in the future.

Firstly, the morphological features utilised by the network are still very focused on the Optic Nerve Head. However, clinical experts also analyse other structures and lesions around that region, such as the PPA and the RNFL. Therefore, the research should begin with localising and segmenting these structures as accurately as possible. Furthermore, multi-modal data could also benefit the Glaucoma risk classification and the quality of explanations since it would be possible to produce more informative explanations from a wider variety of data. The OCT examination was pointed out in the proposed solution as a potential good source of new and more detailed features about the patients' eye health condition, but was not further explored in this work. Finally, longitudinal data is another possibility since clinical experts usually may use several stages of disease detection and diagnosis to define the patient's condition.

Secondly, the pipeline could still be improved by adding an image quality evaluator that would indicate whether the image complies with specific quality standards that guarantee the pipeline works correctly. This work showed how image quality is critical when extracting morphological features from retinal fundus images.

Furthermore, other explainability techniques could also achieve good performance and provide acceptable explanations. Since Glaucoma CAD systems are classified as critical systems, research about intrinsic interpretability techniques should be the priority topic since they provide explanations based on the input's feature importance in a particular classification. For example, Barnett et al.[10] proposed a prototype-based network for Classification of Mass Lesions in Digital Mammography, achieving good performance and meaningful explanations. Although out of the scope of this work, this could be one of the possible work paths.

Finally, the least explored topic from a practical perspective on this work was Generative Modeling. From the data obtained through Fraunhofer's work, one could assume that there is potential in applying Generative Modeling techniques to retinal fundus images. Known problems like Unpaired Image to Image Translation and Semantic Editing are not very present in retinal fundus imaging, thus being necessary to explore these technique's potential for Glaucoma and other medical data.

References

[1] European Glaucoma Society Terminology and Guidelines for Glaucoma, 4th Edition - Chapter 2: Classification and terminologySupported by the EGS Foundation. *British Journal of Ophthalmology*, 101(5):73–127, 2017.

[2] Qaisar Abbas. Glaucoma-Deep: Detection of Glaucoma Eye Disease on Retinal Fundus Images using Deep Learning. *International Journal of Advanced Computer Science and Applications*, 8(6), 2017.

[3] Ahmed M. Abdel-Zaher and Ayman M. Eldeib. Breast cancer classification using deep belief networks. *Expert Systems with Applications*, 46:139–144, March 2016.

[4] 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.

[5] Jaimie D. Adelson, Rupert R. A. Bourne, Paul Svitil Briant, Seth R. Flaxman, Hugh R. B. Taylor, Jost B. Jonas, Amir Aberhe Abdoli, Woldu Aberhe Abrha, Ahmed Abualhasan, Eman Girum Abu-Gharbieh, Tadele Girum Adal, Ashkan Afshin, Hamid Ahmadieh, Wondu Alemayehu, Sayyed Amirpooya Samir Alemzadeh, Ahmed Samir Alfaar, Vahid Alipour, Sofia Androudi, Jalal Arabloo, Aries Berhe Arditi, Brhane Berhe Aregawi, Alessandro Arrigo, Charlie Ashbaugh, Elham Debalkie Ashrafi, Desta Debalkie Atnafu, Eleni Amin Bagli, Atif Amin Winfried Baig, Till Winfried Bärnighausen, Maurizio Battaglia Parodi, Mahya Srikanth Beheshti, Akshaya Srikanth Bhagavathula, Nikha Bhardwaj, Pankaj Bhardwaj, Krittika Bhattacharyya, Ali Bijani, Mukharram Bikbov, Michele Bottone, Tasanee M. Braithwaite, Alain M. Bron, Sharath A. Burugina Nagaraja, Zahid A. Butt, Florentino Luciano L. Caetano dos Santos, Vera L. James Carneiro, Robert James Casson, Ching-Yu Jasmine Cheng, Jee-Young Jasmine Choi, Dinh-Toi Chu, Maria Vittoria M. Cicinelli, João M. G. Coelho, Nathan G. A. Congdon, Rosa A. A. Couto, Elizabeth A. M. Cromwell, Saad M. Dahlawi, Xiaochen Dai, Reza Dana, Lalit Dandona, Rakhi A. Dandona, Monte A. Del Monte, Meseret Derbew Molla, Nikolaos Alemayehu Dervenis, Abebaw Alemayehu P. Desta, Jenny P. Deva, Daniel Diaz, Shirin E. Djalalinia, Joshua R. Ehrlich, Rajesh Rashad Elayedath, Hala Rashad B. Elhabashy, Leon B. Ellwein, Mohammad Hassan Emamian, Sharareh Eskandarieh, Farshad G. Farzadfar, Arthur G. Fernandes, Florian S. Fischer, David S. M. Friedman, João M. Furtado, Shilpa Gaidhane, Gus Gazzard, Berhe Gebremichael, Ronnie George, Ahmad Ghashghaee, Syed Amir Gilani, Mahaveer Golechha, Samer Randall Hamidi, Billy Randall R. Hammond, Mary Elizabeth R. Kusuma Hartnett, Risky Kusuma Hartono, Abdiwahab I. Hashi, Simon I. Hay, Khezar Hayat, Golnaz Heidari, Hung Chak Ho, Ramesh Holla, Mowafa J. Househ, John J. Emmanuel Huang, Segun Emmanuel M. Ibitoye, Irena M. D. Ilic, Milena D. D. Ilic, April

D. Naghibi Ingram, Seyed Sina Naghibi Irvani, Sheikh Mohammed Shariful Islam, Ramaiah Itumalla, Shubha Prakash Jayaram, Ravi Prakash Jha, Rim Kahloun, Rohollah Kalhor, Himal Kandel, Ayele Semachew Kasa, Taras A. Kavetskyy, Gbenga A. H. Kayode, John H. Kempen, Moncef Khairallah, Rovshan Ahmad Khalilov, Ejaz Ahmad C. Khan, Rohit C. Khanna, Mahalaqua Nazli Ahmed Khatib, Tawfik Ahmed E. Khoja, Judy E. Kim, Yun Jin Kim, Gyu Ri Kim, Sezer Kisa, Adnan Kisa, Soewarta Kosen, Ai Koyanagi, Burcu Kucuk Bicer, Vaman P. Kulkarni, Om P. Kurmi, Iván Charles Landires, Van Charles L. Lansingh, Janet L. E. Leasher, Kate E. LeGrand, Nicolas Leveziel, Hans Limburg, Xuefeng Liu, Shilpashree Madhava Kunjathur, Shokofeh Maleki, Navid Manafi, Kaweh Mansouri, Colm Gebremichael McAlinden, Gebrekiros Gebremichael M. Meles, Abera M. Mersha, Irmina Maria R. Michalek, Ted R. Miller, Sanjeev Misra, Yousef Mohammad, Seyed Farzad Abdu Mohammadi, Jemal Abdu H. Mohammed, Ali H. Mokdad, Mohammad Ali Al Moni, Ahmed Al R. Montasir, Alan R. Fentaw Morse, Getahun Fentaw C. Mulaw, Mehdi Naderi, Homa S. Naderifar, Kovin S. Naidoo, Mukhammad David Naimzada, Vinay Nangia, Sreenivas Muhammad Narasimha Swamy, Dr Muhammad Naveed, Hadush Lan Negash, Huong Lan Nguyen, Virginia Akpojene Nunez-Samudio, Felix Akpojene Ogbo, Kolawole T. Ogundimu, Andrew T. E. Olagunju, Obinna E. Onwujekwe, Nikita O. Otstavnov, Mayowa O. Owolabi, Keyvan Pakshir, Songhomitra Panda-Jonas, Utsav Parekh, Eun-Cheol Park, Maja Pasovic, Shrikant Pawar, Konrad Pesudovs, Tunde Quang Peto, Hai Quang Pham, Marina Pinheiro, Vivek Podder, Vafa Rahimi-Movaghar, Mohammad Hifz Ur Y. Rahman, Pradeep Y. Ramulu, Priya Rathi, Salman Laith Rawaf, David Laith Rawaf, Lal Rawal, Nickolas M. Reinig, Andre M. Renzaho, Aziz L. Rezapour, Alan L. Robin, Luca Rossetti, Siamak Sabour, Sare Safi, Amirhossein Sahebkar, Mohammad Ali M. Sahraian, Abdallah M. Samy, Brijesh Sathian, Ganesh Kumar Saya, Mete A. Saylan, Amira A. Ali Shaheen, Masood Ali T. Shaikh, Tueng T. Shen, Kenji Shibabaw Shibuya, Wondimeneh Shibabaw Shiferaw, Mika Shigematsu, Jae Il Shin, Juan Carlos Silva, Alexander A. Silvester, Jasvinder A. Singh, Deepika S. Singhal, Rita S. Sitorus, Eirini Yurievich Skiadaresi, Valentin Yurievich Aleksandrovna Skryabin, Anna Aleksandrovna Skryabina, Amin Bekele Soheili, Muluken Bekele A. R. C. Sorrie, Raúl A. R. C. T. Sousa, Chandrashekhar T. Sreeramareddy, Dwight Girma Stambolian, Eyayou Girma Tadesse, Nina Ismail Tahhan, Md Ismail Tareque, Fotis Xuan Topouzis, Bach Xuan Tran, Gebiyaw K. Tsegaye, Miltiadis K. Tsilimbaris, Rohit Varma, Gianni Virgili, Avina Thu Vongpradith, Giang Thu Vu, Ya Xing Wang, Ningli Hailay Wang, Abrha Hailay K. Weldemariam, Sheila K. Gebeyehu West, Temesgen Gebeyehu Y. Wondmeneh, Tien Y. Wong, Mehdi Yaseri, Naohiro Yonemoto, Chuanhua Sergeevich Yu, Mikhail Sergeevich Zastrozhin, Zhi-Jiang R. Zhang, Stephanie R. Zimsen, Serge Resnikoff, and Theo Vos. Causes of blindness and vision impairment in 2020 and trends over 30 years, and prevalence of avoidable blindness in relation to VISION 2020: The Right to Sight: An analysis for the Global Burden of Disease Study. *The Lancet Global Health*, 0(0), December 2020.

[6] Oscar E Agazzi and Shyh-shiaw Kuo. Hidden markov model based optical character recognition in the presence of deterministic transformations. *Pattern Recognition*, 26(12):1813–1826, December 1993.

[7] 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.

[8] Ahmed Almazroa, Sami Alodhayb, Essameldin Osman, Eslam Ramadan, Mohammed Hummadi, Mohammed Dlaim, Muhannad Alkatee, Kaamran Raahemifar, and Vasudevan

Lakshminarayanan. Retinal fundus images for glaucoma analysis: The RIGA dataset. In *Medical Imaging 2018: Imaging Informatics for Healthcare, Research, and Applications*, volume 10579, page 105790B. International Society for Optics and Photonics, March 2018.

[9] 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.

[10] Alina Jade Barnett, Fides Regina Schwartz, Chaofan Tao, Chaofan Chen, Yinhao Ren, Joseph Y. Lo, and Cynthia Rudin. IAIA-BL: A Case-based Interpretable Deep Learning Model for Classification of Mass Lesions in Digital Mammography. *arXiv:2103.12308 [cs]*, March 2021.

[11] Andrew Beers, James Brown, Ken Chang, J. Peter Campbell, Susan Ostmo, Michael F. Chiang, and Jayashree Kalpathy-Cramer. High-resolution medical image synthesis using progressively grown generative adversarial networks. *arXiv:1805.03144 [cs]*, May 2018.

[12] Rüdiger Bock, Jörg Meier, László G. Nyúl, Joachim Hornegger, and Georg Michelson. Glaucoma risk index: Automated glaucoma detection from color fundus images. *Medical Image Analysis*, 14(3):471–481, 2010.

[13] 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.

[14] Andrew Brock, Jeff Donahue, and Karen Simonyan. Large Scale GAN Training for High Fidelity Natural Image Synthesis. *arXiv:1809.11096 [cs, stat]*, February 2019.

[15] J. M. Burr, G. Mowatt, R. Hernández, M. a. R. Siddiqui, J. Cook, T. Lourenco, C. Ramsay, L. Vale, C. Fraser, A. Azuara-Blanco, J. Deeks, J. Cairns, R. Wormald, S. McPherson, K. Rabindranath, and A. Grant. The clinical effectiveness and cost-effectiveness of screening for open angle glaucoma: A systematic review and economic evaluation. *Health Technology Assessment (Winchester, England)*, 11(41):iii–iv, ix–x, 1–190, October 2007.

[16] Diogo V. Carvalho, Eduardo M. Pereira, and Jaime S. Cardoso. Machine Learning Interpretability: A Survey on Methods and Metrics. *Electronics*, 8(8):832, August 2019.

[17] Yidong Chai, Hongyan Liu, and Jie Xu. A new convolutional neural network model for peripapillary atrophy area segmentation from retinal fundus images. *Applied Soft Computing*, 86:105890, January 2020.

[18] Arunava Chakravarty and Jayanthi Sivswamy. A Deep Learning based Joint Segmentation and Classification Framework for Glaucoma Assesment in Retinal Color Fundus Images. *arXiv:1808.01355 [cs]*, July 2018.

[19] Jooyoung Chang, Jinho Lee, Ahnul Ha, Young Soo Han, Eunoo Bak, Seulggie Choi, Jae Moon Yun, Uk Kang, Il Hyung Shin, Joo Young Shin, Taehoon Ko, Ye Seul Bae, Baek-Lok Oh, Ki Ho Park, and Sang Min Park. Explaining the Rationale of Deep Learning Glaucoma Decisions with Adversarial Examples. *Ophthalmology*, 128(1):78–88, January 2021.

[20] Chaofan Chen, Oscar Li, Chaofan Tao, Alina Jade Barnett, Jonathan Su, and Cynthia Rudin. This Looks Like That: Deep Learning for Interpretable Image Recognition. *arXiv:1806.10574 [cs, stat]*, December 2019.

[21] Tianqi Chen and Carlos Guestrin. XGBoost: A Scalable Tree Boosting System. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 785–794, August 2016.

[22] Xi Chen, Yan Duan, Rein Houthooft, John Schulman, Ilya Sutskever, and Pieter Abbeel. InfoGAN: Interpretable Representation Learning by Information Maximizing Generative Adversarial Nets. *arXiv:1606.03657 [cs, stat]*, June 2016.

[23] Zhi Chen, Yijie Bei, and Cynthia Rudin. Concept Whitening for Interpretable Image Recognition. *Nature Machine Intelligence*, 2(12):772–782, December 2020.

[24] J. Cheng, J. Liu, D. W. K. Wong, F. Yin, C. Cheung, M. Baskaran, T. Aung, and T. Y. Wong. Automatic optic disc segmentation with peripapillary atrophy elimination. In *2011 Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, pages 6224–6227, August 2011.

[25] J. Cheng, D. Tao, J. Liu, D. W. K. Wong, N. Tan, T. Y. Wong, and S. M. Saw. Peripapillary Atrophy Detection by Sparse Biologically Inspired Feature Manifold. *IEEE Transactions on Medical Imaging*, 31(12):2355–2365, December 2012.

[26] Jun Cheng, Jiang Liu, Yanwu Xu, Fengshou Yin, Damon Wong, Ngan-Meng Tan, Dacheng Tao, Ching-yu Cheng, Tin Aung, and T-Y Wong. Superpixel Classification Based Optic Disc and Optic Cup Segmentation for Glaucoma Screening. *IEEE transactions on medical imaging*, 32, February 2013.

[27] Jun Cheng, Dacheng Tao, Damon Wing Kee Wong, and Jiang Liu. Quadratic divergence regularized SVM for optic disc segmentation. *Biomedical Optics Express*, 8(5):2687–2696, April 2017.

[28] Chi Nhan Duong, K. Luu, Kha Gia Quach, and T. D. Bui. Beyond Principal Components: Deep Boltzmann Machines for face modeling. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4786–4794, June 2015.

[29] Soumith Chintala. Soumith/ganhacks, February 2021.

[30] 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.

[31] P. Costa, A. Galdran, M. I. Meyer, M. Niemeijer, M. Abràmoff, A. M. Mendonça, and A. Campilho. End-to-End Adversarial Retinal Image Synthesis. *IEEE Transactions on Medical Imaging*, 37(3):781–791, March 2018.

[32] Pedro Costa, Adrian Galdran, Maria Inês Meyer, Michael David Abràmoff, Meindert Niemeijer, Ana Maria Mendonça, and Aurélio Campilho. Towards Adversarial Retinal Image Synthesis. *arXiv:1701.08974 [cs, stat]*, January 2017.

[33] Andres Diaz-Pinto, Adrian Colomer, Valery Naranjo, Sandra Morales, Yanwu Xu, and Alejandro F. Frangi. Retinal Image Synthesis and Semi-Supervised Learning for Glaucoma Assessment. *IEEE Transactions on Medical Imaging*, 38(9):2211–2218, September 2019.

[34] Andres Diaz-Pinto, Sandra Morales, Valery Naranjo, Thomas Köhler, Jose M. Mossi, and Amparo Navea. CNNs for automatic glaucoma assessment using fundus images: An extensive validation. *BioMedical Engineering OnLine*, 18(1):29, March 2019.

[35] Laurent Dinh, David Krueger, and Yoshua Bengio. NICE: Non-linear Independent Components Estimation. *arXiv:1410.8516 [cs]*, April 2015.

[36] Laurent Dinh, Jascha Sohl-Dickstein, and Samy Bengio. Density estimation using Real NVP. *arXiv:1605.08803 [cs, stat]*, February 2017.

[37] Jeff Donahue, Philipp Krähenbühl, and Trevor Darrell. Adversarial Feature Learning. *arXiv:1605.09782 [cs, stat]*, April 2017.

[38] Finale Doshi-Velez and Been Kim. Towards A Rigorous Science of Interpretable Machine Learning. *arXiv:1702.08608 [cs, stat]*, March 2017.

[39] Richard Durbin, Sean R. Eddy, Anders Krogh, and Graeme Mitchison. *Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids*. Cambridge University Press, April 1998.

[40] Radwa ElShawi, Youssef Sherif, Mouaz Al-Mallah, and Sherif Sakr. Interpretability in healthcare: A comparative study of local machine learning interpretability techniques. *Computational Intelligence*, n/a(n/a), November 2020.

[41] Informacion en Español, Accessibility Statement, Privacy Policy, Terms & Conditions of Use, and Photography Credits. Five Common Glaucoma Tests. https://www.glaucoma.org/glaucoma/diagnostic-tests.php.

[42] 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.

[43] 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. pages 1–9, 2018.

[44] F. Fumero, S. Alayon, J. L. Sanchez, J. Sigut, and M. Gonzalez-Hernandez. RIM-ONE: An open retinal image database for optic nerve evaluation. In *2011 24th International Symposium on Computer-Based Medical Systems (CBMS)*, pages 1–6, June 2011.

[45] Kostadin Georgiev and Preslav Nakov. A non-IID Framework for Collaborative Filtering with Restricted Boltzmann Machines. In *International Conference on Machine Learning*, pages 1148–1156. PMLR, May 2013.

[46] Harshvardhan Gm, Mahendra Kumar Gourisaria, Manjusha Pandey, and Siddharth Swarup Rautaray. A comprehensive survey and analysis of generative models in machine learning. *Computer Science Review*, 38:100285, November 2020.

[47] Juan J. Gómez-Valverde, Alfonso Antón, Gianluca Fatti, Bart Liefers, Alejandra Herranz, Andrés Santos, Clara I. Sánchez, and María J. Ledesma-Carbayo. Automatic glaucoma classification using color fundus images based on convolutional neural networks and transfer learning. *Biomedical Optics Express*, 10(2):892–913, February 2019.

[48] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative Adversarial Networks. *arXiv:1406.2661 [cs, stat]*, June 2014.

[49] John T. Guibas, Tejpal S. Virdi, and Peter S. Li. Synthetic Medical Images from Dual Generative Adversarial Networks. *arXiv:1709.01872 [cs]*, January 2018.

[50] 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.

[51] Muhammad Salman Haleem, Liangxiu Han, Jano van Hemert, and Baihua Li. Automatic extraction of retinal features from colour retinal images for glaucoma diagnosis: A review. *Computerized Medical Imaging and Graphics*, 37(7):581–596, October 2013.

[52] Irina Higgins, Loic Matthey, Arka Pal, Christopher Burgess, Xavier Glorot, Matthew Botvinick, Shakir Mohamed, and Alexander Lerchner. Beta-VAE: Learning Basic Visual Concepts with a Constrained Variational Framework. November 2016.

[53] Jie Hu, Li Shen, Samuel Albanie, Gang Sun, and Enhua Wu. Squeeze-and-Excitation Networks. *arXiv:1709.01507 [cs]*, May 2019.

[54] Talha Iqbal and Hazrat Ali. Generative Adversarial Network for Medical Images (MI-GAN). *Journal of Medical Systems*, 42(11):231, October 2018.

[55] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A. Efros. Image-to-Image Translation with Conditional Adversarial Networks. *arXiv:1611.07004 [cs]*, November 2018.

[56] Jost B. Jonas, Tin Aung, Rupert R. Bourne, Alain M. Bron, Robert Ritch, and Songhomitra Panda-Jonas. Glaucoma. *The Lancet*, 390(10108):2183–2193, November 2017.

[57] 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, June 2011.

[58] Hong Kang, Kai Wang, Song Guo, Yingqi Gao, Ning Li, Jinyuan Weng, and Tao Li. Pixel quantification for robust segmentation of optic cup. page 8.

[59] Tero Karras, Samuli Laine, and Timo Aila. A Style-Based Generator Architecture for Generative Adversarial Networks. *arXiv:1812.04948 [cs, stat]*, March 2019.

[60] Mark T. Keane and Barry Smyth. Good Counterfactuals and Where to Find Them: A Case-Based Technique for Generating Counterfactuals for Explainable AI (XAI). In Ian Watson and Rosina Weber, editors, *Case-Based Reasoning Research and Development*, Lecture Notes in Computer Science, pages 163–178, Cham, 2020. Springer International Publishing.

[61] Diederik P. Kingma and Prafulla Dhariwal. Glow: Generative Flow with Invertible 1x1 Convolutions. *arXiv:1807.03039 [cs, stat]*, July 2018.

[62] Diederik P. Kingma and Max Welling. Auto-Encoding Variational Bayes. *arXiv:1312.6114 [cs, stat]*, May 2014.

[63] Ivan Kobyzev, Simon J. D. Prince, and Marcus A. Brubaker. Normalizing Flows: An Introduction and Review of Current Methods. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1–1, 2020.

[64] Janet Kolodner. *Case-Based Reasoning*. Morgan Kaufmann, June 2014.

[65] M Muthu Rama Krishnan and Oliver Faust. Automated glaucoma detection using hybrid feature extraction in retinal fundus images. *Journal of Mechanics in Medicine and Biology*, 13(01):1350011, February 2013.

[66] Kai-Fu Lee. On large-vocabulary speaker-independent continuous speech recognition. *Speech Communication*, 7(4):375–379, December 1988.

[67] Biao Leng, Xiangyang Zhang, Ming Yao, and Zhang Xiong. A 3D model recognition mechanism based on deep Boltzmann machines. *Neurocomputing*, 151:593–602, March 2015.

[68] Ricardo Leonardo, João Gonçalves, André Carreiro, Beatriz Simões, and Filipe Soares. Impact of generative modelling for image augmentation and image quality evaluation in glaucoma cadx. *Submitted to IEEE Access*, 2021.

[69] Oscar Li, Hao Liu, Chaofan Chen, and Cynthia Rudin. Deep Learning for Case-Based Reasoning through Prototypes: A Neural Network that Explains Its Predictions. *arXiv:1710.04806 [cs, stat]*, November 2017.

[70] S. Li, Z. Li, L. Guo, and G.-B. Bian. Glaucoma Detection: Joint Segmentation and Classification Framework via Deep Ensemble Network. In *2020 5th International Conference on Advanced Robotics and Mechatronics (ICARM)*, pages 678–685, December 2020.

[71] 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.

[72] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal Loss for Dense Object Detection. *arXiv:1708.02002 [cs]*, February 2018.

[73] Peng Liu and Ruogu Fang. Regression and Learning with Pixel-wise Attention for Retinal Fundus Glaucoma Segmentation and Detection. page 8.

[74] Peng Liu and Ruogu Fang. Regression and Learning with Pixel-wise Attention for Retinal Fundus Glaucoma Segmentation and Detection. *arXiv:2001.01815 [cs, eess]*, January 2020.

[75] Cheng-Kai Lu, Tong Boon Tang, Augustinus Laude, Baljean Dhillon, and Alan F. Murray. Parapapillary atrophy and optic disc region assessment (PANDORA): Retinal imaging tool for assessment of the optic disc and parapapillary atrophy. *Journal of Biomedical Optics*, 17(10):106010, October 2012.

[76] Andreas Lugmayr, Martin Danelljan, Luc Van Gool, and Radu Timofte. SRFlow: Learning the Super-Resolution Space with Normalizing Flow. *arXiv:2006.14200 [cs, eess]*, July 2020.

[77] Dwarikanath Mahapatra and Behzad Bozorgtabar. Retinal Vasculature Segmentation Using Local Saliency Maps and Generative Adversarial Networks For Image Super Resolution. *arXiv:1710.04783 [cs]*, May 2018.

[78] 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.

[79] José Martins, Jaime S. Cardoso, and Filipe Soares. Offline computer-aided diagnosis for Glaucoma detection using fundus images targeted at mobile devices. *Computer Methods and Programs in Biomedicine*, 192:105341, August 2020.

[80] Tim Miller. Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence*, 267:1–38, February 2019.

[81] Yao Ming, Panpan Xu, Huamin Qu, and Liu Ren. Interpretable and Steerable Sequence Learning via Prototypes. *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 903–913, July 2019.

[82] Delaram Mirzania, Atalie C Thompson, and Kelly W Muir. Applications of deep learning in detection of glaucoma: A systematic review. *European Journal of Ophthalmology*, page 1120672120977346, December 2020.

[83] Anirban Mitra, Priya Shankar Banerjee, Sudipta Roy, Somasis Roy, and Sanjit Kumar Setua. The region of interest localization for glaucoma analysis from retinal fundus image using deep learning. *Computer Methods and Programs in Biomedicine*, 165:25–35, 2018.

[84] 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.

[85] Ji Eun Oh, Hee Kyung Yang, Kwang Gi Kim, and Jeong Min Hwang. Automatic computer-aided diagnosis of retinal nerve fiber layer defects using fundus photographs in optic neuropathy. *Investigative Ophthalmology and Visual Science*, 56(5):2872–2879, 2015.

[86] Sejong Oh, Yuli Park, Kyong Jin Cho, and Seong Jae Kim. Explainable Machine Learning Model for Glaucoma Diagnosis and Its Interpretation. *Diagnostics*, 11(3):510, March 2021.

[87] José Ignacio Orlando, Huazhu Fu, João Barbossa Breda, Karel van Keer, Deepti R. Bathula, Andrés Diaz-Pinto, Ruogu Fang, Pheng-Ann Heng, Jeyoung Kim, JoonHo Lee, Joonseok Lee, Xiaoxiao Li, Peng Liu, Shuai Lu, Balamurali Murugesan, Valery Naranjo, Sai Samarth R. Phaye, Sharath M. Shankaranarayana, Apoorva Sikka, Jaemin Son, Anton van den Hengel, Shujun Wang, Junyan Wu, Zifeng Wu, Guanghui Xu, Yongli Xu, Pengshuai Yin, Fei Li, Xiulan Zhang, Yanwu Xu, Xiulan Zhang, and Hrvoje Bogunović. REFUGE Challenge: A Unified Framework for Evaluating Automated Methods for Glaucoma Assessment from Fundus Photographs. *Medical Image Analysis*, 59:101570, January 2020.

[88] Z. Pan, W. Yu, X. Yi, A. Khan, F. Yuan, and Y. Zheng. Recent Progress on Generative Adversarial Networks (GANs): A Survey. *IEEE Access*, 7:36322–36333, 2019.

[89] Taesung Park, Alexei A. Efros, Richard Zhang, and Jun-Yan Zhu. Contrastive Learning for Unpaired Image-to-Image Translation. In Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm, editors, *Computer Vision – ECCV 2020*, Lecture Notes in Computer Science, pages 319–345, Cham, 2020. Springer International Publishing.

[90] D. Povey, L. Burget, M. Agarwal, P. Akyazi, K. Feng, A. Ghoshal, O. Glembek, N. K. Goel, M. Karafiát, A. Rastrow, R. C. Rose, P. Schwarz, and S. Thomas. Subspace Gaussian Mixture Models for speech recognition. In *2010 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 4330–4333, March 2010.

[91] Haonan Qiu, Chaowei Xiao, Lei Yang, Xinchen Yan, Honglak Lee, and Bo Li. SemanticAdv: Generating Adversarial Examples via Attribute-conditional Image Editing. *arXiv:1906.07927 [cs, eess]*, December 2019.

[92] 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.

[93] Alec Radford, Luke Metz, and Soumith Chintala. Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks. *arXiv:1511.06434 [cs]*, January 2016.

[94] 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.

[95] An Ran Ran, Clement C. Tham, Poemen P. Chan, Ching-Yu Cheng, Yih-Chung Tham, Tyler Hyungtaek Rim, and Carol Y. Cheung. Deep learning in glaucoma with optical coherence tomography: A review. *Eye*, 35(1):188–201, January 2021.

[96] Danilo Jimenez Rezende and Shakir Mohamed. Variational Inference with Normalizing Flows. *arXiv:1505.05770 [cs, stat]*, June 2016.

[97] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-Net: Convolutional Networks for Biomedical Image Segmentation. *arXiv:1505.04597 [cs]*, May 2015.

[98] Shalinder Sabherwal, Denny John, Suneeta Dubey, Saptarshi Mukherjee, Geetha R. Menon, and Atanu Majumdar. Cost-effectiveness of glaucoma screening in cataract camps versus opportunistic and passive screening in urban India: A study protocol. *F1000Research*, 8, April 2019.

[99] Ruslan Salakhutdinov, Andriy Mnih, and Geoffrey Hinton. Restricted Boltzmann machines for collaborative filtering. In *Proceedings of the 24th International Conference on Machine Learning*, ICML '07, pages 791–798, New York, NY, USA, June 2007. Association for Computing Machinery.

[100] Kathryn Schutte, Olivier Moindrot, Paul Hérent, Jean-Baptiste Schiratti, and Simon Jégou. Using StyleGAN for Visual Interpretability of Deep Learning Models on Medical Images. *arXiv:2101.07563 [cs, eess]*, January 2021.

[101] Clinical Sciences. The ISNT Rule and Differentiation of Normal From Glaucomatous Eyes. 124, 2014.

[102] Ramprasaath R. Selvaraju, Abhishek Das, Ramakrishna Vedantam, Michael Cogswell, Devi Parikh, and Dhruv Batra. Grad-CAM: Why did you say that? *arXiv:1611.07450 [cs, stat]*, January 2017.

[103] Sourya Sengupta, Amitojdeep Singh, Henry A. Leopold, Tanmay Gulati, and Vasudevan Lakshminarayanan. Application of Deep Learning in Fundus Image Processing for Ophthalmic Diagnosis – A Review. *Artificial Intelligence in Medicine*, 102:101758, January 2020.

[104] Anindita Septiarini, Agus Harjoko, Reza Pulungan, and Retno Ekantini. Automated Detection of Retinal Nerve Fiber Layer by Texture-Based Analysis for Glaucoma Evaluation. *Healthcare Informatics Research*, 24(4):335–345, October 2018.

[105] Sharath M. Shankaranarayana, Keerthi Ram, Kaushik Mitra, and Mohanasankar Sivaprakasam. Joint Optic Disc and Cup Segmentation Using Fully Convolutional and Adversarial Networks. In M. Jorge Cardoso, Tal Arbel, Andrew Melbourne, Hrvoje Bogunovic, Pim Moeskops, Xinjian Chen, Ernst Schwartz, Mona Garvin, Emma Robinson, Emanuele Trucco, Michael Ebner, Yanwu Xu, Antonios Makropoulos, Adrien Desjardin, and Tom Vercauteren, editors, *Fetal, Infant and Ophthalmic Medical Image Analysis*, Lecture Notes in Computer Science, pages 168–176, Cham, 2017. Springer International Publishing.

[106] Sanjivani Shantaiya, Shruti Gorasia, and Rida Anwar. Early Detection of Glaucoma Using Retinal Fundus Images. (6):1525–1528, 2016.

[107] Naoto Shibata, Masaki Tanito, Keita Mitsuhashi, Yuri Fujino, Masato Matsuura, Hiroshi Murata, and Ryo Asaoka. Development of a deep residual learning algorithm to screen for glaucoma from fundus photography. *Scientific Reports*, 8(1):14665, October 2018.

[108] 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.

[109] Amitojdeep Singh, Sourya Sengupta, and Vasudevan Lakshminarayanan. Explainable Deep Learning Models in Medical Image Analysis. *Journal of Imaging*, 6(6):52, June 2020.

[110] 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.

[111] Vivek Kumar Singh, Hatem Rashwan, Farhan Akram, Nidhi Pandey, Md Mostaf Kamal Sarker, Adel Saleh, Saddam Abdulwahab, Najlaa Maaroof, Santiago Romani, and Domenec Puig. Retinal Optic Disc Segmentation using Conditional Generative Adversarial Network. *arXiv:1806.03905 [cs]*, June 2018.

[112] Jayanthi Sivaswamy, Subbaiah Krishnadas, Arunava Chakravarty, Datt Gopal, Gopal Joshi, Ujjwal, and Tabish Syed. A Comprehensive Retinal Image Dataset for the Assessment of Glaucoma from the Optic Nerve Head Analysis. *JSM Biomedical imaging data papers*, April 2015.

[113] Daniel Smilkov, Nikhil Thorat, Been Kim, Fernanda Viégas, and Martin Wattenberg. SmoothGrad: Removing noise by adding noise. *arXiv:1706.03825 [cs, stat]*, June 2017.

[114] European Glaucoma Society. Terminology and guidelines for glaucoma, 2021.

[115] Jaemin Son, Sang Jun Park, and Kyu-Hwan Jung. Retinal Vessel Segmentation in Fundoscopic Images with Generative Adversarial Networks. *arXiv:1706.09318 [cs]*, June 2017.

[116] 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 & Laser Technology*, 110:87–98, February 2019.

[117] Gregor Stiglic, Primoz Kocbek, Nino Fijacko, Marinka Zitnik, Katrien Verbert, and Leona Cilar. Interpretability of machine learning-based prediction models in healthcare. *WIREs Data Mining and Knowledge Discovery*, 10(5):e1379, 2020.

[118] Jianlin Su and Guang Wu. F-VAEs: Improve VAEs with Conditional Flows. *arXiv:1809.05861 [cs, stat]*, September 2018.

[119] Y-W Teh. Rate-coded Restricted Boltzmann Machines for Face Recognition. page 7.

[120] 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.

[121] Yih Chung Tham, Xiang Li, Tien Y. Wong, Harry A. Quigley, Tin Aung, and Ching Yu Cheng. Global prevalence of glaucoma and projections of glaucoma burden through 2040: A systematic review and meta-analysis. *Ophthalmology*, 121(11):2081–2090, 2014.

[122] Erico Tjoa and Cuntai Guan. A Survey on Explainable Artificial Intelligence (XAI): Towards Medical XAI. *IEEE Transactions on Neural Networks and Learning Systems*, pages 1–21, 2020.

[123] Too Chen, Chao Huang, E. Chang, and Jingehan Wang. Automatic accent identification using Gaussian mixture models. In *IEEE Workshop on Automatic Speech Recognition and Understanding, 2001. ASRU '01.*, pages 343–346, December 2001.

[124] 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.

[125] 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.

[126] Shujun Wang, Lequan Yu, and Pheng-Ann Heng. Optic Disc and Cup Segmentation with Output Space Domain Adaptation. page 8.

[127] Brandon J. Wong, Benjamin Y. Xu, and Mingguang He. Examination of the Optic Nerve in Glaucoma. In Rohit Varma, Benjamin Y. Xu, Grace M. Richter, and Alena Reznik, editors, *Advances in Ocular Imaging in Glaucoma*, Essentials in Ophthalmology, pages 59–69. Springer International Publishing, Cham, 2020.

[128] Xiangyu Chen, Yanwu Xu, Damon Wing Kee Wong, Tien Yin Wong, and Jiang Liu. Glaucoma detection based on deep convolutional neural network. *Conference proceedings of IEEE engineering and medical biological society*, pages 715–718, 2015.

[129] Yanwu Xu, Lixin Duan, Stephen Lin, Xiangyu Chen, Damon Wing Kee Wong, Tien Yin Wong, and Jiang Liu. Optic Cup Segmentation for Glaucoma Detection Using Low-Rank Superpixel Representation. In Polina Golland, Nobuhiko Hata, Christian Barillot, Joachim Hornegger, and Robert Howe, editors, *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2014*, Lecture Notes in Computer Science, pages 788–795, Cham, 2014. Springer International Publishing.

[130] 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. In *2011 Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, pages 2626–2629, August 2011.

[131] Pengshuai Yin, Guanghui Xu, Jingwen Wang, Yuguang Yan, Qingyao Wu, and Mingkui Tan. Optic Disc and Cup Segmentation using Ensemble Deep Neural Networks. page 6.

[132] Zekuan Yu, Qing Xiang, Jiahao Meng, Caixia Kou, Qiushi Ren, and Yanye Lu. Retinal image synthesis from multiple-landmarks input with generative adversarial networks. *BioMedical Engineering OnLine*, 18(1):62, May 2019.

[133] Han Zhang, Tao Xu, Hongsheng Li, Shaoting Zhang, Xiaogang Wang, Xiaolei Huang, and Dimitris Metaxas. StackGAN: Text to Photo-realistic Image Synthesis with Stacked Generative Adversarial Networks. *arXiv:1612.03242 [cs, stat]*, August 2017.

[134] Rui Zhang, Tomas Pfister, and Jia Li. Harmonic Unpaired Image-to-image Translation. *arXiv:1902.09727 [cs]*, February 2019.

[135] X. Zhang and J. Wu. Deep Belief Networks Based Voice Activity Detection. *IEEE Transactions on Audio, Speech, and Language Processing*, 21(4):697–710, April 2013.

[136] 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. In *2010 Annual International Conference of the IEEE Engineering in Medicine and Biology*, pages 3065–3068, August 2010.

[137] H. Zhao, H. Li, S. Maurer-Stroh, Y. Guo, Q. Deng, and L. Cheng. Supervised Segmentation of Un-Annotated Retinal Fundus Images by Synthesis. *IEEE Transactions on Medical Imaging*, 38(1):46–56, January 2019.

[138] He Zhao, Huiqi Li, Sebastian Maurer-Stroh, and Li Cheng. Synthesizing retinal and neuronal images with generative adversarial nets. *Medical Image Analysis*, 49:14–26, October 2018.

[139] 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.

[140] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A. Efros. Unpaired Image-to-Image Translation using Cycle-Consistent Adversarial Networks. *arXiv:1703.10593 [cs]*, August 2020.

[141] Zhuo Zhang, Jiang Liu, Fengshou Yin, Beng-Hai Lee, Damon Wing Kee Wong, and Kyung Rim Sung. ACHIKO-K: Database of fundus images from glaucoma patients. In *2013 IEEE 8th Conference on Industrial Electronics and Applications (ICIEA)*, pages 228–231, Melbourne, VIC, June 2013. IEEE.

[142] Karel Zuiderveld. Contrast limited adaptive histogram equalization. In *Graphics Gems IV*, pages 474–485. Academic Press Professional, Inc., USA, August 1994.