FACULDADE DE ENGENHARIA DA UNIVERSIDADE DO PORTO

Characterization and Assessment of Vessel Tortuosity in Retinal Images

Hugo Alexandre Azedo Sousa

Master in Bioengineering

Supervisor: Ana Maria Mendonça

Co-supervisor: Tânia Fernandes de Melo

October 31, 2024

Characterization and Assessment of Vessel Tortuosity in Retinal Images

Hugo Alexandre Azedo Sousa

Master in Bioengineering

October 31, 2024

Resumo

A tortuosidade dos vasos sanguíneos da retina é um biomarcador amplamente conhecido por auxiliar o diagnóstico e a monitorização de várias condições oculares e sistémicas. Esta dissertação investiga a utilização de modelos de aprendizagem profunda para a classificação automática da tortuosidade dos vasos da retina, avaliando simultaneamente o impacto de diferentes técnicas de pré-processamento de imagens. São explorados quatro métodos de pré-processamento diferentes: recorte, 'padding', uniformização do fundo e aplicação de *Contrast Limited Adaptive Histogram Equalization* (CLAHE). Os modelos de aprendizagem profunda baseados na arquitetura Xception foram treinados e validados usando um conjunto de dados que consiste em imagens de fundo da retina anotadas manualmente. Os modelos gerados foram ainda testados num conjunto de dados externo e em imagens de OCTA. A avaliação do desempenho foi realizada através de um processo de validação cruzada com cinco *folds* e mapas de saliência gerados para interpretar o foco do modelo durante a classificação. Os resultados indicaram uma variabilidade considerável entre *folds*, sem que uma única técnica de pré-processamento superasse consistentemente as restantes. Os mapas de saliência mostraram que o 'padding' melhorou a visibilidade dos vasos sanguíneos, enquanto o CLAHE resultou num foco mais disperso nas regiões da imagem. Além disso, o tamanho reduzido do dataset e os *ground truth* inconsistentes, causados por discrepâncias inerentes às anotações de especialistas, limitaram o desempenho geral e a generalização dos modelos. Os resultados destacam a necessidade de protocolos padronizados de aquisição de imagens e conjuntos mais diversificados para melhorar a precisão e a fiabilidade dos modelos em aplicações clínicas. O trabalho futuro deve centrar-se na melhoria da qualidade do conjunto de dados, no refinamento das técnicas de pré-processamento e na investigação de arquiteturas alternativas para otimizar, ainda mais, a classificação da tortuosidade.

Palavras-Chave: Imagens Coloridas de Fundo, Tortuosidade dos Vasos da Retina, Aprendizagem Profunda

Abstract

Retinal vessel tortuosity is a biomarker widely known to aid the diagnosis and monitoring of various ocular and systemic conditions. This dissertation investigates the use of deep learning models for automatic classification of retinal vessel tortuosity, while simultaneously assessing the impact of different image preprocessing techniques. Four different preprocessing methods are explored: cropping, padding, uniformization of the background, and aplication of *Contrast Limited Adaptive Histogram Equalization* (CLAHE). Deep learning models based on the Xception architecture were trained and validated using a dataset consisting of manually annotated retinal fundus images and further tested on an external dataset and in OCTA images. Performance evaluation was conducted through a five-fold cross-validation process, with saliency maps generated to interpret the model's focus during classification. The results indicated considerable variability across folds, with no single preprocessing technique consistently outperforming the others. Padding improved the blood vessels visibility, while CLAHE resulted in more scattered focus on image regions. Furthermore, the small dataset and inconsistent ground truth labels, caused by inherent discrepancies in expert annotations, limited the overall performance and generalizability of the models. The findings highlight the need for standardized image acquisition protocols and larger, more diverse datasets to improve model accuracy and reliability in clinical applications. Future work should focus on enhancing dataset quality, refining preprocessing techniques, and investigating alternative architectures to further optimize tortuosity classification.

Keywords: Color Fundus Imaging, Retinal Vascular Tortuosity, Deep Learning

Acknowledgments

I would like to extend my heartfelt gratitude to Professor Ana Maria Mendonça and Tânia Melo for their invaluable support throughout the development of this work.

A special thanks to Diana, Margarida, and Bárbara, whose unwavering presence over the past five years has been a source of immense strength and encouragement.

I am also deeply grateful to my family for their continuous support and to Henrique, whose uplifting presence has been instrumental in my journey. Your contributions and encouragement have made this endeavor possible and are sincerely appreciated.

Hugo

Contents

List of Figures

List of Tables

Abbreviations

AD	Alzheimer's Disease
AI	Artificial Intelligence
AMD	Age-related Macular Degeneration
AUC	Area Under the Curve
AV	Artery-Vein
BA	Bilinear Attention
BACC	Balanced Accuracy
CLAHE	Contrast Limited Adaptive Histogram Equalization
CCM	Corneal Confocal Microscopy
CNS	Central Nervous System
CNNs	Convolutional Neural Networks
DNNs	Deep Neural Networks
EIARG	Eye Images Analysis Research Group
FFA	Fundus Fluorescein Angiography
FOV	Field of View
FRAT	Familial Retinal Arteriolar Tortuosity
ICGA	Indocyanine Green Angiography
JSEIC	Joint Shantou International Eye Centre
OCT	Optical Coherence Tomography
OCTA	Optical Coherence Tomography Angiography
ReLU	Rectified Linear Unit
ROC	Receiver Operating Characteristic
ROI	Region of Interest
RPE	Retinal Pigment Epithelium
VTI	Vessel Tortuosity Index

Chapter 1

Introduction

This chapter details fundamental concepts regarding the eye, particularly focusing on the retina. It also examines the clinical importance of retinal tortuosity, the diseases associated with it, and the imaging techniques currently used in both diagnosis and research. Additionally, the chapter explains the motivation for this study and defines its main objectives.

1.1 Context

1.1.1 The Eye

The human eye is a vital sensory organ, essential for perceiving and interacting with the world. It consists of three distinct layers, illustrated in Figure 1.1.

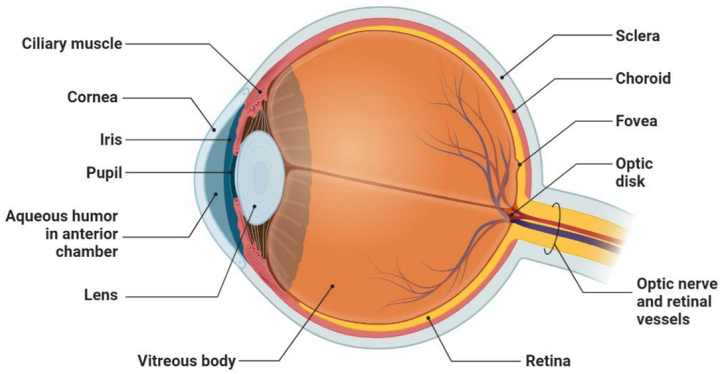


Figure 1.1: Main components of the eye.

The outer layer includes the cornea and sclera, while the middle layer comprises the iris, ciliary body, and choroid. The innermost layer of the eye is the retina. The cornea, located at the front of the eye, protects against infections and refracts light to the lens and retina. The limbus connects the cornea and sclera, providing structural stability. The iris controls the pupil size, regulating the amount of light reaching the retina. The ciliary body adjusts the lens's shape and power, while the choroid supplies oxygen and nutrients to the retina's outer layers [Navarro (2009)].

The retina is a complex structure of neurons that capture and process light, converting it into electrical signals sent to the brain, via the optic nerve. It contains photoreceptors, which have some of the fastest metabolic rates in the body, underscoring the retina's crucial role in visual processing. As part of the central nervous system (CNS), the retina is protected by the specialized blood-retinal barriers. The inner barrier, formed by tight junctions between endothelial cells of retinal blood vessels, and the outer barrier, made of tight junctions between retinal pigment epithelium (RPE) cells. These barriers, illustrated in Figure 1.2, maintain the delicate environment necessary for optimal retinal function and visual acuity, by safeguarding the retina from harmful substances [UCL (2020)].

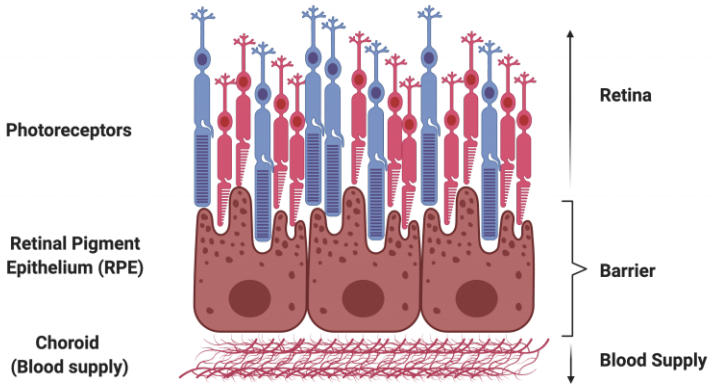


Figure 1.2: The blood-retinal barrier is formed between the retina and the choroid [UCL (2020)].

There are some relevant regions within the retina, namely the fovea, the macula and the optic disc, which Figure 1.3 depicts as observed directly by a physician. The fovea is a small indentation within the macula and is essential for sharp central vision. Measuring about 1.5 mm in diameter, it is densely packed with cone photoreceptors. The macula, encompassing the fovea, is vital for high-resolution vision and color perception, filtering short wavelengths to protect against bright light damage. Damage to the macula, for instance from age-related macular degeneration, can lead to significant vision loss. The optic disc is a bright spot with size of about 3 mm, that serves as the exit point for the optic nerve and blood vessels. This characteristic generates a natural 'blind spot' due to the absence of light-sensitive cells. The retinal circulation faces unique challenges due to the eye's small opening at the optic nerve head, where different pressures can impact blood flow and nutrient delivery. Maintaining precise pressure regulation is critical for retinal health, highlighting the intricate balance required for the eye's complex visual processes [Semerád and Drahanský (2020)].

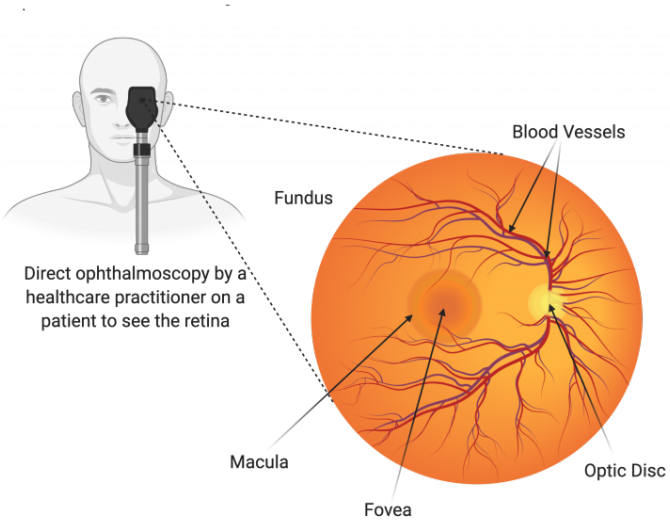


Figure 1.3: Schematic visualization of the elements of the retina as seen directly from an ophthalmologist: macula, fovea, optic disc and the blood vessels [UCL (2020)].

1.1.2 Imaging Techniques for Retinal Visualization

The eye offers a unique glimpse into the vascular network *in vivo* without the need for invasive procedures, unlike most parts of the human body. Since the retina is transparent, the blood vessels that supply it can be studied directly by physicians. The recent advancements in retinal imaging have significantly improved the ophthalmology field, by advancing the ability to diagnose and treat retinal conditions. High-resolution fundus cameras and Optical Coherence Tomography (OCT) have provided deeper insights into the retina's structure and function, ultimately improving patient care, even in regions where expensive devices are not available for all.

- **Color Fundus** — Fundus imaging creates a 2D image of the eye's internal 3D surface using a low-power microscope and a camera. The patient stabilizes their head on a chin and forehead rest, while the operator aligns and focuses the camera to capture the image. This results in an upright, magnified view of the fundus, typically with angles of 30°, 45°, or 60° and a magnification of ×2.5. Pupil dilation with mydriatic drops can improve image quality by widening the field of view but is not always necessary. White light is generally used to illuminate the retina for full-color images, although red-free images can be created by filtering out red light, offering better contrast of retinal and choroidal blood vessels [MacGillivray et al. (2014)].

- **Fundus Fluorescein Angiography (FFA)** — With this technique the subject receives an intravenous injection of a sodium fluorescent dye, while the retina is illuminated at an excitation wavelength (465–490nm). This enables a high-contrast image of the blood vessels and also highlights areas of damage where the dye escapes into the surrounding tissue. Flow dynamics and associated diseases are shown by taking a timed sequence of photographs showing the movement of the dye into the arteries. FFA is a moderately invasive procedure

which has several important contraindications, such as dye allergy and pregnancy [Ruia and Tripathy (2024)].

- **Indocyanine Green Angiography (ICGA)** — Visualizes the choroidal circulation and detects any abnormalities. While FFA is effective for detailed imaging of retinal circulation, it has limitations in capturing the choroidal circulation due to poor fluorescence transmission through RPE cells, media opacities, and retinal exudates [Muraleedharan and Tripathy (2024)].

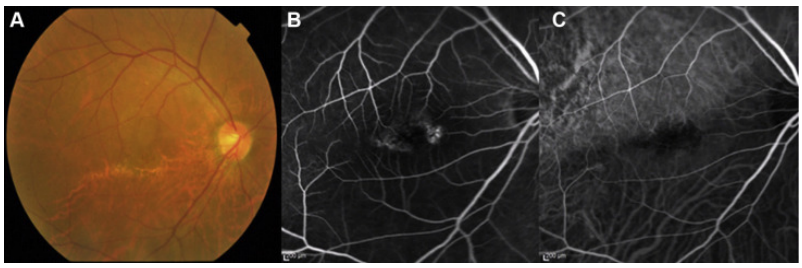


Figure 1.4: A serous retinal detachment caused by tilted disc syndrome. (A) The color fundus image reveals retinal pigment epithelium atrophy spreading from the disc to the lower posterior pole. (B) Fluorescein angiography reveals hyperfluorescence and fluorescein leaking surrounding the fovea. (C) Indocyanine green angiography reveals hypofluorescence around the fovea [Kubota et al. (2019)].

- **Optical Coherence Tomography Angiography (OCTA)** — In OCT, infrared light is divided and focused on an internal mirror and the target tissue. The ensuing phase differences provide detailed pictures. OCTA extends this by generating en-face and cross-sectional pictures using flow data. En-face OCTA provides a top-down view of the arterial layers, whereas cross-sectional B-scans provide comprehensive structure and flow information, which aids in the detection of vascular anomalies. OCTA is advantageous because it is more time-efficient than earlier approaches, taking around 6 seconds per image, and is technically simpler to execute [Javed et al. (2023)].

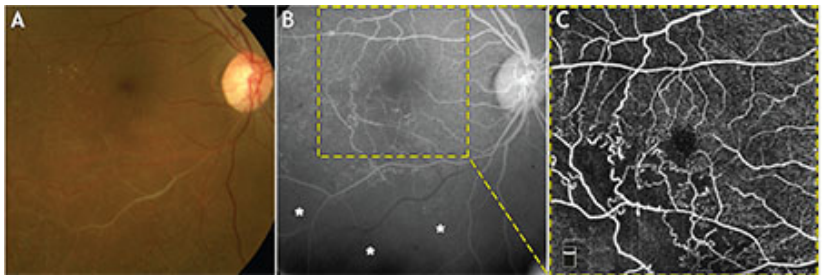


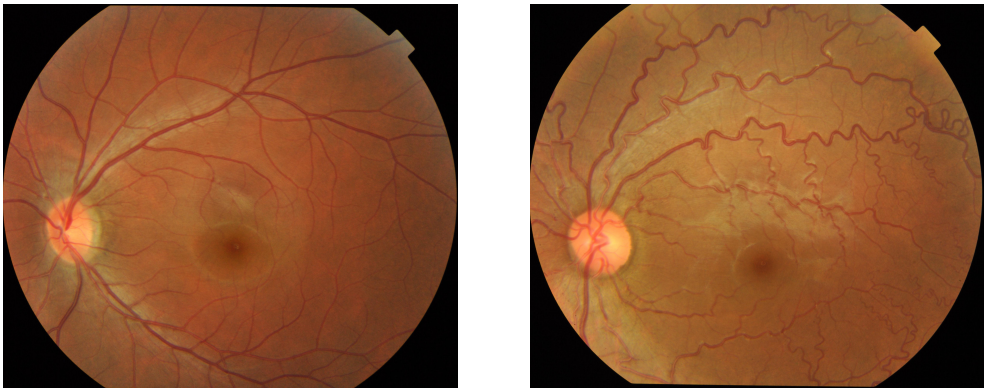
Figure 1.5: (A) Color fundus photography shows non-perfused vessels and laser scars in the inferior temporal retina. (B) Fluorescein angiography identifies the 6x6 mm area captured by OCTA, with white asterisks marking ischemic regions outside the small field of view. (C) En face OCTA of the superficial retinal plexus reveals a non-perfused area temporal-inferior to the fovea [Novais (2017)].

Using a variety of imaging modalities is critical for obtaining a clearer view of the retinal and choroidal vasculature. Each strategy has advantages and disadvantages, and depending just on one might yield insufficient results. A combined approach not only enhances the understanding of the disease and its course, but it also helps measure how effectively medicines work, making it vital for optimal patient care [Javed et al. (2023)].

The acquistion method is an aspect that is often disregarded between studies involving the retina. Bias can arise from differences in image acquisition methods and camera angle settings, and make comparisons across protocols invalid. Using a larger grid for tortuosity evaluation tends to capture primarily parent vessels, omitting crucial details about smaller arterioles, venules, and the capillary network. Larger vessels might endure pressure changes longer than smaller ones, but once the capillary network is compromised, they can rapidly deteriorate [Abràmoff et al. (2010)].

1.1.3 Retinal Tortuosity and Related Pathologies

The retina's blood arteries are arranged in a traditional pattern that optimizes blood flow and oxygen delivery to the retinal tissues. The four quadrants or lobes of the tissue are reached by the four main artery branches and the veins that correspond to them. First-order branches are typically more arched when they are found in the temporal hemisphere, and they are typically slightly sinuous. This subtle arching and sinuosity allows the vascular network to remain structurally intact without undue strain while also accommodating the eye's flexibility and movement. Particularly, the temporal arteries' arching guarantees effective blood flow to the retina's periphery, which is essential for peripheral vision [Han (2012); Abràmoff et al. (2010); Vilela et al. (2021)].



(a) Normal (b) Tortuous

Figure 1.6: Side by side comparison of normal retina and one with tortuosity.

A slight degree of tortuosity in retinal vessels is normal and varies with each vessel. However, excessive tortuosity is abnormal and compromises the integrity of the vascular system. When vessels become overly twisted, the smooth flow of blood is disrupted, reducing the efficiency of oxygen and nutrient delivery to retinal tissues. The heightened tortuosity can jeopardize retinal

health and therefore is associated with multiple retinal diseases. Figure 1.6 clearly showcases retinal fundus images with and without tortuosity.

A tortuous vessel, being longer than a straight vessel over the same distance, enhances metabolic input to tissues. The changes detected during retina assessment serve as biomarkers for disease severity, longevity, prognosis, and therapeutic response. Other key indicators include vascular caliber, arteriovenous crossing points, occlusions, embolisms, and capillary permeability [Vilela et al. (2021)]. Tortuosity can assume different forms and be categorized as follows: curving, curling, angulation, twisting and looping, as shown in Figure 1.7 [Han (2012)].

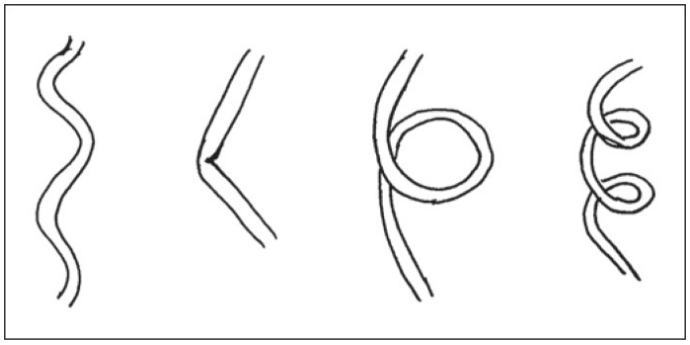


Figure 1.7: Various phenotypes of tortuous vessels. Left to right: curving, angulation, looping and spiral twisting.

One disorder that affects the retinal vasculature is hypertension, which thickens the arteriolar walls through autonomic processes that shield the capillary bed from elevated systemic blood pressure. Hence, this thickening and heightened vascular tortuosity are typical. Uncontrolled hypertension can cause consequences such retinal edema, cotton wool spots, and hemorrhages because high pressures are greater than what retinal arterioles can regulate [Henderson et al. (2011)].

Retinal vascular integrity is also becoming more and more associated with Alzheimer's disease (AD). Although CNS plaques and tangles are the traditional signs of AD, measuring vessel dropout and keeping an eye on changes in blood flow, retinal thickness, and flow regulation are also becoming more and more popular [Pead et al. (2023); Liao et al. (2018); Dumitrascu and Qureshi (2018); MacCormick et al. (2015)].

High oxygen levels in neonatal care, while necessary for survival, interfere with the proper development of the blood vessels in the peripheral retina, leading to retinopathy of prematurity. The portions of the retina that do not receive enough blood flow become hypoxic as the baby grows and his oxygen levels stabilize. At the border between the vascularized and non-vascularized retina, this oxygen deficiency promotes the formation of new blood vessels, which frequently results in severe retinal detachment and scarring [Wigdahl et al. (2017)].

In sickle cell anemia, tortuosity is a frequent and early sign. Evaluations using OCTA images revealed that it mainly affects the perifoveal capillaries and parafoveal areas.[Cano et al. (2020)].

Research has also linked increased tortuosity to several other pathologies and genetic disorders, including: diabetic retinopathy [Sasongko et al. (2011)], gestational diabetes mellitus [Li et al. (2017); Sasongko et al. (2012)], familial retinal arteriolar tortuosity (FRAT) [Saraf et al.

(2019)], Fabry Disease [Atiskova et al. (2021)] and facioscapulohumeral muscular dystrophy [Goselink et al. (2019)]. Figure 1.8 presents retinal images of patients with pathologies associated with retinal tortuosity.

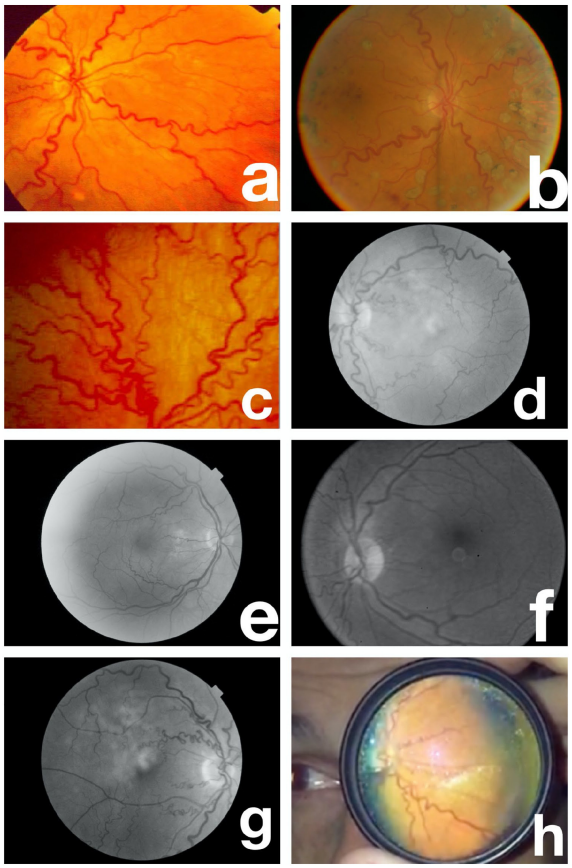


Figure 1.8: Various conditions associated with retinal vascular tortuosity: Images (a, b) illustrate the long-term progression and late-stage development of diabetic retinopathy. Image (c) shows the vascular pattern characteristic of plus disease in retinopathy of prematurity. Image (d) represents Fabry disease, while image (e) displays congenital cyanotic heart disease and associated small-vessel changes. Image (f) highlights systemic arterial hypertension. Image (g) depicts familial retinal arterial tortuosity (FRAT) in a young boy, and image (h) shows the same condition in his parent [Vilela et al. (2021)].

1.2 Motivation

The excessive twisting of blood vessels in the retina, designed as retinal vascular tortuosity, is an important indicator of various health conditions. These changes in the retinal vessels often mirror issues in the microvasculature of other organs and may even signal systemic vascular problems. Therefore, understanding and accurately measuring this tortuosity is vital for diagnosing broader health issues.

Despite its potential, using retinal vascular tortuosity as a clinical biomarker faces some barriers, mainly because achieving objective and reliable assessments is no easy feat. Manual evaluation from color fundus images is labor-intensive and time-consuming, requiring the identification and contextual assessment of tortuous vessels. Clinicians might interpret slight twisting in some vessels as normal, while minor tortuosity in others may be overvalued. These issues highlight the need for automated image analysis algorithms to provide more accurate and less subjective assessments. Deep learning models that receive retinal images with minimal processing as input and output tortuosity predictions can be the way forward to achieve that purpose.

1.3 Research Objectives

The main objective of this research is to evaluate the effectiveness of the Xception deep learning architecture for the classification of the retinal vascular tortuosity in three levels: 'No Tortuosity, 'Some Tortuosity', and 'High Tortuosity'. This involves training and testing the model on the available labeled retinal image datasets. A key aim is to assess how distinct preprocessing techniques impact the model's performance by applying different methods to retinal images and evaluating their influence on classification accuracy and reliability. The research also explores the potential of transferring the Xception model, trained on fundus images, to classify retinal vascular tortuosity in OCTA images.

Ultimately, the goal is to identify the optimal combination of deep learning architecture and preprocessing method to achieve the highest accuracy in tortuosity classification, enhancing automated retinal vessel analysis.

1.4 Dissertation Structure

This document begins with a comprehensive review of the state of the art in retinal vascular tortuosity analysis, discussing classic tools, hybrid techniques that integrate deep learning for segmentation, and the use of deep learning for direct tortuosity analysis (Chapter 2). Chapter 3 details the datasets used, the preprocessing steps and the deep learning architecture employed. It also covers implementation details, like the training procedure, data augmentation, the hyperparameter settings and the validation process with performance metrics. The results chapter (Chapter 4) showcases the outcomes of tortuosity grading using the developed algorithm and includes visualizations and case studies to demonstrate effectiveness. It interprets the findings, evaluates the strengths and limitations of the deep learning approach, and suggests areas for future research. The dissertation concludes in Chapter 5 with the key findings of this study and remarks regarding future paths for tortuosity evaluation using deep learning.

Chapter 2

State of the Art

This chapter starts with a brief introduction to deep learning applications in ophthalmology. Additionally, it explores the evolution of tortuosity assessment, covering its past, present, and future. First, classic metrics widely applied in research will be presented. Following that, some hybrid methods that partially use deep learning will be showcased. Lastly, sole deep learning applied for classification of retinal tortuosity is briefly discussed.

2.1 Deep Learning in Ophthalmology

The adoption of deep learning in therapeutic is wrapped with a number of important challenges. Two major concerns are the security and privacy of sensitive patient data, as large datasets are required to train these algorithms, yet safeguarding this information remains difficult. Another major hurdle is model generalizability, because ensuring that deep learning models perform well across diverse demographics and imaging devices is crucial. Ongoing efforts focus on improving the robustness and adaptability of these models to ensure reliable performance in varied clinical settings. Regulatory compliance is also critical. Deep learning models must adhere to strict legal standards and ethical guidelines before they can be safely integrated into clinical practice. Innovative solutions are currently being explored to address these issues. For example, federated learning allows models to be trained on distributed datasets without the need for direct data sharing, thus preserving data privacy. Similarly, blockchain technology offers a more secure way to manage and audit data, which might boost transparency and confidence in Artificial Intelligence (AI) technologies used in healthcare [Moulahi et al. (2023)].

Deep learning has disrupted Ophthalmology by advancing the detection and diagnosis of eye disorders. In particular, convolutional neural networks (CNNs) have demonstrated high accuracy in diagnosing some eye conditions from ophthalmic images. For instance, models trained on extensive retinal image datasets can detect diabetic retinopathy with precision comparable to experienced ophthalmologists, as shown by Gulshan et al. (2016). Similarly, deep learning models have proven effective in diagnosing age-related macular degeneration (AMD) and glaucoma, achieving

diagnostic accuracy on par with human experts [Dong et al. (2021)]. Deep learning is also enhancing predictive analytics, including forecasting disease progression and treatment outcomes. Models can predict visual field changes in glaucoma patients, facilitating early intervention and better disease management [Ashtari-Majlan et al. (2024)].

2.2 Computational Approaches for Tortuosity Evaluation

Understanding tortuosity has traditionally been done using classic computational methods. While these methods are somewhat reliable, they often struggle to fully capture the intricate details of complex networks. In recent years, the integration of machine learning models with traditional metrics has become increasingly common. This blend improves the precision and depth of tortuosity assessments, making it simultaneously less time consuming. Typically, these hybrid methods start with deep learning tools like U-Net to segment and map out the vascular structures. After this, traditional metrics come into play to measure tortuosity. Lastly, still few approaches use pure machine learning without relying on traditional methods to assess tortuosity. Figure 2.1 is a schematic that illustrates the usual pipeline for the classification of a certain eye disorder. The attempt with this work is to disrupt the usual pipeline and skip image segmentation and feature extraction, moving from image processing directly to classification.

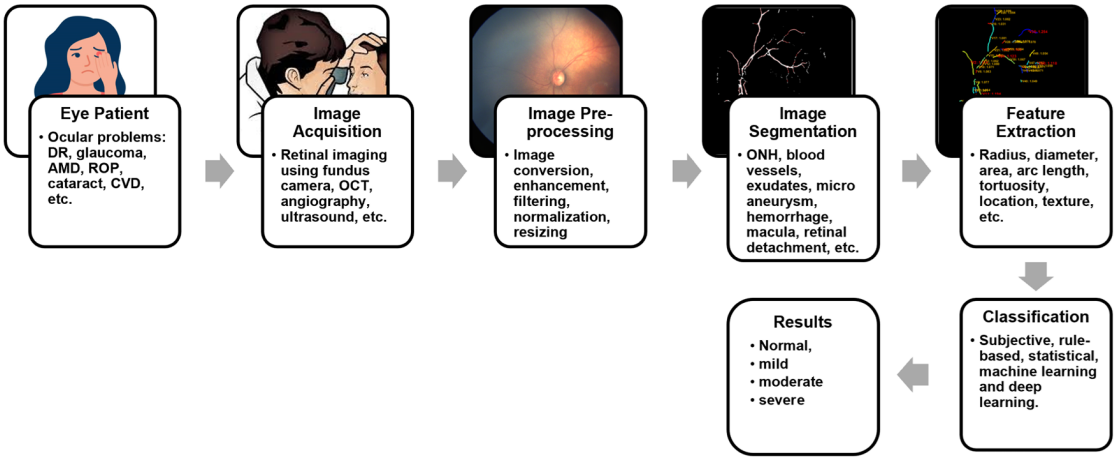


Figure 2.1: Pipeline for grading of an eye disease includes several stages: Image acquisition, Image Preprocessing, Image Segmentation, Feature Extraction, and lastly Classification [Kumar and Paul (2023)].

2.2.1 Classical Computational Approaches

The efforts to objectively quantify the clinicians' intuitive notion of tortuosity are ongoing with several methods being proposed, but so far the metrics are still far from being perfect. Many of these approaches focus on developing quantitative metrics to measure the retinal tortuosity typically based on the mathematical representations of the vessels path. As a result, the tortuosity

of a vessel segment is often defined by its geometric features, such as curvature, inflection points, and the lengths of arcs and chords.

The simplest and most widely used index was first discussed by Smedby et al. (1993) and further expanded by Hart et al. (1999). It computes the tortuosity of a vessel by examining how long the curve is relative to its chord length, as follows:

$$\tau_{\text{Hart}} = \frac{L_c}{L_x} - 1 \tag{2.1}$$

where L_c is the arc length or curve length obtained by counting all the points from the start of the vessel to its end, and L_x is the length of the underlying chord, that is, the Euclidean distance between the two end points of the vessel. A perfectly straight vessel will have a score of 0. Therefore, the greater the tortuosity of the blood vessel, the higher the value of this metric. This dimensionless measure is advantageous for comparisons between studies or eyes. However, the method is flawed and can potentially underestimate the blood vessels true tortuosity: vessels that bend gradually can yield the same result as vessels presenting more frequent turns [Ramos et al. (2018)].

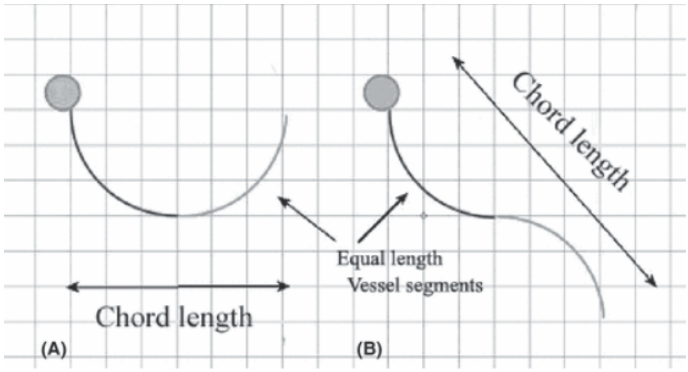


Figure 2.2: Visual explanation of underestimation of tortuosity with metric based on arc and chord length: (A) is intuitively a more tortuous path, however (B) will have a higher tortuosity score if only taken into account the curve and chord lengths [Ramos et al. (2018)].

Another tortuosity metric has been proposed by Grisan et al. (2008). This approach subdivides each vessel into n segments of constant-sign curvature and then combines the evaluation of such segments and their number as follows:

$$\tau_{\text{Grisan}} = \frac{n-1}{n} \frac{1}{L_c} \sum_{i=1}^{n} \left[\frac{L_{csi}}{L_{xsi}} - 1 \right] \tag{2.2}$$

where L_c corresponds to the arc length of the vessel whereas L_{csi} and L_{xsi} represent the arc length and the chord length of each subsegment. Using this method, two vessels with the same geometry and different sizes will have different values for the tortuosity. Although this limitation does not affect the comparison between vessels of similar size, it considerably compromises any attempt to quantify overall vascular tortuosity. When this metric computes retinal fundus global tortuosity,

short vessels with high tortuosity will increase the tortuosity of the entire image. In fact, those short vessels might appear insignificant compared to other vessels in the image.

Trucco et al. (2010) presented a tortuosity measure that only depends on the vessel skeleton curvature. This metric is a generalized version of the curvature-based metrics presented in Hart et al. (1999). It is defined by:

$$\tau_{\text{Trucco}} = \left(\sum_j |k_s(j)|^p \right)^{\frac{1}{p}} \tag{2.3}$$

where p is a strictly positive integer and $k_s(j)$ is the curvature at the j point of the vessel s, defined as follows:

$$k_s(j) = \frac{x'(j)y''(j) - x''(j)y'(j)}{[y'(j)^2 + x'(j)^2]^{3/2}}, \tag{2.4}$$

The curvature measures the variation of the slope of the line tangent to the curve at each point along the segment. A significant difference in slope between the point and its surrounding neighbors implies a high curvature.

Khansari et al. (2017) proposed a tortuosity measurement method designated as the Vessel Tortuosity Index (VTI). This measure is sensitive to small changes in tortuosity, therefore it is appropriate for detecting small tortuosity changes in retinal vessels of OCTA images.

$$VTI = \frac{0.1SD_\theta \cdot N \cdot M \cdot L_A}{L_c} \tag{2.5}$$

where SD_θ is defined as the standard deviation of the absolute angle values between the line tangent to the centerline and a predetermined reference axis, N is the number of the curve's critical points, M is the magnitude of the vessel's centerline, L_A is the arc length, and L_C is the chord length. Changes in the sign of curvature determines the number of inflection points, which can be inaccurate for curves with many small curves.

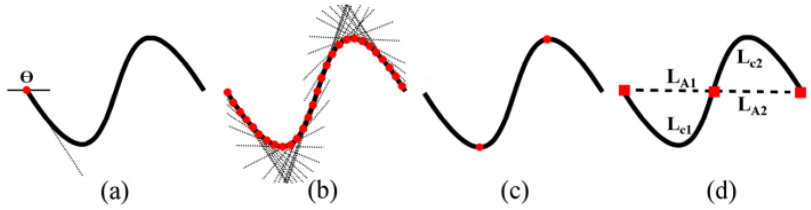


Figure 2.3: (a) Angle between a line tangent to the centerline and a reference axis for the first centerline point. (b) Tangent lines for points along the centerline. SD of angles between each tangent line and the reference axis was computed. (c) Critical points (red circles) were determined based on changes in sign of slope of the tangent lines. (d) Magnitude of curve as ratio of arch length (LA) to the chord length (LC) between pairs of inflection points including centerline end points (red squares) [Khansari et al. (2017)].

Previous methods have proven capable of adequately assessing retinal vascular tortuosity automatically. However, they are constrained by classical image processing techniques, such as the Otsu thresholding to segment the blood vessels and other anatomical structures. This task is particularly challenging in color fundus images due to their photographic nature and the diverse characteristics of retinal anatomy. The variations in capture devices, lighting conditions, and operator expertise can alter the appearance of these images. Additionally, the presence of multiple diseases can affect the appearance of anatomical structures. Vascular tortuosity itself significantly changes the morphology of blood vessels, impacting algorithm performance. Compared to modern CNNs, classical methods typically offer lower performance and are less robust to image acquisition variations and the presence of diseases [Hervella et al. (2024)].

2.2.2 Hybrid Techniques Integrating Deep Learning for Segmentation

Current approaches to assess retinal vascular tortuosity frequently begin with an automatic segmentation of the vascular network through the deep learning algorithms. These methods usually apply U-Nets to extract and categorize the anatomical components from fundus images, such as blood vessels, veins, and arteries. Classic tortuosity indices provide the degree of blood vessel contortion once the vascular network has been subdivided. These indices offer a more uniform way to quantify tortuosity. Those metrics are based on vessel length and chord length ratios, accumulated tangent angle discrepancies, or curvature analysis. Certain techniques adjust widely-used metrics, adding logistic functions or other mathematical models to better represent vessel morphology and curvature inflection points in order to increase accuracy even more. The combination of deep learning for segmentation and traditional tortuosity metrics has shown to improve the precision and reliability of retinal vascular assessments.

Wang et al. (2021) introduced a novel multiple subdivision-based algorithm for quantifying retinal vascular tortuosity, which involves segmenting the retinal vascular network using a multi-path recurrent U-Net and subdividing vessel segments into overlapping sub-segments with continuously increasing chord lengths. The tortuosity index is calculated based on accumulated tangent angle differences, enhanced by a logistic function modeling curvature inflection points. Validation on the RET-TORT dataset yielded high correlation coefficients with expert grading (0.931 for arteries and 0.925 for veins). In a clinical dataset of 104 retinal fundus images from diabetic patients, the algorithm achieved a Spearman's rank correlation coefficient of 0.82 with expert grading and an Area Under the Curve (AUC) of 0.968 in Receiver operating characteristic (ROC) analysis, indicating high accuracy for grading vascular tortuosity .

Khanal and Estrada (2023) developed a fully automated method for measuring the artery-vein (AV) ratio and vascular tortuosity in retinal fundus images using topology extraction and U-Net-based CNNs. The method includes generating a vessel likelihood map, creating a refined topology graph, and employing U-Nets for optic disk extraction and AV classification. The technique achieved a balanced accuracy (BACC) of 0.934, a sensitivity of 0.943, and a specificity of 0.925 on the DRIVE dataset for AV classification. For the INSPIRE-AV dataset, it achieved a BACC of

0.924. The average error in AV ratio measurement on the INSPIRE-AVR dataset was 0.065±0.058 compared to a reference grader.

Hervella et al. (2024) presents a novel explainable AI methodology for assessing retinal vascular tortuosity from color fundus images. Deep neural networks are used to extract anatomical structures, segmenting arteries, veins, and blood vessels, and detecting the optic disc and fovea. Individual vessel segments are identified by computing the skeleton of the vascular tree, and several anatomical factors, such as vessel caliber, AV distinction, and distances to the optic disc and fovea, are calculated. The tortuosity of each vessel segment is measured using Grisan et al. (2008) metric, and the global tortuosity score is computed as a weighted average of these values. Validated on a dataset of 200 color fundus images from diabetic patients, the methodology achieved an AUC of 95.21%, outperforming previous methods and matching clinical experts' performance. The inclusion of anatomical factors significantly improved accuracy, with AV distinction providing the greatest improvement. The approach also offers visual explanations of the tortuosity values, enhancing its potential for clinical adoption.

2.2.3 Deep Learning for Tortuosity Classification

Deep learning algorithms have demonstrated effectiveness and versatility in addressing various computer vision challenges. However, despite their success, direct retinal tortuosity classification remains scarce in the literature. In order to address this gap, the application of deep learning to classify tortuosity in other organs has been explored, as the ultimate goal is similar.

Mou et al. (2022) presents a fully automated deep learning method for grading corneal nerve fiber tortuosity in corneal confocal microscopy (CCM). addressing a key challenge in clinical decision-making and disease understanding. The method aims to improve the accuracy and interpretability of tortuosity quantification by leveraging segmented corneal nerves and a two-stage process. The first stage uses a pre-trained feature extraction model, fine-tuned with a novel bilinear attention (BA) module to capture global context and predict regions of interest (ROIs), allowing for coarse grading of the image. The second stage introduces an auxiliary grading network AuxNet that refines the grading of these ROIs, and the combined results enhance overall accuracy. Figure 2.4 presents the architecture used for this purpose. Experimental results show the method achieves an 85.64% accuracy in four-level classification, outperforming previous techniques. Additionally, the model's performance is validated on a clinical dataset, demonstrating significant differences in tortuosity between healthy individuals and patients with diabetes.

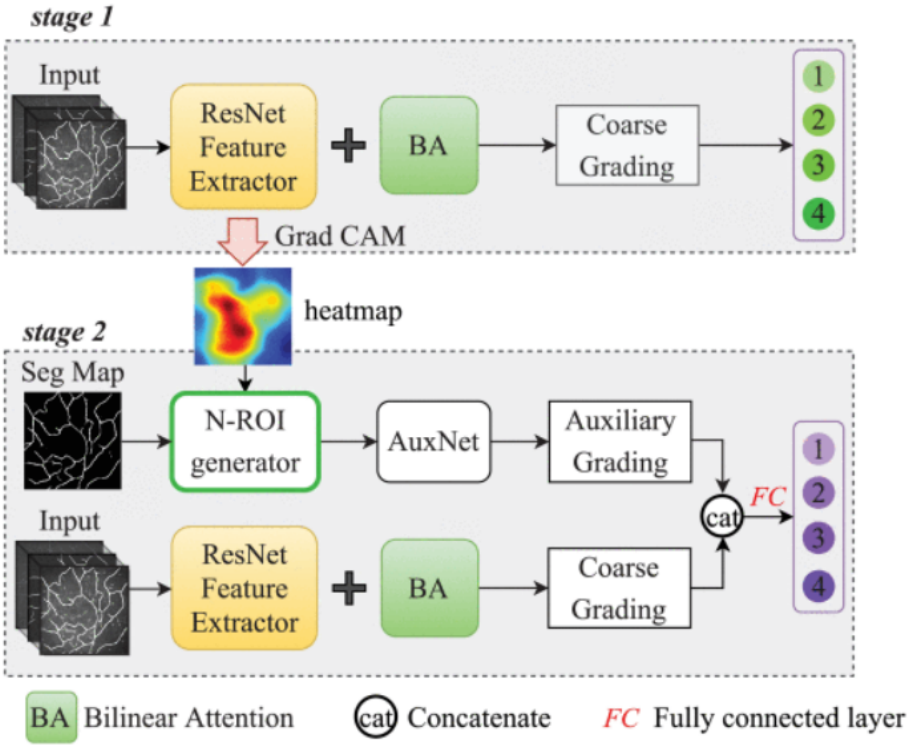


Figure 2.4: Schematic diagram of architecture used for grading tortuosity in corneal nerve fibers. The green and purple circles represent the tortuosity levels graded at two distinct stages. Although the same backbone network is used in both stages, their functions differ: stage 1 outputs N-ROIs, while stage 2 provides the coarse grading [Mou et al. (2022)].

Cobo et al. (2023) focused on detecting coronary artery tortuosity, a condition often overlooked in patients undergoing coronary angiography. Given the importance of understanding coronary artery morphology for planning interventions like stenting, the researchers aimed to develop an AI-based algorithm to automatically identify this condition. Using deep learning techniques, specifically CNNs, the model classified patients as having tortuous or non-tortuous arteries based on coronary angiographies. Figure 2.5 showcases the procedure implemented. The model was trained on 658 angiographies and achieved a test accuracy of 87%, with a high AUC of 0.96. Additionally, the sensitivity, specificity, and predictive values of the model were comparable to expert visual assessments. Since this method achieved higher accuracy, a close approach was followed for this work.

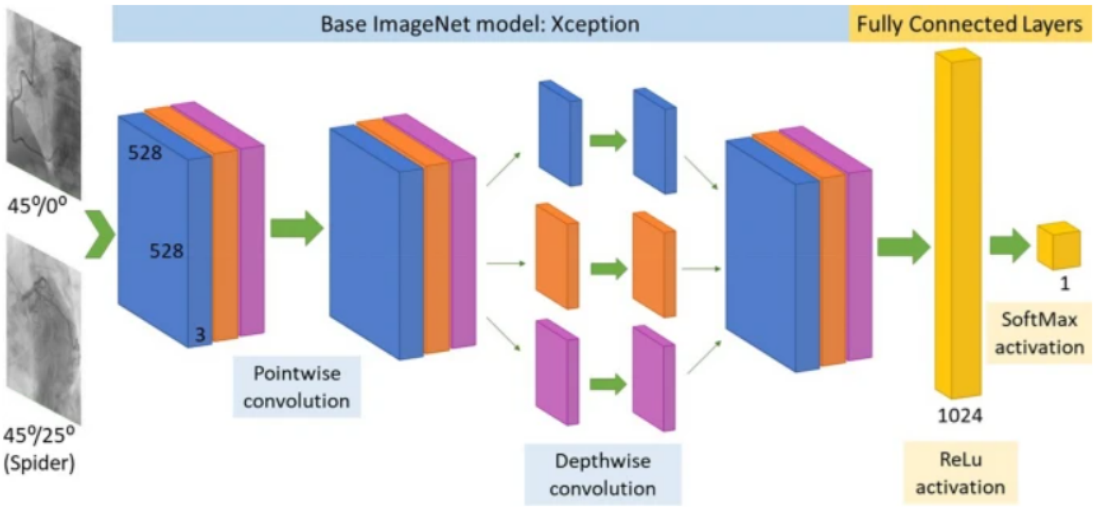


Figure 2.5: CNN architecture for coronary artery tortuosity detection [Cobo et al. (2023)].

Chapter 3

Tortuosity Grading

This work aims to develop a reliable deep learning model to analyze retinal fundus images, classifying retinal tortuosity into three categories: 'No Tortuosity', 'Some Tortuosity', and 'High Tortuosity'. Additionally, different preprocessing techniques are applied in an attempt to improve the distinction of blood vessels from other structures. This approach aims to ensure that the model not only performs well on the training data but also generalizes effectively to other datasets and potentially to different imaging techniques, such as OCTA images.

3.1 Datasets

Two datasets of fundus images were used: EIARG2 [Pourreza] and JSEIC [Cen et al. (2021)]. The first one was used for training, validation and testing of the deep learning models, while the second was applied to investigate the generalization of the models to images with other specifications, like the field of view.

The Eye Images Analysis Research Group 2 (EIARG2) dataset comprises 120 digital retinal images and has been assembled for research purposes by experts from the Retina Research Center at Mashhad University of Medical Science, specifically from the Khatam-al-Anbya hospital in Mashhad, Iran. The retinal photographs in this dataset were captured using the RetCam 3 imaging system, developed by Clarity Medical System Inc. The RetCam 3 provides a wide field of view (FOV) of 130 degrees, allowing for comprehensive visualization of the retina. Four ophthalmologists independently graded the tortuosity of the retinal images on three separate rounds. Their assessments were based on the presence and extent of tortuosity in the retinal vessels. The final unique ground truth for each image was given by majority voting the annotations. The dataset includes a comprehensive grading system for tortuosity, divided into three classes: 'No Tortuosity', 'Some Tortuosity', and 'High Tortuosity'. Additionally, there is a fourth class for images that were deemed non-gradable due to low quality. This additional classification led to the exclusion of some images. In the end, the dataset is composed of 114 images, categorized as follows: 58 images fall into the high tortuosity class, 31 images exhibit some tortuosity, and 25 images are labeled as having no tortuosity. All the images have dimensions 480 by 640 pixels.

Figure 3.1: Samples regarding the 3 classes of retinal images on the EIARG2 dataset: (A)'No tortuosity', (B) 'Some Tortuosity' , and (C) 'High Tortuosity'.

The external test dataset has 1,000 fundus images categorized into 39 distinct classes and is sourced from the Joint Shantou International Eye Centre (JSIEC) in Shantou city, China. These images represent a small subset of a larger collection of 209,494 fundus images intended for training, validating, and testing deep learning models for retinal disease detection. The images were captured using a Topcon TRC-50DX Mydriatic Retinal Camera from 2013 to 2018, with a field setting of 50 degrees. This dataset serves as a secondary dataset used for external testing of the developed models and therefore will only be partially used, specifically the 38 normal images and the 14 images with tortuosity. Two samples are represented in Figure 3.2.

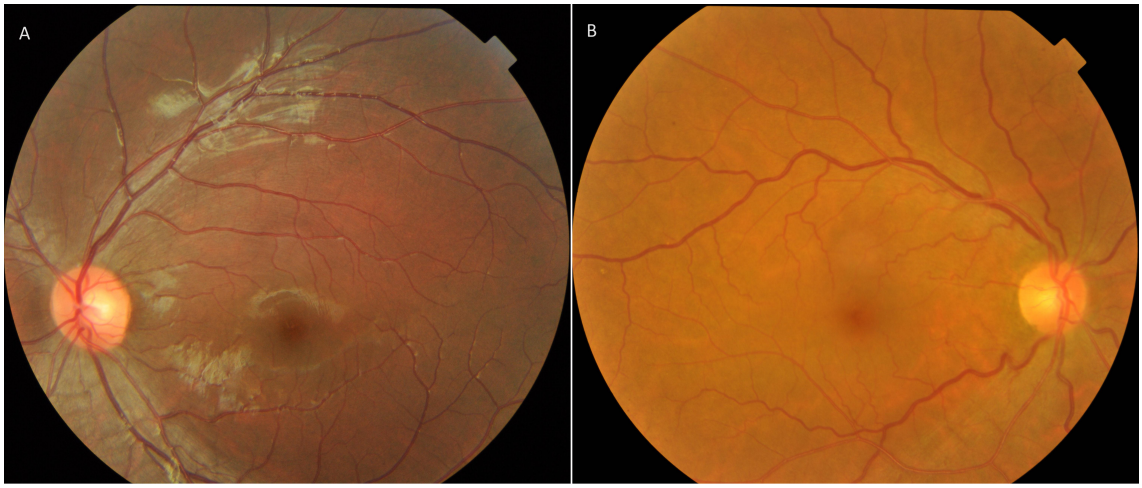


Figure 3.2: Samples regarding the 2 classes of retinal images on the JSEIC dataset: (A)'No tortuosity' and (B) 'High Tortuosity'.

3.2 Methodology

3.2.1 Data Preprocessing

Deep learning methods, as well as more traditional approaches, are frequently combined with preprocessing techniques. Researchers have observed that preprocessing images often enhances results, though there is often a lack of explanation for the choice of methods and no consensus on the most effective techniques. Improving vascular visibility and correcting intensity heterogeneity

are the main objectives of preprocessing retinal pictures. Because green channel extraction provides a higher contrast between the vessels and background than other methods, it is usually the first step and leads to better segmentation results.

One common preprocessing method is Contrast Limited Adaptive Histogram Equalization (CLAHE) used in both supervised and unsupervised approaches for vessel enhancement. CLAHE enhances local contrast while minimizing noise amplification in homogeneous regions [Sonali et al. (2019)]. For example, Owler and Rockett (2021) explored the automation of retinal vessel segmentation from fundus images using a patch-based deep learning framework built on a modified U-Net architecture. The researchers compared pre-processing techniques, including no processing, N4 bias field correction [Tustison et al. (2010)], CLAHE, and a combination of N4 and CLAHE, to mitigate background unevenness and enhance segmentation accuracy. The N4 bias field correction technique rectifies low frequency intensity, usually in MRI image data, but has also been applied to retinal images [Kaba et al. (2014)]. The results, tested in three public available datasets - DRIVE Staal et al. (2004), STARE [Hoover et al. (2000)], and CHASE [Fraz et al. (2014)] - demonstrated that while CLAHE alone had the most significant positive effect, combining it with N4 bias correction improved performance only in images with highly uneven backgrounds.

Since the EIARG2 dataset had originally dimensions 480 by 640 pixels and the final dimensions of the images should be 528 by 528, it was necessary to handle the changes in dimensions carefully, leading to the consideration of multiple approaches. Initially, a straightforward resizing was attempted. However, this method distorted the blood vessels, potentially skewing conclusions about tortuosity. As a result, this approach was scrapped. Two alternatives were selected: first cropping and then resizing, or first padding the images and then downsizing. Both were implemented and independently used for training the deep learning models.

Several steps are consistently applied to the EIARG2 dataset in order to prepare the image data for further analysis. Initially, the green channel is extracted and negated to emphasize the blood vessels.

3.2.1.1 Cropping

The cropping starts from the points that position the largest possible square in the middle of the image. This means finding the smaller dimension (in this case height) and using it to set the boundaries for both the horizontal and vertical cuts. The result is a centered square slice of the image of 480 by 480. Once the central portion is cropped out, it is then resized to 528 by 528 pixels.

3.2.1.2 Padding

The padding process begins by obtaining the dimensions of the original image and creating a new blank canvas of 528 by 528 pixels filled with ones. The smaller image is then centered within

this larger canvas by calculating the horizontal and vertical offsets, ensuring the image is perfectly placed in the middle. Padding is preferred over cropping in preprocessing, particularly when dealing with image data in machine learning. Padding maintains the original content and aspect ratio of the image by adding extra pixels. This approach ensures that no important information is lost. On the other hand, cropping removes parts of the image, potentially discarding significant details and altering the context. Therefore, further preprocessing steps were built on top of this padding.

3.2.1.3 Mean background

To handle noise present in the image borders, a mask is created by converting the image to binary, followed by an erosion process. Each connected component within the binary image is then labeled, and components exceeding the specified area threshold (0.05) are selected for further processing. These selected components are dilated to smooth their edges. The inverted green channel is then combined with the refined mask to calculate the mean value, which normalizes the images. The processed green channel is padded to fit a standard size of 640 by 640 pixels. Additionally, a new mask is created and padded to the same size, filled with the calculated mean value to maintain consistency. The final processed image is produced by combining the normalized green channel with the new mask, ensuring a uniform background tone. This final image is resized to 528 by 528, suited to go through the deep learning models. Maintaining a mean background and removing border noise in the dataset are steps intended to improve accuracy in the classification of tortuosity. A consistent background ensures uniformity, allowing the classifier to focus on the vascular structures without being misled by varying background intensities. Additionally, eliminating noise present in the borders is expected to also remove irrelevant artifacts. Together, these preprocessing steps might contribute to better model training and testing, reducing the risk of overfitting and enhancing overall classifier performance.

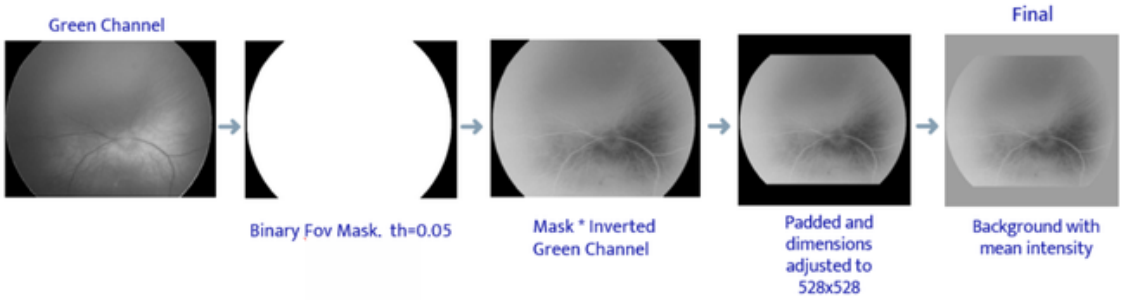


Figure 3.3: Preprocessing pipeline of Mean Background method.

3.2.1.4 CLAHE

Later on, it was decided that CLAHE should be incorporated into the image preprocessing pipeline based on its proven effectiveness in the literature regarding segmentation tasks. CLAHE brings out subtle details in vascular structures that might be lost in the shadows or highlights of the original image, without over-amplifying noise, which ensures that the focus remains on the true anatomical details rather than the background artifacts. From a human examiner stand point, the blood vessels become clearer using this preprocessing technique. Figure 3.4 synthesizes the methods applied to the EIARG2 dataset.

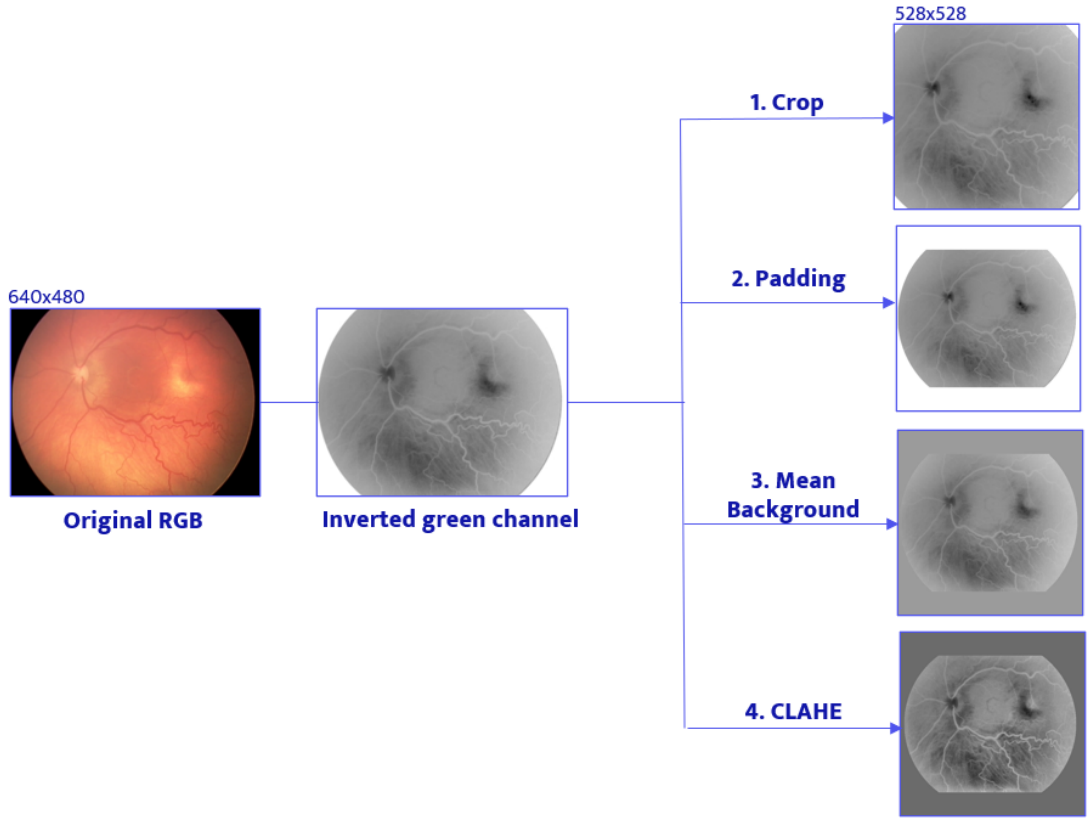


Figure 3.4: Schematic of the four methods applied to the EIARG2 dataset: Cropping, Padding, Mean Background, and CLAHE.

JSEIC dataset

Some of the preprocessing techniques previously presented were also applied to the JSEIC dataset. Padding was applied to standardize the image sizes to square dimensions, ensuring uniformity across the dataset while preserving the important features without distortion or cropping. Mean background correction was utilized to minimize variations in lighting and reduce background noise, which could otherwise obscure critical details in the images. Additionally, CLAHE was implemented for better image contrast, improving the visibility of the blood vessels and making it easier for the models to detect subtle tortuosity features.

The main differential of the preprocessing procedure in this dataset is that three resizing factors were applied for each one of the methods aforementioned: 0.9, 0.75 and 0.6. Essentially, nine sets of data were generated from the original set. Since the JSEIC dataset is composed of images with a smaller FOV than the dataset used for training, the blood vessels have different size and more capillaries are observed. Therefore, it was hypothesized that submitting these images with finer details to the deep learning models, trained on images with a wider FOV, would lead to under performance when it comes to grading the tortuosity. Figure 3.5 presents the scheme of the JSEIC preprocessing.

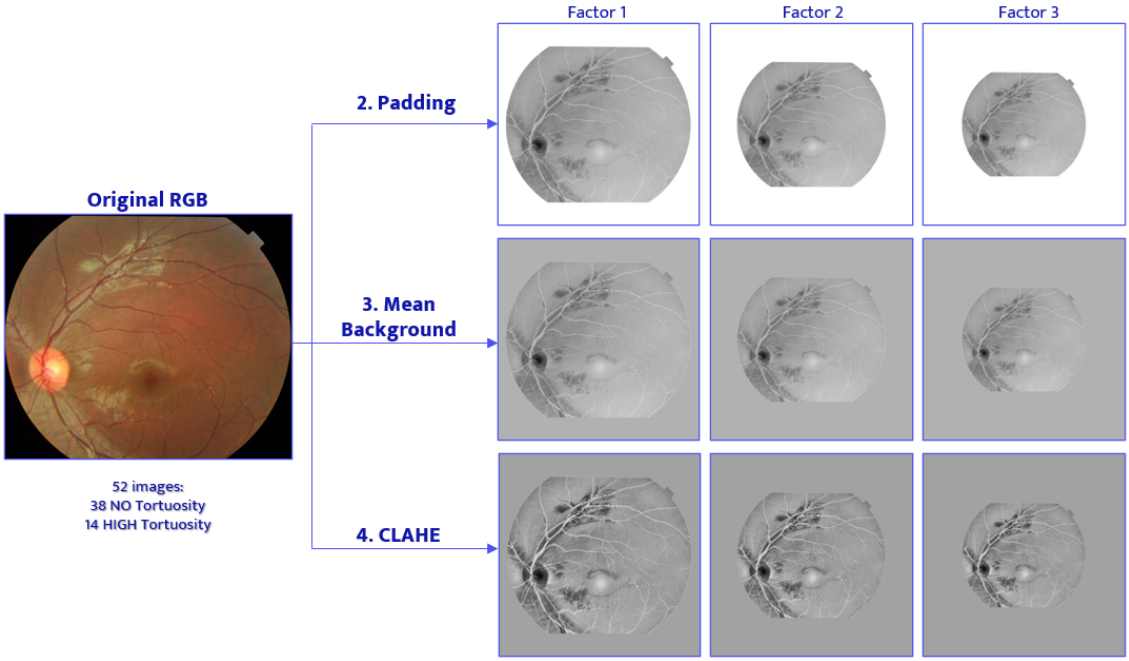


Figure 3.5: Pre-processing schematic of JSEIC dataset: each row represents a different preprocessing method and each column a different resizing factor. Factors 1, 2 and 3 correspond respectively to resizing factors of 0.9, 0.75 and 0.6.

3.2.2 Model Architecture

Using deep learning models requires preparing adequate datasets, which can be a significant challenge, especially when data is scarce or not publicly available. This is where transfer learning plays a crucial role. By allowing a model trained on one task to be adapted for a new, but similar task, transfer learning reduces the need for large amounts of labeled data. Instead of starting the learning process from scratch, this method is sustained by the knowledge gained from a previous task with abundant data and applies it to a new problem with fewer resources.

The Xception architecture is commonly used due to its feature extraction capabilities, having been pre-trained on the extensive ImageNet dataset. By adapting the pre-trained Xception model, it can be fine-tuned for specific tasks through the addition of custom layers, making it effective in scenarios with limited data.

The Xception model, developed by Chollet (2017), is a deep learning architecture that enhances the Inception [Szegedy et al. (2014)] model by using depthwise separable convolutions. These convolutions consist of two parts: depthwise convolution, which applies a single filter per input channel, and pointwise convolution, which combines the outputs of the depthwise convolution with 1x1 convolutions. This approach improves efficiency and performance.

The architecture of Xception is divided into three main flows shown in Figure 3.6 : Entry, Middle, and Exit. The Entry Flow starts with the input that usually has equal height and width, because makes handling the data easier and less computationally costly. For this particular case and in order to preserve the maximum information about the blood vessels topology, the dimensions of the input images were settled on 528 by 528 pixels. This entry data passes through a convolutional layer with 32 filters and a 3x3 kernel, using a stride of 2, followed by batch normalization and rectified linear unit (ReLU) activation. This is followed by another convolutional layer with 64 filters, a 3x3 kernel, and a stride of 1, again followed by batch normalization and ReLU. This flow includes three residual blocks, each comprising depthwise separable convolutions with intermediate batch normalization and ReLU activations. Each block has a stride of 2 at its first layer to reduce spatial dimensions.

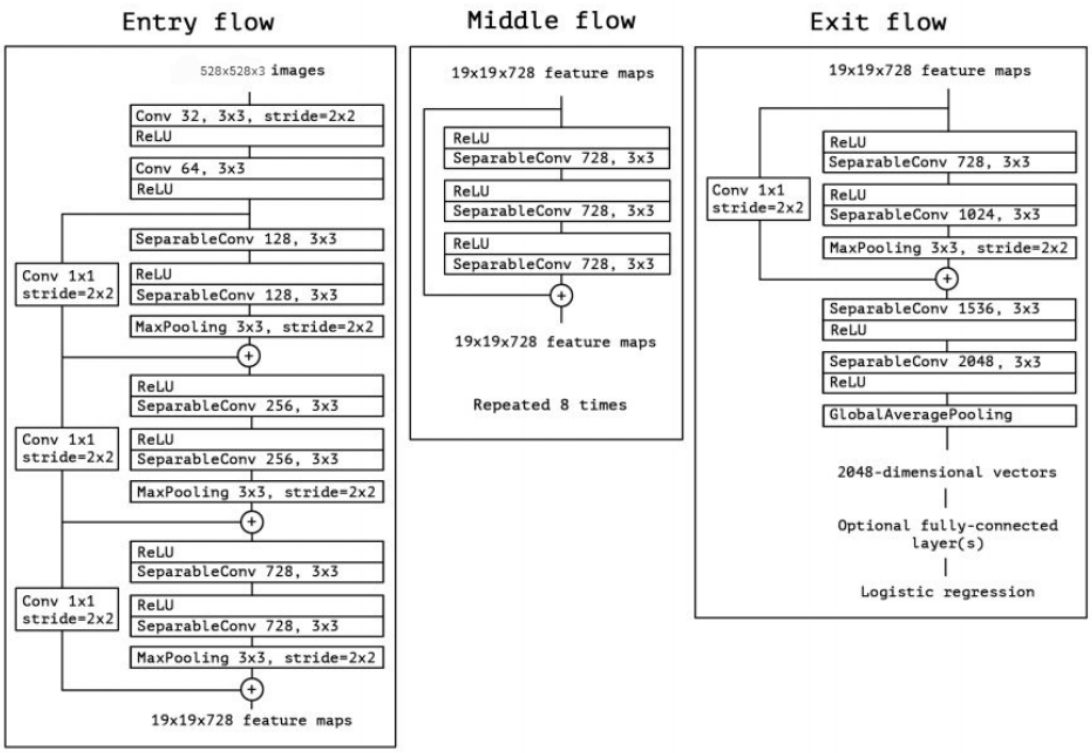


Figure 3.6: The Xception architecture: the data first goes through the entry flow, then through the middle flow which is repeated eight times, and finally through the exit flow [Chollet (2017)].

The Middle Flow contains eight residual blocks, each made up of three depthwise separable convolution layers, followed by batch normalization and ReLU activations. The input and output dimensions remain consistent across these blocks, ensuring that the feature maps' size and depth

remain unchanged throughout this section. The Exit Flow begins with two residual blocks. The first block includes a depthwise separable convolution followed by a 1x1 convolution to match dimensions, while the second block further reduces the spatial dimensions. This is followed by the global average pooling, which reduces the dimensions to 1x1x2048, and a fully connected layer with typically 1000 units for ImageNet classification, ending with a softmax activation.

Xception offers several advantages when compared with other deep learning architectures, including enhanced efficiency and performance due to the use of depthwise separable convolutions, which reduce the number of parameters and computational cost. It achieves better accuracy on benchmarks compared to the Inception architecture. After extracting features using the Xception model, a custom set of layers is added to adapt it for the classification task at hands. First, a Global Average Pooling layer is applied, which reduces the dimensionality of the feature maps by computing the average across their spatial dimensions. This acts as a way to compress the feature maps into a more compact form without introducing too many new parameters. Following that, a fully connected Dense layer with 1024 neurons and a ReLU activation function is added to introduce non-linearity. Finally, a Dense output layer is used, where the number of neurons corresponds to the number of target classes, in this case 3. This layer employs a softmax activation function to output a probability distribution over the different classes.

3.2.3 Training Procedure

3.2.3.1 Cross-Validation

A five-fold stratified cross-validation approach was employed for the Xception model training, dividing the dataset into five equal parts (Figure 3.7). In each round, the model was trained on four of these parts, each containing 23 images (except fold 5 which has 22 images). The remaining one, was reserved for validation and testing, half is the subset for validation and the other half for testing. The stratification procedure refers to maintaining the distribution of classes consistently across each part.

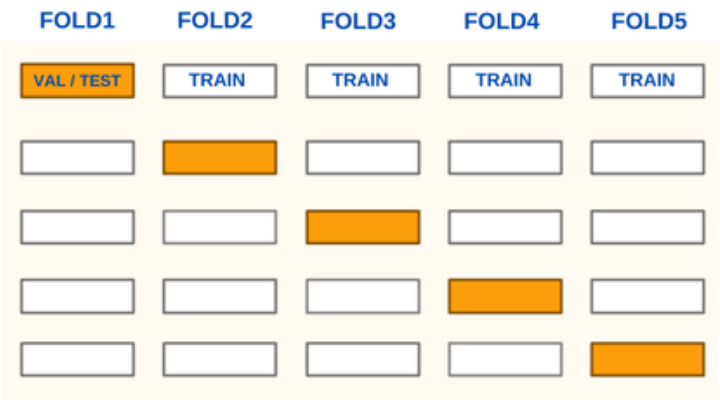


Figure 3.7: Five-fold stratified cross validation representation.

3.2.3.2 Augmentations

The augmentation procedure for training involves several transformations applied with specified probabilities to enhance the dataset's diversity and improve model generalization. Horizontal and vertical flips are performed with a 50% probability. Similarly, a 50% chance is assigned to image rotation, with a maximum rotation limit of 90 degrees, introducing varied perspectives and angles. Stretching, cropping, zooming, blurring, pixel noise, and cutout are set to zero, indicating no modifications in these aspects, ensuring the integrity of the images.. However, pixel saturation adjustments are applied with a 10% probability, subtly altering the image colors and improving the model's robustness to color variations. This combination of augmentations, particularly flips and rotations, effectively expands the training dataset's variability without compromising the original image content.

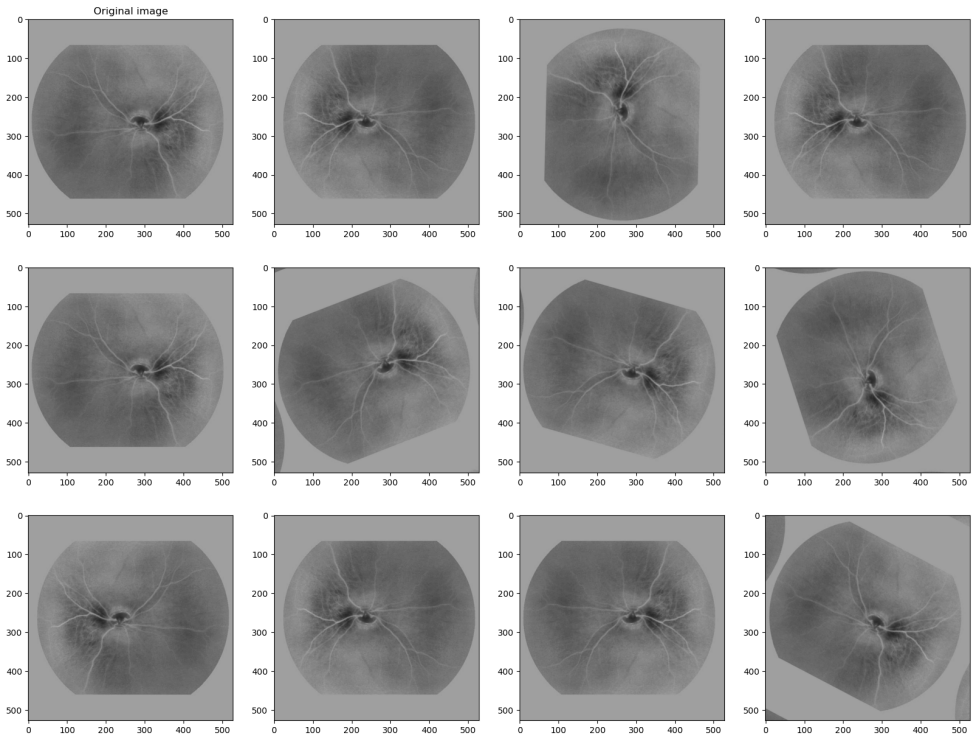


Figure 3.8: Example of random augmentations on the EIARG2 dataset.

3.2.3.3 Hyperparameter Settings

The training procedure employs categorical cross-entropy as loss function to measure the discrepancy between predicted labels and the ground truth annotations of the deep learning models. Due to the class imbalance of the dataset, there is class weights that attribute more focus on the underrepresented images, namely the 'High Tortuosity' class.

$$\mathscr{L}_{wCCE} = -\left[\sum_{i=1}^{3} w_i \cdot y^i_{\text{true}} \cdot \log(y^i_{\text{pred}})\right] \tag{3.1}$$

Where $\mathscr{L}_{wCCE}$ represents the weighted categorical cross-entropy loss. The summation passes through all 3 classes i. The variable w_i is the weight assigned to class i, while the variable y^i_{true} denotes the true label. Finally, $\log(y^i_{\text{pred}})$ is the natural logarithm of the predicted probability for class i. The overall loss $\mathscr{L}_{wCCE}$ is the negative weighted sum of these terms across the three classes.

Through experimentation and due to the computer power available, the batch size was set to 8, which determines that 8 images are processed in each iteration.

An appropriately chosen learning rate prevents overshooting the optimal solution and ensures steady progress towards minimal loss. For this case, the starting learning rate was defined as 0.01. Additionally, the learning rate was managed through a schedule, set to 'step' mode. This step schedule involved decaying the learning rate by a factor of 0.1 at the specified points during training. Specifically, the learning rate was diminished at 70 % and 90 % of the total number of epochs.

The activation function ReLU is used in each layer to introduce non-linearity, enabling the network to learn complex patterns. L2 regularization with a coefficient of 0.0001 is applied to the last two Dense layers. This technique penalizes large weights, encouraging the model to maintain simpler parameters. Additionally, dropout with a rate of 0.5 is employed to randomly deactivate neurons during training.

The training process ran for 300 epochs and the Adam optimizer updated the model's parameters based on the gradients of the loss function. The models were initialized with the weights from the ImageNet dataset, excluding the top classification layers.

3.2.4 Framework

Python 3.9.19 is chosen for its extensive range of libraries, making it well-suited for the tasks at hand. The Keras API streamlines transfer learning, enabling the easy loading and preparation of pre-trained models. The efficiency of training was boosted by the GPU-accelerated version of TensorFlow, making use of Nvidia's CUDA and cuDNN libraries for faster computation. Aditionally, essential libraries such as Numpy and Matplotlib are used for tasks like numerical computation and data visualization. For image preprocessing, MATLAB is also incorporated to handle the data processing steps mention in Section 3.2.1.

3.2.5 Performance Metrics

In the context of evaluating classification models, some metrics commonly assess performance. These metrics include accuracy, macro-recall, macro-precision, and the F1 score. Each metric provides different insights into the performance of a classifier, especially in the presence of imbalanced datasets.

Accuracy

Accuracy is the proportion of correctly classified instances out of the total instances. It gives an overall measure of the model's performance but can be misleading in cases of class imbalance.

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{3.2}$$

where:

- TP (True Positives) are the instances correctly predicted as positive.
- TN (True Negatives) are the instances correctly predicted as negative.
- FP (False Positives) are the instances incorrectly predicted as positive.
- FN (False Negatives) are the instances incorrectly predicted as negative.

Macro-Recall

Macro-recall is the arithmetic mean of recall values for each class. It treats all classes equally by giving them equal weight, making it useful in evaluating the performance on each class independently.

$$\text{Recall}_i = \frac{TP_i}{TP_i + FN_i} \tag{3.3}$$

$$\text{Macro-Recall} = \frac{1}{C} \sum_{i=1}^{C} \text{Recall}_i \tag{3.4}$$

where:

- C is the number of classes.
- TP_i and FN_i are the true positives and false negatives for class i.

Macro-Precision

Macro-precision is the arithmetic mean of precision values for each class. Like macro-recall, it gives equal weight to each class, making it effective in assessing the performance across all classes.

$$\text{Precision}_i = \frac{TP_i}{TP_i + FP_i} \tag{3.5}$$

$$\text{Macro-Precision} = \frac{1}{C} \sum_{i=1}^{C} \text{Precision}_i \tag{3.6}$$

where:

- FP_i is the false positives for class i.

F1 Score

The F1 score is the harmonic mean of precision and recall. It is a single measure that balances both the precision and recall, providing a useful metric when there is an uneven class distribution. The F1 score is especially relevant in scenarios where both false positives and false negatives carry significant costs.

$$\text{F1}_i = 2 \cdot \frac{\text{Precision}_i \cdot \text{Recall}_i}{\text{Precision}_i + \text{Recall}_i} \tag{3.7}$$

$$\text{Macro-F1} = \frac{1}{C} \sum_{i=1}^{C} \text{F1}_i \tag{3.8}$$

These metrics collectively offer a comprehensive view of a classifier's performance, each highlighting different aspects and potential areas for improvement.

3.2.6 Saliency Maps

In deep learning models have become increasingly complex and opaque, often termed as 'black boxes'. Saliency maps improve the explainability of these models by visualizing the parts of the input that most influence predictions. They reveal which regions of an image or text are critical for decision-making, using techniques like gradient-based methods (e.g., vanilla gradients [Erhan et al. (2009)], Grad-CAM [Selvaraju et al. (2019)]), occlusion-based approaches, perturbation-based techniques (e.g., LIME [Shi et al. (2020)]), and activation-based methods (e.g., Class Activation Maps [Zeiler and Fergus (2014)]). Saliency maps are valuable for model debugging, improving trust and transparency, and detecting biases, as they highlight important features and provide insights into model behavior. However, they can be challenging to interpret, sensitive to input changes, and computationally expensive, often offering only partial explanations. In this study, gradient saliency maps, particularly Gradient Saliency and Guided Backprop [Springenberg et al. (2015), were used to iteratively refine deep learning models by identifying key features that influenced predictions.

Chapter 4

Results and Discussion

This chapter presents the results of the classification of retinal tortuosity in fundus images using multiple deep learning models, comparing their performance across the four previously detailed preprocessing methods. The results focus on the evaluation metrics, like the F1-score and accuracy, as well as illustrative examples of model predictions, saliency maps, and analysis of misclassifications. Additionally, the chapter comprises the results on an external test set, which contains images captured with a different FOV. Furthermore, it also evaluates model performance on OCTA images.

4.1 Retinal Fundus Photographs

The performance evaluation focuses on the four datasets originated from the preprocessing methods: cropping, padding, mean background, and CLAHE. For each method, graphs displaying the evolution of training and validation loss, along with accuracy throughout the epochs, will be presented. The models trained on retinal fundus images are assessed using a 5-fold cross-validation strategy, that led to the training of five distinct models. By comparing the accuracy, precision, recall, and F1-score across the five models for each preprocessing method, the consistency and robustness of the models can be determined. Since the dataset is imbalanced, closer attention was paid to F1-score, instead of accuracy, which in this situation can be deceiving.

4.1.1 Cropping

Figure 4.1 depicts the loss for both the training and validation sets for the five models made with fold rotation and using the cropping dataset. It reveals a trend where the validation loss experiences several spikes before stabilizing. The main driver of instability might be the very small number of samples used for validation. However, as the training progresses, these spikes diminish, and the loss eventually reaches a plateau. This stabilization is largely attributed to the learning rate scheduler, which adjusts the learning rate dynamically throughout the training process, at epoch 210 and 270. By reducing the learning rate, the scheduler helps the model converge, mitigating the erratic behavior.

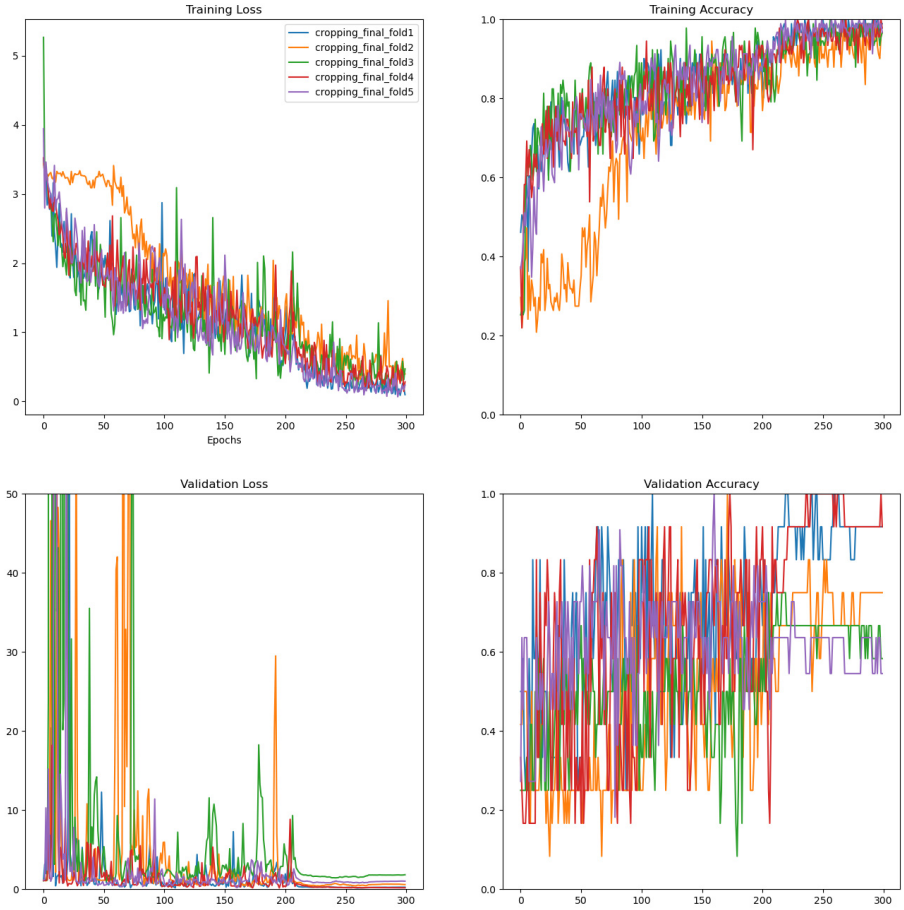


Figure 4.1: Training and validation loss over the epochs (left), along with training and validation accuracy (right), regarding the cropping preprocessing.

Table 4.1: Performance metrics for validation and test sets under the cropping configuration for the 5 folds in percentage (%).

Cropping	**Metric**	**Fold1**	**Fold2**	**Fold3**	**Fold4**	**Fold5**	**Mean ± STD**
Validation	Accuracy	91.7	75.0	58.3	91.7	54.5	74.2 ± 14.5
	Precision	**91.7**	78.3	38.9	**93.3**	44.4	69.3 ± 21.2
	Recall	88.9	83.3	44.4	83.3	48.9	69.8 ± 17.4
	F1-score	**88.6**	75.8	40.0	85.2	46.5	67.2 ± 18.4
Test	Accuracy	45.5	63.6	**81.8**	**81.8**	**81.8**	70.9 ± 13.3
	Precision	20.8	57.1	**84.1**	82.2	**84.1**	65.7 ± 22.5
	Recall	27.8	55.6	72.2	**82.2**	72.2	62.0 ± 17.4
	F1-score	23.8	55.6	75.2	**82.2**	75.2	62.4 ± 19.4

Table 4.1 presents the F1-scores for the validation set that vary significantly across folds. The model performs best when Fold 1 is the validation set (see section 3.2.3.1) with an F1-score of 88.6%, but drops sharply with Fold 3 to 40.0 %, indicating an imbalance between precision

and recall. The average F1-score is 67.2 %, with a high standard deviation of 18.4%, showing inconsistent performance across the validation data. On the test set, F1-scores are generally lower and vary as well. Fold 1 scores just 23.8%, while Folds 3, 4, and 5 improve to 75.2%, 82.2%, and 75.2%. The average test score is 62.4%, with a high standard deviation of 19.4%, again reflecting inconsistency.

4.1.2 Padding

The results displayed in Figure 4.2 show the training and validation loss, as well as the training and validation accuracy, for the five folds trained with Padding preprocessed images. In the top-left plot, the training loss decreases consistently across all folds as the epochs progress, indicating that the model is learning effectively during training. The training accuracy, shown in the top-right, follows an upward trend over the course of the training process for each fold. On the other hand, the validation loss in the bottom-left plot exhibits higher volatility, particularly in the earlier epochs, before stabilizing at lower values towards the end.

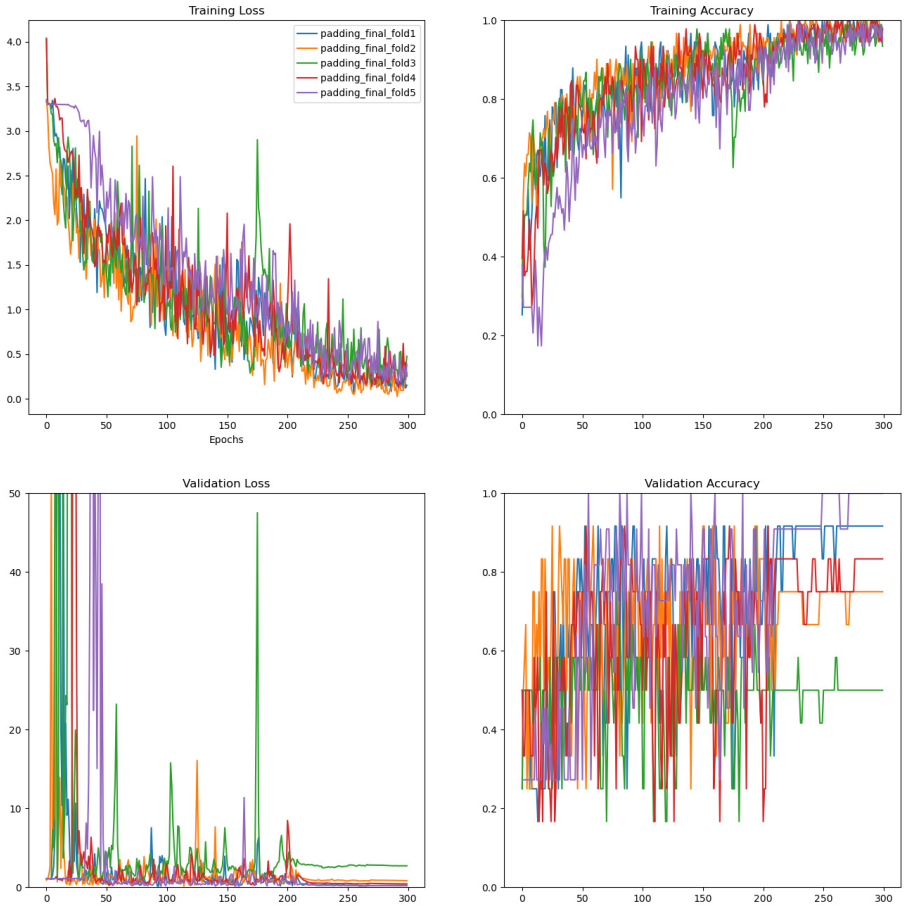


Figure 4.2: Training and validation loss over the epochs (left), along with training and validation accuracy (right), regarding the padding preprocessing.

The results for the five models generated with the padding preprocessing method with fold rotation are presented in Table 4.2. For the validation set, the model performs well in Fold 1 with an F1-score of 88.6%, indicating a strong balance between precision and recall. Fold 2 sees a slight decrease to 76.6%, but performance remains robust. In Fold 3, the F1-score drops sharply to 37.1%, revealing significant issues with that subset. Fold 4 recovers slightly with an F1-score of 60%, while Fold 5 achieves a perfect 100%. The average F1-score across all validation folds is 72.5%, with a high standard deviation of 20.3%, indicating high variability. The test set shows similar variability. Fold 1 has an F1-score of 61.0%, Fold 2 drops to 49.8%, Fold 3 remains at 55.6%, and Fold 4 improves slightly to 61.9%. Fold 5 also scores a perfect 100%. The average test F1-score is 65.7%, with a standard deviation of 16.2%, reflecting slightly more consistent performance than the validation set.

Table 4.2: Performance metrics for validation and test sets under the padding configuration for the 5 folds in percentage (%).

Padding	**Metric**	**Fold1**	**Fold2**	**Fold3**	**Fold4**	**Fold5**	**Mean ± STD**
Validation	Accuracy	91.7	75.0	50	83.3	**100**	80.0 ± 15.7
	Precision	91.7	76.7	37.5	55.6	**100**	72.3 ± 21.0
	Recall	88.9	77.8	38.9	66.7	**100**	74.5 ± 19.1
	F1-score	88.6	76.6	37.1	60	**100**	72.5 ± 20.3
Test	Accuracy	54.5	72.7	63.6	63.6	**100**	70.9 ± 14.3
	Precision	61.7	45.2	55.6	61.1	**100**	64.7 ± 17.0
	Recall	61.1	55.6	55.6	64.4	**100**	67.3 ± 15.2
	F1-score	61.0	49.8	55.6	61.9	**100**	65.7 ± 16.2

4.1.3 Mean Background

The results displayed in Figure 4.3 present the training and validation loss and accuracy for the mean background preprocessing. The training loss decreases consistently across all folds as the epochs progress, while accuracy approaches 100 %. In contrast, the validation loss shows considerable instability, especially in the earlier epochs, before stabilizing at lower values towards the end. Paying close attention to validation loss and accuracy, it is noticeable that Fold 3 (green) generates the worst-performing model.

The F1-scores for the mean background configuration reveals once again performance variability (see Table 4.3). In the validation set, Fold 1 has a moderate F1-score of 65.7%, while Fold 2 has good performance with 90.8%. The average F1-score is 69.7% with a high standard deviation of 20.3%, indicating significant inconsistency. For the test set, Fold 1 scores low at 35.2% and Fold 5 achieves the highest score of 90.8%. The average test F1-score is 59.7% with a standard deviation of 16.2%, showing variability, though slightly less than the validation set.

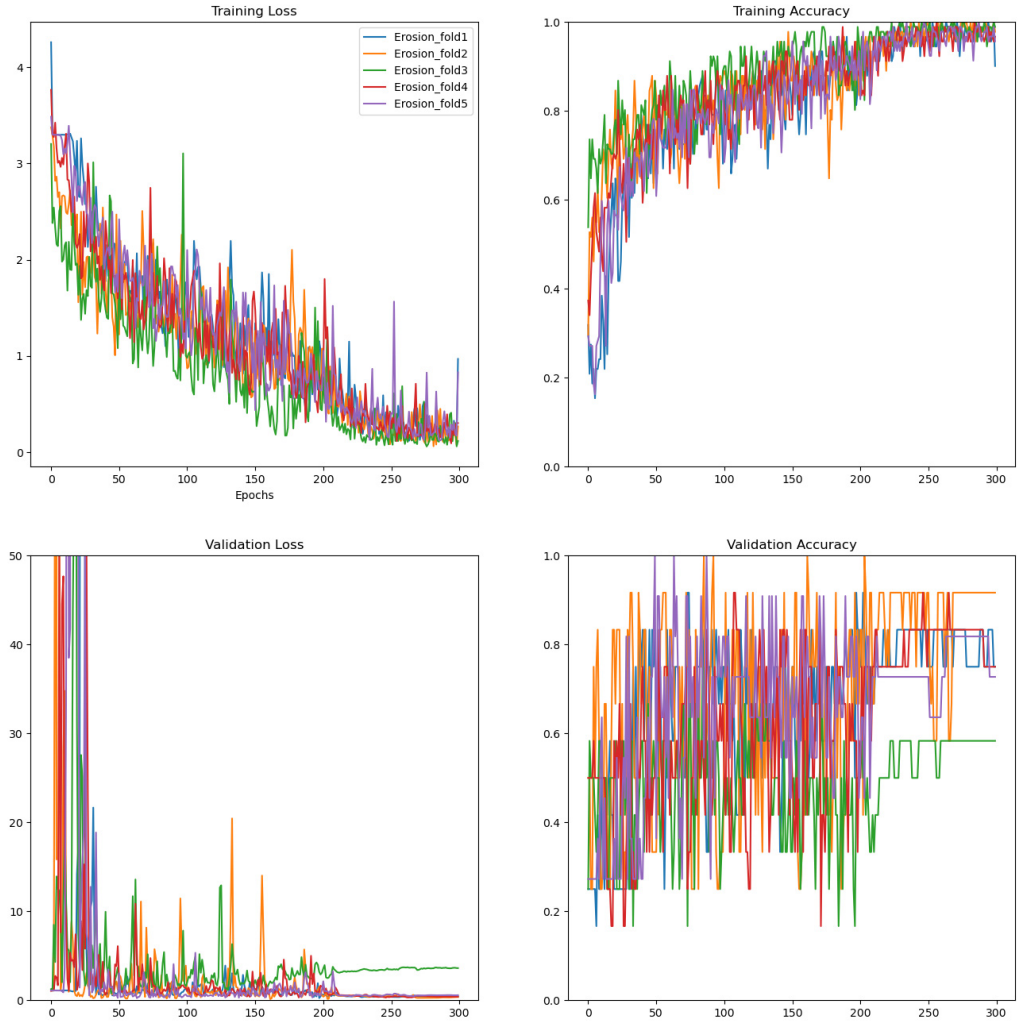


Figure 4.3: Training and validation loss over the epochs (left), along with training and validation accuracy (right), regarding the mean background preprocessing.

Table 4.3: Performance metrics for validation and test sets under the mean background configuration for the 5 folds in percentage (%).

Mean Background	Metric	Fold1	Fold2	Fold3	Fold4	Fold5	Mean ± STD
Validation	Accuracy	75.0	**91.7**	58.3	75.0	72.7	74.5 ± 15.7
	Precision	66.7	**95.2**	57.8	70.0	68.3	71.6 ± 21.0
	Recall	66.7	**88.9**	55.6	69.4	71.1	70.3 ± 19.1
	F1-score	65.7	**90.8**	54.4	69.2	68.6	69.7 ± 20.3
Test	Accuracy	54.5	72.7	36.4	72.7	**90.9**	65.4 ± 14.3
	Precision	32.1	61.9	55.6	68.3	**95.2**	62.6 ± 17.0
	Recall	38.9	61.1	38.9	71.1	**88.9**	59.8 ± 15.2
	F1-score	35.2	60.8	43.0	68.6	**90.8**	59.7 ± 16.2

4.1.4 CLAHE

The final image preprocessing method, CLAHE, exhibits a pattern similar to the previous methods, with noticeable spikes as shown in Figure 4.4. In this case, the model developed with Fold 3 as validation demonstrates lower performance. At the end of the 300 epochs, loss is much higher comparatively with other models.

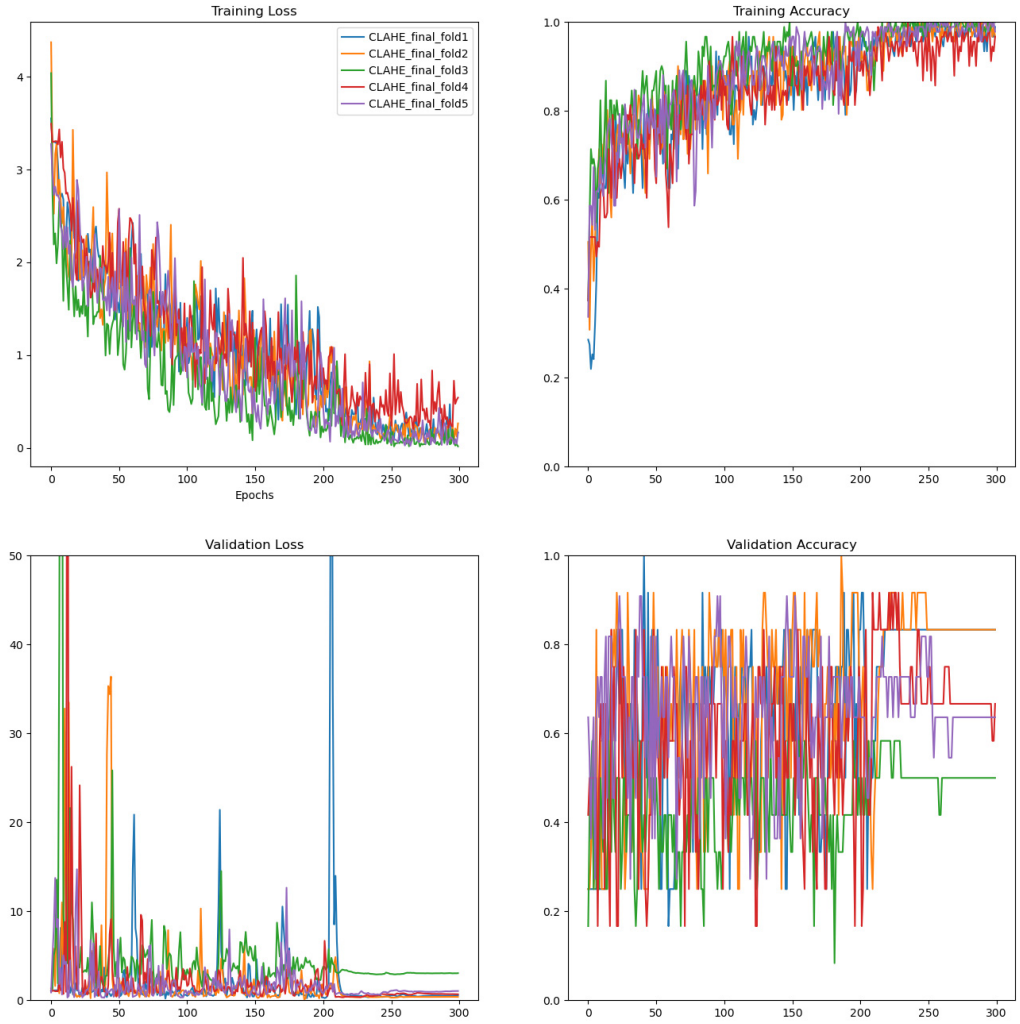


Figure 4.4: Training and validation loss over the epochs (left), along with training and validation accuracy (right), regarding the CLAHE preprocessing.

Table 4.4 presents the results for the CLAHE preprocessing. Some folds show very low accuracy and F1-scores, like Fold 3 in the validation set and Fold 2 in the test set. Conversely, Fold 4 achieves higher scores in validation set, while Fold 4 set achieves better performance. The high standard deviations for both sets indicate significant variability, which is unsatisfactory.

Table 4.4: Performance metrics for validation and test sets under the CLAHE configuration for the 5 folds in percentage (%).

CLAHE	Metric	Fold1	Fold2	Fold3	Fold4	Fold5	Mean ± STD
Validation	Accuracy	**83.3**	**83.3**	50.0	66.7	63.6	69.4 ± 11.6
	Precision	**86.7**	83.3	25.0	66.7	60.0	64.3 ± 20.1
	Recall	77.8	**83.3**	33.3	63.9	60.0	63.7 ± 15.9
	F1-score	75.0	**83.3**	28.6	63.3	60.0	62.0 ± 17.0
Test	Accuracy	72.7	54.5	**81.8**	**81.8**	**81.8**	74.5 ± 9.7
	Precision	73.8	25	**86.7**	82.2	84.1	70.4 ± 21.1
	Recall	72.2	33.3	77.8	**82.2**	72.2	67.5 ± 16.0
	F1-score	72.3	28.6	77.5	**82.2**	75.2	67.2 ± 17.8

4.1.5 Overall Considerations

Table 4.5 shows the best and worst-performing models for each of the four preprocessing techniques. Fold 3 consistently emerges as the worst model across all metrics, raising concerns about the images used for that validation set. A closer examination revealed that these images had dubious ground truth annotations, due to challenging cases where even experts struggled to distinguish between adjacent classes. This discrepancy suggests that the issues with Fold 3 may be linked to the quality of the annotations rather than the preprocessing techniques themselves.

Table 4.5: Identification of best and worst performing models.

Performance		Cropping	Padding	MeanBackground	CLAHE
Validation	Best	Fold1	Fold5	Fold2	Fold2
	Worst	Fold3	Fold3	Fold3	Fold3
Test	Best	Fold4	Fold5	Fold5	Fold4
	Worst	Fold1	Fold2	Fold1	Fold2

In figure 4.5 two retinal images are shown along with the predictions of tortuosity given by the deep learning model developed with CLAHE dataset and Fold 4 as validation set. Both are annotated as having some tortuosity (numerically, 1). However visually they are distinct from each other, with the first one having less tortuous vessels and the second presenting higher degree of tortuosity. The predictions in the right upper corner of each retinal image translate exactly that. The first one is classified with probability 99.2% as 'No Tortuosity', while the second as 'High Tortuosity' with probability 98.4%. Some of the experts also gave these cases this classification, but overall the ground truth was not defined as that (see section 3.1).

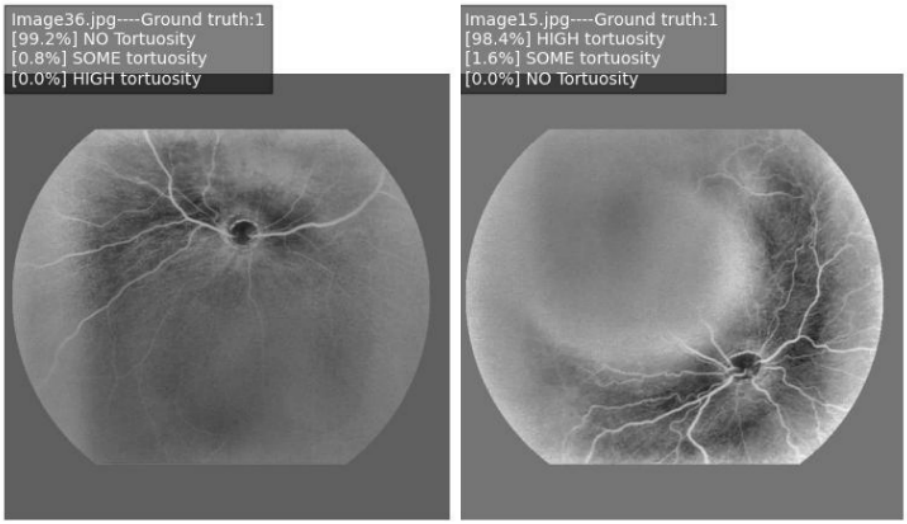


Figure 4.5: Examples of wrong predictions for images labeled as 'Some Tortuosity' in the validation set of Fold 3.

In figure 4.6 there is two others retinal images, but in this case both have as ground truth 'High Tortuosity' (numerically, 2). The prediction of the model classifies them as having only some degree of twisted blood vessels, which visually seems plausible. To sum up, this fold is a particular difficult one due to the images that were randomly selected to be part of it. Since the dataset is small, they cause a severe impact in performance.

Figure 4.6: Examples of wrong predictions for images labeled as 'High Tortuosity' in the validation set of Fold 3.

Figure 4.7 presents the global confusion matrix for each one of the models implementing different preprocessing methods. The global confusion matrix is computed by adding the five confusions matrixes obtained with fold rotation. This way a better analysis per class can be performed.

The results are satisfactory, specially because misclassifications happen almost exclusively between two adjacent classes for the models generated with every preprocessing technique.

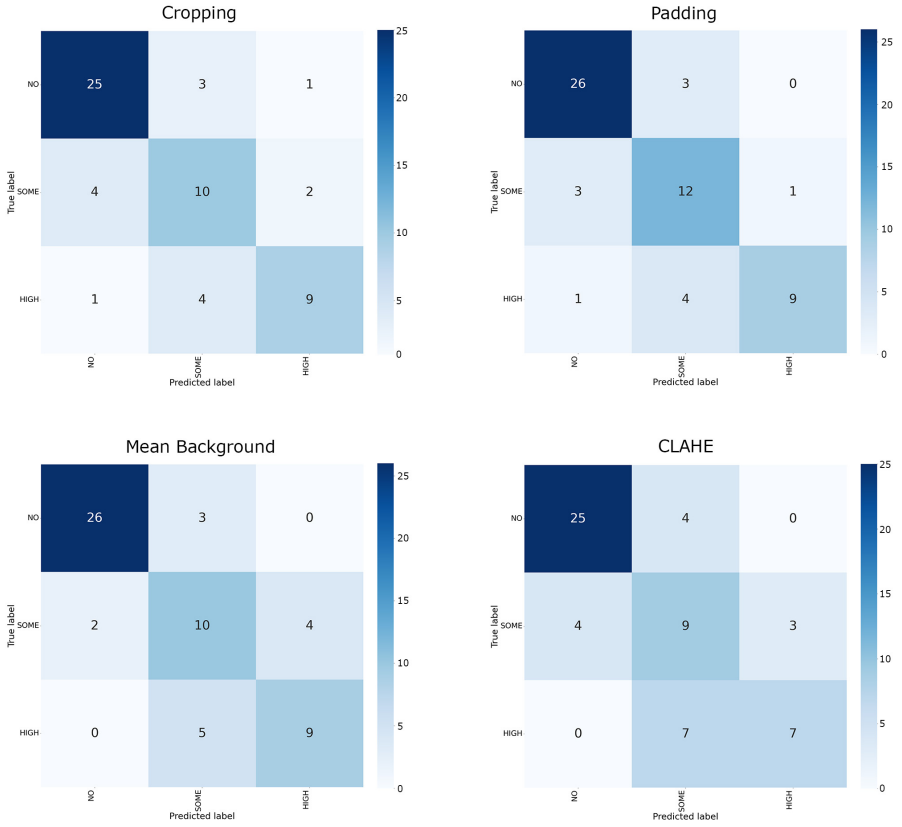


Figure 4.7: Global confusion matrix of validation sets for each preprocessing method.

Table 4.6: Metrics obtained from each one of the global confusion matrixes in Figure 4.7.

Metric		**Cropping**	**Padding**	**MeanBackground**	**CLAHE**
Global Accuracy		74.6	79.7	74.6	69.5
Precision	NO	83.3	86.7	92.6	86.2
	SOME	58.8	63.2	55.6	45.0
	HIGH	75.0	90.0	64.3	70.0
Recall	NO	86.2	89.7	86.2	86.2
	SOME	62.5	75.0	62.5	56.3
	HIGH	64.3	64.3	64.3	50.0
F1-score	NO	84.7	88.1	89.3	86.2
	SOME	60.6	68.6	58.8	50.0
	HIGH	69.2	75.0	64.3	58.3

Table 4.6 shows the results derived from the global confusion matrixes of Figure 4.7 for each of the preprocessing method. The highest accuracy, 79.7% belongs to the Padding preprocessing, while the lowest was 69.5% obtained with images with CLAHE. This result is unexpected, because it was hypothesised that this additional step would improve performance. The most challenging class to correctly predict was, for all methods, the intermediate level of tortuosity ('Some Tortuosity'), the F1-score is as low as 50.0% for this class with CLAHE and has the highest value for the Padding with 68.6%, which is the best preprocessing to predict this class. The 'No Tortuosity' class has the better performance regarding F1-score, ranging from 84.7% (Cropping) and 89.3% (CLAHE), which maybe be due to being the label with most samples. The 'High Tortuosity' class is also a little bit better than the intermediate grading, with lowest value being CLAHE with 58.3% and the best Padding with 75.0 %, which is an unexpected result. Overall for the validation set, since the differences in performance don't follow a strict pattern and are not clearly evident, it is not possible to definitively discern a better preprocessing method based on these results.

Figure 4.8 shows the confusion matrixes obtained by summing each confusion matrix of the five folds for the test set. The results are again satisfactory, but visually there is less demarcation of the intermediate level of tortuosity, which suggests less accuracy in correctly classifying the images as having 'Some Tortuosity'.

Table 4.7 shows the results obtained from the global confusion matrixes for each preprocessing method. The highest accuracy is 74.5% which belongs to the CLAHE preprocessing, while the lowest was 65.5% obtained with images with Mean Background. This result is not in line to what was observed for the validation set. The most challenging class to correctly predict was once again, for all methods, the intermediate level of tortuosity ('Some Tortuosity'). For this class, the F1-score is as low as 38.7% with Mean Background and does not go beyond 55.2% with CLAHE. The 'No Tortuosity' class has the better performance, F1-score ranging from 78.0% with Mean Background and 85.2% with CLAHE, which maybe be due to being the label with most samples.

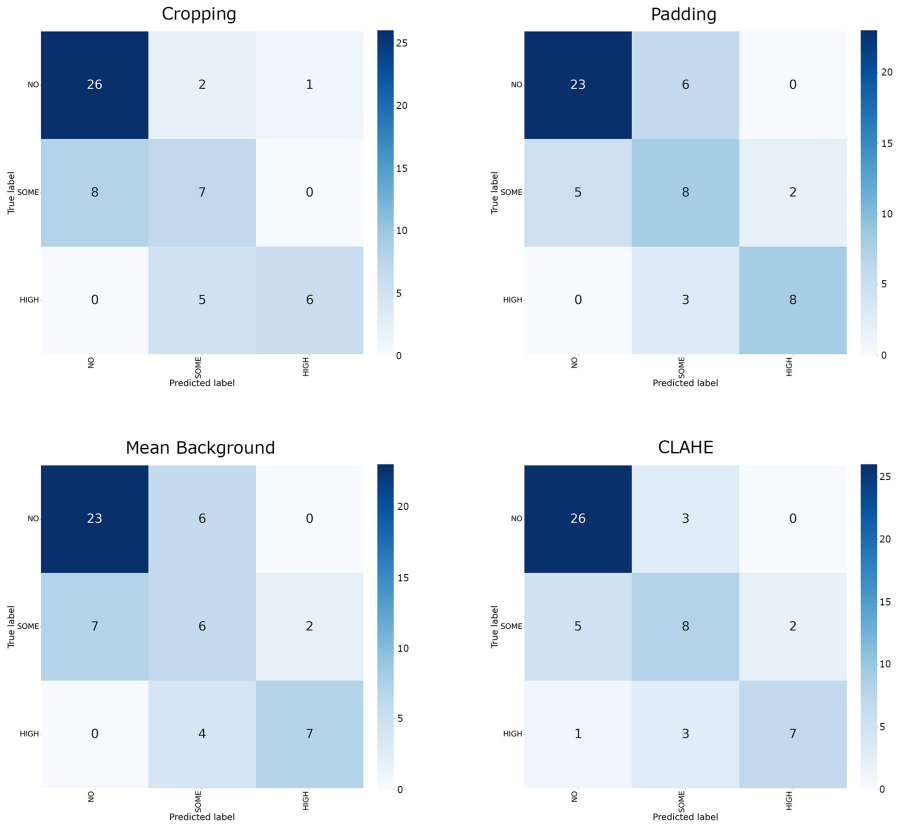


Figure 4.8: Global confusion matrix of test sets for each preprocessing method.

Table 4.7: Metrics obtained from each one of the global confusion matrixes in Figure 4.8.

Metric		**Cropping**	**Padding**	**MeanBackground**	**CLAHE**
Global Accuracy		70.9	70.9	65.5	74.5
Precision	NO	76.5	82.1	76.7	81.3
	SOME	50.0	47.1	37.5	57.1
	HIGH	85.7	80.0	77.8	77.8
Recall	NO	89.7	79.3	79.3	89.7
	SOME	46.7	53.3	40.0	53.3
	HIGH	54.5	72.7	63.6	63.6
F1-score	NO	82.5	80.7	78.0	85.2
	SOME	48.3	50.0	38.7	55.2
	HIGH	66.7	76.2	70.0	70.0

Figure 4.9 shows the four distinct preprocessing techniques applied to a retinal image annotated as not having tortuosity. Models trained on each preprocessed set generated the saliency maps. These maps reveal that each method emphasizes unique aspects of the images: with Padding, the blood vessels are more prominently highlighted, while the Mean Background approach directs attention toward the optic disk. CLAHE, in contrast, shows a more scattered focus, pinpointing various areas across the image.

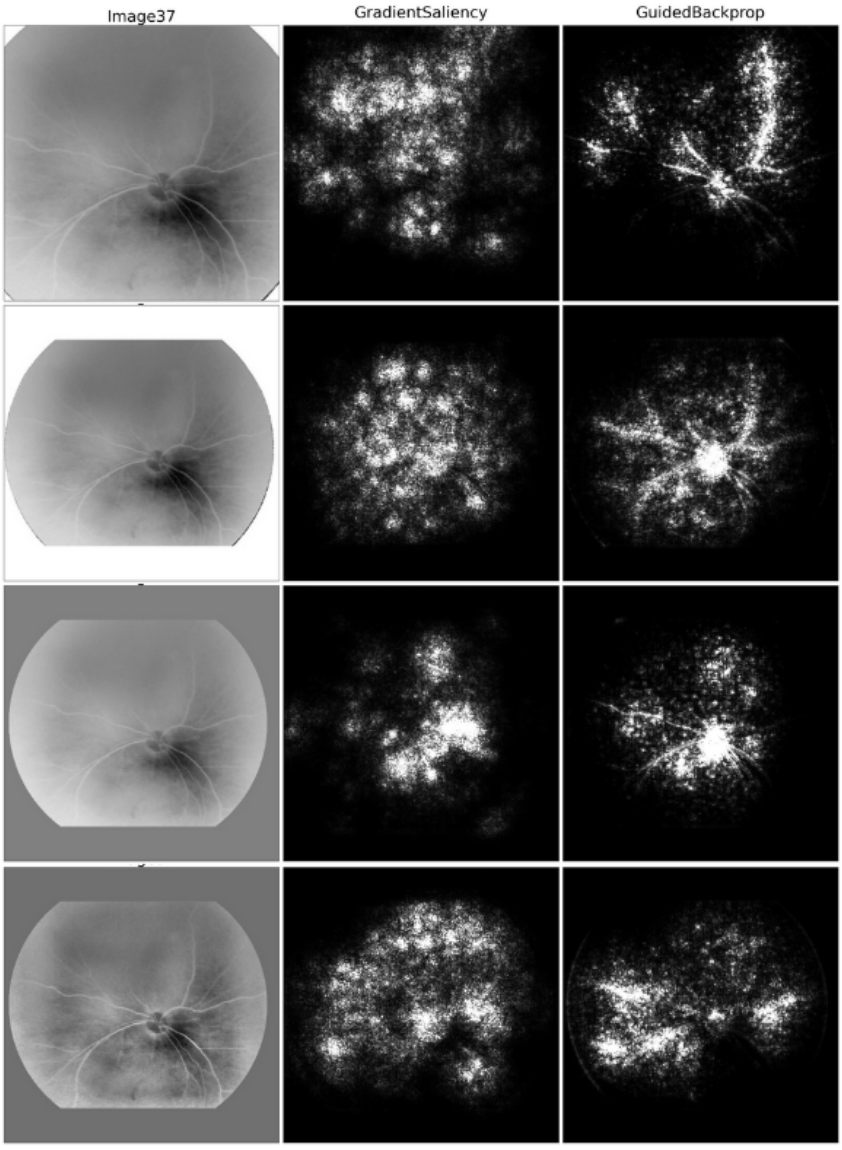


Figure 4.9: Retinal image with 'No Tortuosity' and corresponding saliency maps for each preprocessing method (from top to bottom): Cropping, Padding, Mean Background, and CLAHE.

Figure 4.10 shows that with Cropping the focus of the deep learning model is more restrict to the main branches. For the other techniques the attention changes places in the image. With Padding there is an undesired effect of the deep learning model focusing on the border of the retinal area of the image.

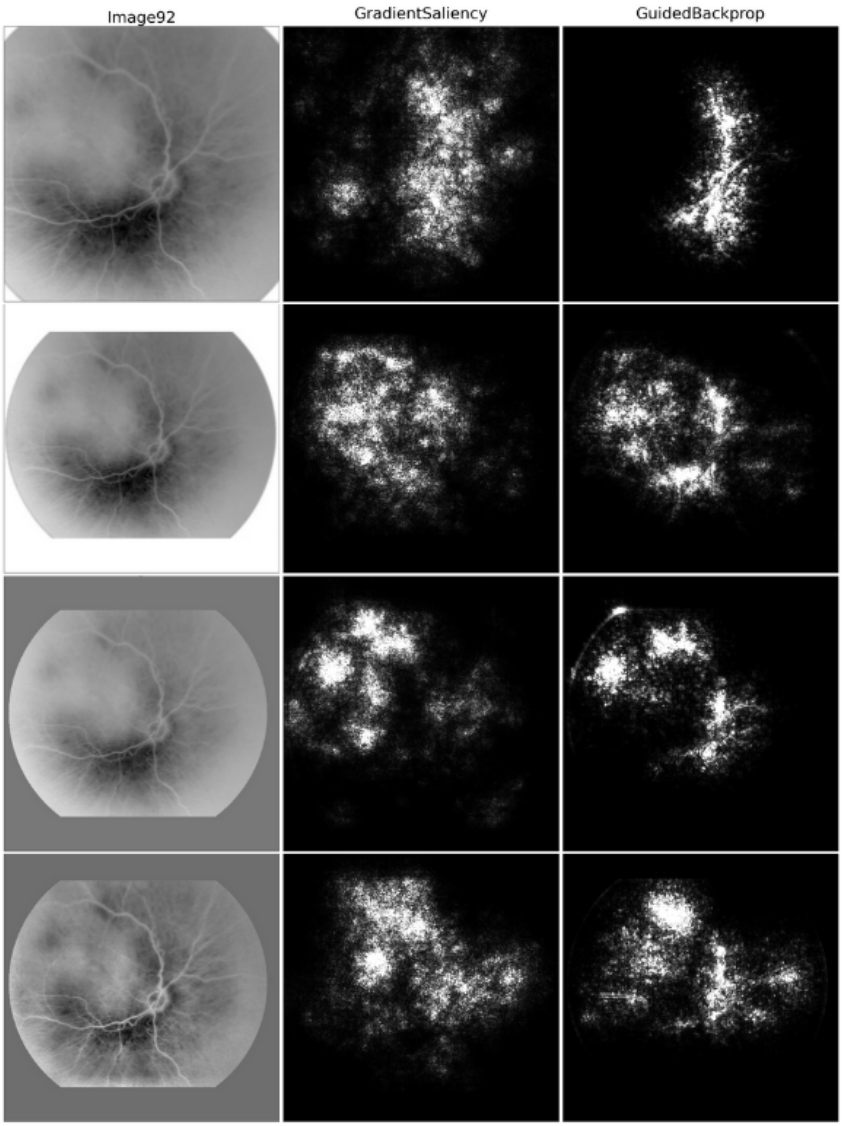


Figure 4.10: Retinal image with 'Some Tortuosity' and corresponding saliency maps for each preprocessing method (from top to bottom): Cropping, Padding, Mean Background, and CLAHE.

Figure 4.11 shows the four distinct preprocessing techniques applied to a retinal image with High Tortuosity. In this case, it's clearly visible that every deep learning model was focusing on the twisted blood vessels. Padding and CLAHE also focus on regions with no blood vessels. The saliency map for Cropping clearly delineates the blood vessels and its tortuosity nature.

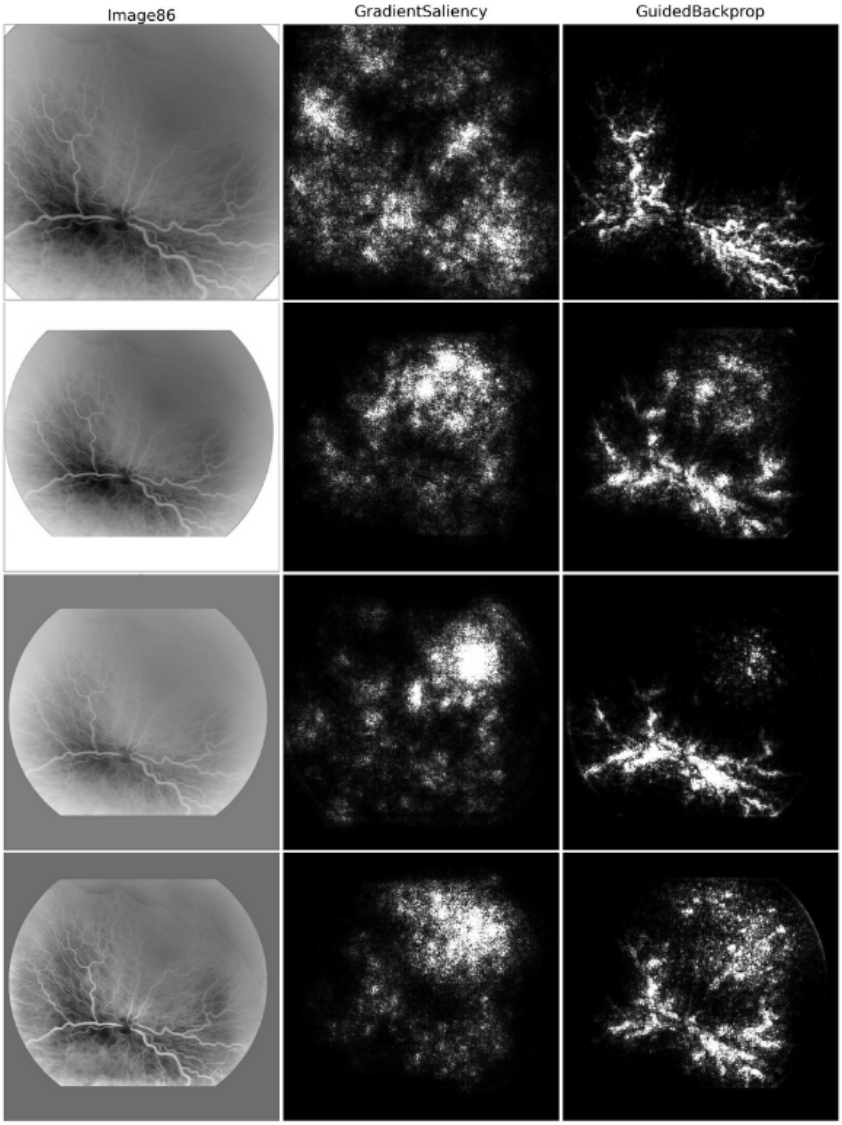


Figure 4.11: Retinal image with 'High Tortuosity' and corresponding saliency maps for each preprocessing method (from top to bottom): Cropping, Padding, Mean Background, and CLAHE.

4.2 JSEIC dataset

The results of the experiments are presented in Table 4.8 and reflect the performance of each metric after applying maximum voting to the predictions. The experiments involved running the 3 different preprocessing techniques and 3 distinct resizing factors, which comprised nine datasets (see section 3.1). In terms of relabeling, the JSEIC dataset presented a challenge, as it only had two distinct classes. This required the relabeling of the 'Some Tortuosity' class (originally labeled as 1) to either 'No Tortuosity' (0) or 'High Tortuosity' (2). Through a combination of visual inspection and model predictions, it became evident that the images misclassified into this middle category generally exhibited a low degree of tortuosity. The data in table 4.8 supports this observation, as metrics improved in nearly all preprocessing and resizing configurations when relabeling class

1 to 0. This suggests that the characteristics of the 'Some Tortuosity' class were closer to 'No Tortuosity', thereby making this relabeling decision more appropriate for the task at hand.

One key finding is that the resizing factor 3 consistently yielded the best overall results for the different preprocessing methods. This was anticipated, as resizing the blood vessel images by this factor made them more similar to the images in the EIARG2 dataset. The EIARG2 dataset had a wider FOV and therefore presented mostly blood vessels of higher caliber.

Table 4.8: Comparison of metrics for three preprocessing methods (Padding, Mean Background and CLAHE) and three resizing factors applied to the JSEIC dataset in percentage (%).

JSEIC	**Metric**	**Factor 1 (0.9)**		**Factor 2 (0.75)**		**Factor 3 (0.6)**	
		1 to 0	1 to 2	1 to 0	1 to 2	1 to 0	1 to 2
Padding	Accuracy	69.2	32.6	88.4	55.7	**92.3**	77.0
	Precision	47.1	29.3	70.3	38.8	**86.8**	54.3
	Recall	100	100	100	100	86.8	100
	F1-score	63.2	44.1	82.4	55.7	**86.8**	72.1
Mean Background	Accuracy	84.6	29.2	**96.3**	46.0	**96.3**	59.1
	Precision	64.4	27.3	100	33.1	100	43.2
	Recall	100	100	86.8	100	86.8	100
	F1-score	78.3	43.5	**92.2**	50.1	**92.2**	57.6
CLAHE	Accuracy	79.7	29.4	87.6	50.1	**94.6**	71.7
	Precision	56.2	27.3	68.2	35.8	**82.4**	48.2
	Recall	100	100	93.1	100	100	100
	F1-score	72.2	43.7	79.1	52.7	**90.2**	65.4

4.3 Results for OCTA Images

Another dataset of a different imaging modality, OCTA, was submitted to the models to grade tortuosity. The dataset was provided and annotated by ophthalmology specialists from the *Centro Hospitalar Universitário de São João*. It has 37 images with 'No Tortuosity', 14 with 'Some Tortuosity' and 6 with 'High Tortuosity'.

In order to test if the models created with retinal fundus images were capable of assessing tortuosity in OCTA images, the five models previously trained on CLAHE images, were applied to these distinct images. Since no preprocessing technique revealed better performance, the choice for the models to be implemented was based on the similarity of the images. CLAHE had most contrast of the blood vessels and could have generated promising deep learning algorithms for OCTA images. Unfortunately, this direct application resulted in all images being predicted as having high tortuosity. This outcome is likely due to the nearly binary nature of the OCTA images, which contain numerous blood vessels.

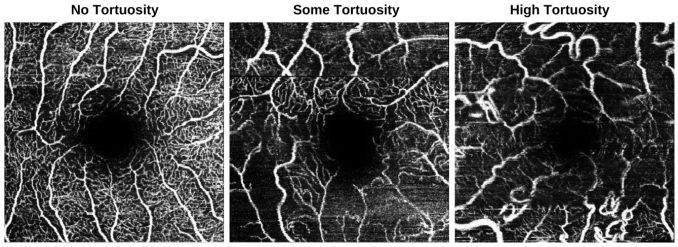


Figure 4.12: Images of 3 classes available in the OCTA dataset: 'No Tortuosity', 'Some Tortuosity' and 'High Tortuosity'.

To address this issue, the next step involved retraining the deep learning models using the same architecture, but with initial weights derived from the fundus image models. This retraining process employed 2-fold stratified cross-validation and applied the same data augmentations previously used (section 3.2.1).

Figure 4.13 is the confusion matrix obtained by adding the confusion matrixes obtained in the two validation folds, after max voting the predictions produced by the 5 models mentioned above.

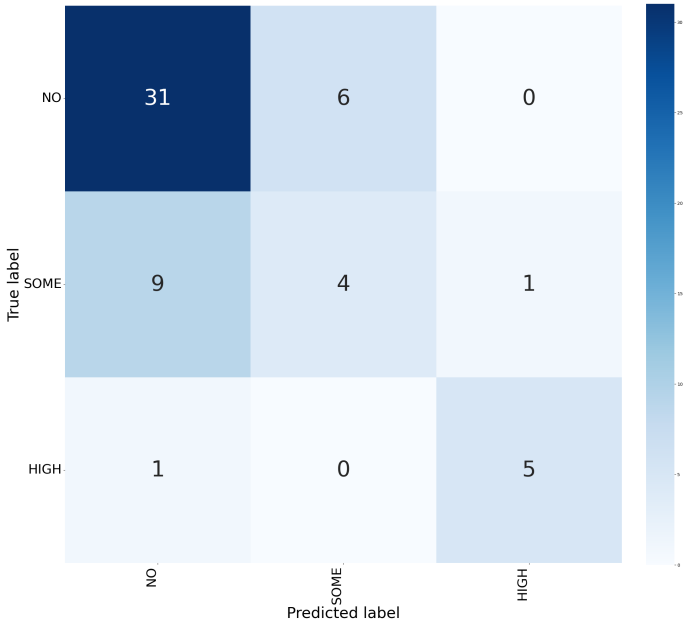


Figure 4.13: Global confusion matrix of validation sets for OCTA images.

Table 4.9 showcases the results regarding the performance of the retrained models. The 'Some Tortuosity' label underperformed and tended to be classified as not having tortuosity, therefore its low F1-score. Overall, the results were positive, especially regarding the extreme classes.

Table 4.9: Metrics obtained from the confusion matrixes in Figure 4.13.

OCTA	Accuracy	Precision			Recall			F1-score		
		NO	SOME	HIGH	NO	SOME	HIGH	NO	SOME	HIGH
	70.2	75.6	40.0	83.3	83.8	28.6	83.3	79.5	33.3	83.3

4.4 Limitations and Challenges

The quality of annotations was a significant challenge. Ground truth annotations should have been established through common sessions among experts to ensure accuracy and consistency. The variability in annotations, stemming from differing expert opinions, affected the robustness of the model and its performance across different data subsets.

The quality of images used for training was another limitation. The dataset, primarily from the EIARG2 collection, included high-caliber blood vessel images with a wide FOV. However, tortuosity often starts in smaller vessels before becoming more pronounced in larger ones. This focus on larger vessels in the training data may have negatively impacted the model's ability to detect early stage tortuosity in smaller vessels. In the future, other datasets that possess a smaller FOV should be used for training. Going towards the standardization in image acquisition would also be very beneficial.

The retinal fundus dataset used for training the deep learning models is very limited in size, consisting of 114 images in total. This small sample size poses a challenge for training generalizable models, as a more extensive dataset is typically required to capture the full variability of retinal features and ensure reliable model performance. Consequently, the restricted dataset underscores the need for caution when interpreting these results and highlights the importance of acquiring a larger and more diverse set of images for future research.

Chapter 5

Conclusions and Future Work

Retinal vessel tortuosity is a biomarker consistently used for diagnosing and monitoring various ocular and systemic conditions. Therefore, accurate evaluation of vessel tortuosity can be useful in early diagnostics, but also in managing disease progression. Medical imaging modalities, such as retinal fundus photography and OCTA, offer valuable insights into vessel structure, and deep learning models have the potential to enhance the precision of tortuosity assessment. This dissertation aimed to explore the effectiveness of the different preprocessing techniques and their impact on deep learning models used for grading tortuosity.

The preprocessing methods included cropping the image, padding, standardizing the background and eroding the edges, and enhancing contrast using CLAHE. The results were underwhelming, as no clear trend emerged, and no single preprocessing method stood out as the best. Additionally, standard deviation between models created with different folds is high, which is likely due to the small sets used for validation and test.

Saliency maps demonstrated that while certain preprocessing techniques, such as Padding, highlighted the blood vessels, CLAHE showed more scattered focus of the deep learning models.

Significant challenges were encountered due to the limited dataset size and quality. The dataset, predominantly featuring larger blood vessels, may not fully represent the early-stage tortuosity often seen in smaller vessels. The results of the models applied to JSEIC dataset (with a smaller FOV) showed that different resizing factors are a determinant aspect in the performance of the models, so an effort should be made to uniformize the acquisition stage and prevent under performance from this component.

The findings suggest that while deep learning models hold promise for tortuosity evaluation, there is a need for larger, more diverse datasets and standardized methodologies to improve model performance and clinical relevance. Future work should focus on expanding the datasets available and their quality, while also exploring other deep learning architectures for this classification task. Further work could also be developed to better the performance on OCTA images.

References

Michael D. Abràmoff, Mona K. Garvin, and Milan Sonka. Retinal Imaging and Image Analysis. *IEEE reviews in biomedical engineering*, 3:169–208, January 2010. ISSN 1937-3333. doi: 10.1109/RBME.2010.2084567. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3131209/.

Mona Ashtari-Majlan, Mohammad Mahdi Dehshibi, and David Masip. Glaucoma diagnosis in the era of deep learning: A survey. *Expert Systems with Applications*, 256:124888, December 2024. ISSN 0957-4174. doi: 10.1016/j.eswa.2024.124888. URL https://www.sciencedirect.com/science/article/pii/S095741742401755X.

Yevgeniya Atiskova, Jan Wildner, Martin Stephan Spitzer, Charlotte Aries, Nicole Muschol, and Simon Dulz. Retinal vessel tortuosity as a prognostic marker for disease severity in Fabry disease. *Orphanet Journal of Rare Diseases*, 16:485, November 2021. ISSN 1750-1172. doi: 10.1186/s13023-021-02080-0. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8605526/.

Jennifer Cano, Shayan Farzad, Maziyar M. Khansari, Ou Tan, David Huang, Jennifer I. Lim, and Mahnaz Shahidi. Relating retinal blood flow and vessel morphology in sickle cell retinopathy. *Eye*, 34(5):886–891, May 2020. ISSN 1476-5454. doi: 10.1038/s41433-019-0604-y. URL https://www.nature.com/articles/s41433-019-0604-y. Publisher: Nature Publishing Group.

Ling-Ping Cen, Jie Ji, Jian-Wei Lin, Si-Tong Ju, Hong-Jie Lin, Tai-Ping Li, Yun Wang, Jian-Feng Yang, Yu-Fen Liu, Shaoying Tan, Li Tan, Dongjie Li, Yifan Wang, Dezhi Zheng, Yongqun Xiong, Hanfu Wu, Jingjing Jiang, Zhenggen Wu, Dingguo Huang, Tingkun Shi, Binyao Chen, Jianling Yang, Xiaoling Zhang, Li Luo, Chukai Huang, Guihua Zhang, Yuqiang Huang, Tsz Kin Ng, Haoyu Chen, Weiqi Chen, Chi Pui Pang, and Mingzhi Zhang. Automatic detection of 39 fundus diseases and conditions in retinal photographs using deep neural networks. *Nature Communications*, 12(1):4828, August 2021. ISSN 2041-1723. doi: 10.1038/s41467-021-25138-w. URL https://www.nature.com/articles/s41467-021-25138-w. Publisher: Nature Publishing Group.

François Chollet. Xception: Deep learning with depthwise separable convolutions, 2017. URL https://arxiv.org/abs/1610.02357.

Miriam Cobo, Francisco Pérez-Rojas, Constanza Gutiérrez-Rodríguez, Ignacio Heredia, Patricio Maragaño-Lizama, Francisca Yung-Manriquez, Lara Lloret Iglesias, and José A. Vega. Novel deep learning method for coronary artery tortuosity detection through coronary angiography. *Scientific Reports*, 13(1):11137, July 2023. ISSN 2045-2322. doi: 10.1038/s41598-023-37868-6. URL https://www.nature.com/articles/s41598-023-37868-6. Publisher: Nature Publishing Group.

Li Dong, Qiong Yang, Rui Heng Zhang, and Wen Bin Wei. Artificial intelligence for the detection of age-related macular degeneration in color fundus photographs: A systematic review and meta-analysis. *eClinicalMedicine*, 35, May 2021. ISSN 2589-5370. doi: 10.1016/j.eclinm.2021.100875. URL https://www.thelancet.com/journals/eclinm/article/PIIS2589-5370(21)00155-3/fulltext. Publisher: Elsevier.

Oana M Dumitrascu and Touseef A Qureshi. Retinal Vascular Imaging in Vascular Cognitive Impairment: Current and Future Perspectives. *Journal of Experimental Neuroscience*, 12: 1179069518801291, January 2018. ISSN 1179-0695. doi: 10.1177/1179069518801291. URL https://doi.org/10.1177/1179069518801291. Publisher: SAGE Publications Ltd STM.

Dumitru Erhan, Yoshua Bengio, Aaron Courville, and Pascal Vincent. Visualizing higher-layer features of a deep network. *University of Montreal*, 1341(3):1, 2009.

Muhammad Moazam Fraz, Alicja R. Rudnicka, Christopher G. Owen, and Sarah A. Barman. Delineation of blood vessels in pediatric retinal images using decision trees-based ensemble classification. *International Journal of Computer Assisted Radiology and Surgery*, 9(5):795–811, September 2014. ISSN 1861-6429. doi: 10.1007/s11548-013-0965-9. URL https://doi.org/10.1007/s11548-013-0965-9.

Rianne J M Goselink, Vivian Schreur, Caroline R van Kernebeek, George W Padberg, Silvère M van der Maarel, Baziel G M van Engelen, Corrie E Erasmus, and Thomas Theelen. Ophthalmological findings in facioscapulohumeral dystrophy. *Brain Communications*, 1(1):fcz023, October 2019. ISSN 2632-1297. doi: 10.1093/braincomms/fcz023. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7425335/.

Enrico Grisan, Marco Foracchia, and Alfredo Ruggeri. A novel method for the automatic grading of retinal vessel tortuosity. *IEEE Transactions on Medical Imaging*, 27(3):310–319, 2008. doi: 10.1109/TMI.2007.904657.

Varun Gulshan, Lily Peng, Marc Coram, Martin C. Stumpe, Derek Wu, Arunachalam Narayanaswamy, Subhashini Venugopalan, Kasumi Widner, Tom Madams, Jorge Cuadros, Ramasamy Kim, Rajiv Raman, Philip C. Nelson, Jessica L. Mega, and Dale R. Webster. Development and Validation of a Deep Learning Algorithm for Detection of Diabetic Retinopathy in Retinal Fundus Photographs. *JAMA*, 316(22):2402–2410, December 2016. ISSN 1538-3598. doi: 10.1001/jama.2016.17216.

Hai-Chao Han. Twisted Blood Vessels: Symptoms, Etiology and Biomechanical Mechanisms. *Journal of Vascular Research*, 49(3):185–197, May 2012. ISSN 1018-1172. doi: 10.1159/000335123. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3369246/.

William E. Hart, Michael Goldbaum, Brad Côté, Paul Kube, and Mark R. Nelson. Measurement and classification of retinal vascular tortuosity. *International Journal of Medical Informatics*, 53(2):239–252, 1999. ISSN 1386-5056. doi: https://doi.org/10.1016/S1386-5056(98)00163-4. URL https://www.sciencedirect.com/science/article/pii/S1386505698001634.

Amanda D. Henderson, Beau B. Bruce, Nancy J. Newman, and Valérie Biousse. Hypertension-related eye abnormalities and the risk of stroke. *Reviews in neurological diseases*, 8(1-2):1–9, 2011. ISSN 1545-2913. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3448945/.

Álvaro S. Hervella, Lucía Ramos, José Rouco, Jorge Novo, and Marcos Ortega. Explainable artificial intelligence for the automated assessment of the retinal vascular tortuosity. *Medical & Biological Engineering & Computing*, 62(3):865–881, March 2024. ISSN 1741-0444. doi: 10.1007/s11517-023-02978-w. URL https://doi.org/10.1007/s11517-023-02978-w.

A.D. Hoover, V. Kouznetsova, and M. Goldbaum. Locating blood vessels in retinal images by piecewise threshold probing of a matched filter response. *IEEE Transactions on Medical Imaging*, 19(3):203–210, 2000. doi: 10.1109/42.845178.

Ahmed Javed, Aishwarya Khanna, Eleanor Palmer, Craig Wilde, Anwar Zaman, Gavin Orr, Dharmalingam Kumudhan, Arun Lakshmanan, and Georgios D Panos. Optical coherence tomography angiography: a review of the current literature. *The Journal of International Medical Research*, 51(7):03000605231187933, July 2023. ISSN 0300-0605. doi: 10.1177/03000605231187933. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC10387790/.

Djibril Kaba, Chuang Wang, Yongmin Li, Ana Salazar-Gonzalez, Xiaohui Liu, and Ahmed Serag. Retinal blood vessels extraction using probabilistic modelling. *Health Information Science and Systems*, 2:2, January 2014. ISSN 2047-2501. doi: 10.1186/2047-2501-2-2. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4376494/.

Aashis Khanal and Rolando Estrada. Fully automated artery-vein ratio and vascular tortuosity measurement in retinal fundus images, 2023. URL https://arxiv.org/abs/2301.01791.

Maziyar M. Khansari, William O'Neill, Jennifer Lim, and Mahnaz Shahidi. Method for quantitative assessment of retinal vessel tortuosity in optical coherence tomography angiography applied to sickle cell retinopathy. *Biomedical Optics Express*, 8(8):3796–3806, July 2017. ISSN 2156-7085. doi: 10.1364/BOE.8.003796. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5560841/.

Fumihiro Kubota, Tetsuyuki Suetsugu, Aki Kato, Fumi Gomi, Seiji Takagi, Takamasa Kinoshita, Hiroto Ishikawa, Yoshinori Mitamura, Mineo Kondo, Chiharu Iwahashi, Soichiro Kuwayama, Yasuo Kurimoto, Yuichiro Ogura, and Tsutomu Yasukawa. Tilted disc syndrome associated with serous retinal detachment: Long-term prognosis. a retrospective multicenter survey. *American Journal of Ophthalmology*, 207:313–318, 2019. ISSN 0002-9394. doi: https://doi.org/10.1016/j.ajo.2019.05.027. URL https://www.sciencedirect.com/science/article/pii/S0002939419302600.

Vijay Kumar and Kolin Paul. Fundus imaging-based healthcare: Present and future. *ACM Trans. Comput. Healthcare*, 4(3), sep 2023. doi: 10.1145/3586580. URL https://doi.org/10.1145/3586580.

Ling-Jun Li, Michael Kramer, Robyn J. Tapp, Ryan E. K. Man, Ngee Lek, Shirong Cai, Fabian Yap, Peter Gluckman, Kok Hian Tan, Yap Seng Chong, Jia Yu Koh, Seang Mei Saw, Yin Bun Cheung, and Tien Yin Wong. Gestational diabetes mellitus and retinal microvasculature. *BMC Ophthalmology*, 17(1):4, January 2017. ISSN 1471-2415. doi: 10.1186/s12886-016-0398-7. URL https://doi.org/10.1186/s12886-016-0398-7.

Huan Liao, Zhuoting Zhu, and Ying Peng. Potential Utility of Retinal Imaging for Alzheimer's Disease: A Review. *Frontiers in Aging Neuroscience*, 10, June 2018. ISSN 1663-4365.

doi: 10.3389/fnagi.2018.00188. URL https://www.frontiersin.org/journals/aging-neuroscience/articles/10.3389/fnagi.2018.00188/full. Publisher: Frontiers.

Ian JC MacCormick, Gabriela Czanner, and Brian Faragher. Developing Retinal Biomarkers of Neurological Disease: An Analytical Perspective. *Biomarkers in Medicine*, 9(7):691–701, July 2015. ISSN 1752-0363. doi: 10.2217/bmm.15.17. URL https://doi.org/10.2217/bmm.15.17. Publisher: Taylor & Francis _eprint: https://doi.org/10.2217/bmm.15.17.

T J MacGillivray, E Trucco, J R Cameron, B Dhillon, J G Houston, and E J R van Beek. Retinal imaging as a source of biomarkers for diagnosis, characterization and prognosis of chronic illness or long-term conditions. *British Journal of Radiology*, 87(1040): 20130832, July 2014. ISSN 0007-1285. doi: 10.1259/bjr.20130832. URL https://doi.org/10.1259/bjr.20130832. _eprint: https://academic.oup.com/bjr/article-pdf/87/1040/20130832/54342003/bjr.20130832.pdf.

Lei Mou, Hong Qi, Yonghuai Liu, Yalin Zheng, Peter Matthew, Pan Su, Jiang Liu, Jiong Zhang, and Yitian Zhao. Deepgrading: Deep learning grading of corneal nerve tortuosity. *IEEE Transactions on Medical Imaging*, 41(8):2079–2091, 2022. doi: 10.1109/TMI.2022.3156906.

Wided Moulahi, Imen Jdey, Tarek Moulahi, Moatsum Alawida, and Abdulatif Alabdulatif. A blockchain-based federated learning mechanism for privacy preservation of healthcare IoT data. *Computers in Biology and Medicine*, 167:107630, December 2023. ISSN 0010-4825. doi: 10.1016/j.compbiomed.2023.107630. URL https://www.sciencedirect.com/science/article/pii/S0010482523010958.

Sabareesh Muraleedharan and Koushik Tripathy. Indocyanine Green (ICG) Angiography. In *StatPearls*. StatPearls Publishing, Treasure Island (FL), 2024. URL http://www.ncbi.nlm.nih.gov/books/NBK580479/.

Rafael Navarro. The optical design of the human eye: a critical review. *Journal of Optometry*, 2 (1):3–18, 2009. ISSN 1888-4296. doi: https://doi.org/10.3921/joptom.2009.3. URL https://www.sciencedirect.com/science/article/pii/S1888429609700184.

Eduardo Novais. The Clinical Utility of OCT Angiography, 2017. URL https://www.reviewofophthalmology.com/article/the-clinical-utility-of-oct-angiography.

James Owler and Peter Rockett. Influence of background preprocessing on the performance of deep learning retinal vessel detection. *Journal of Medical Imaging*, 8:064001, November 2021. ISSN 2329-4302. doi: 10.1117/1.JMI.8.6.064001. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8562352/.

Emma Pead, Atalie C. Thompson, Dilraj S. Grewal, Sarah McGrory, Cason B. Robbins, Justin P. Ma, Kim G. Johnson, Andy J. Liu, Charlene Hamid, Emanuele Trucco, Craig W. Ritchie, Graciela Muniz, Imre Lengyel, Baljean Dhillon, Sharon Fekrat, and Tom MacGillivray. Retinal Vascular Changes in Alzheimer's Dementia and Mild Cognitive Impairment: A Pilot Study Using Ultra-Widefield Imaging. *Translational Vision Science & Technology*, 12(1):13, January 2023. ISSN 2164-2591. doi: 10.1167/tvst.12.1.13. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9838583/.

Hamid Reza Pourreza. EIARG2 database. URL https://eiarg.um.ac.ir/index802d.html?option=com_content&view=article&id=112:eiarg2&catid=14:sample-data-articles.

Lucía Ramos, Jorge Novo, José Rouco, Stephanie Romeo, María D. Álvarez, and Marcos Ortega. Retinal vascular tortuosity assessment: inter-intra expert analysis and correlation with computational measurements. *BMC Medical Research Methodology*, 18(1):144, November 2018. ISSN 1471-2288. doi: 10.1186/s12874-018-0598-3. URL https://doi.org/10.1186/s12874-018-0598-3.

Surabhi Ruia and Koushik Tripathy. Fluorescein Angiography. In *StatPearls*. StatPearls Publishing, Treasure Island (FL), 2024. URL http://www.ncbi.nlm.nih.gov/books/NBK576378/.

Steven S. Saraf, Ariel J. Tyring, Chieh-Li Chen, Thao Phuong Le, Robert E. Kalina, Ruikang K. Wang, and Jennifer R. Chao. Familial retinal arteriolar tortuosity and quantification of vascular tortuosity using swept-source optical coherence tomography angiography. *American Journal of Ophthalmology Case Reports*, 14:74–78, March 2019. ISSN 2451-9936. doi: 10.1016/j.ajoc.2019.03.001. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6425085/.

M. B. Sasongko, T. Y. Wong, T. T. Nguyen, C. Y. Cheung, J. E. Shaw, and J. J. Wang. Retinal vascular tortuosity in persons with diabetes and diabetic retinopathy. *Diabetologia*, 54(9):2409–2416, September 2011. ISSN 1432-0428. doi: 10.1007/s00125-011-2200-y. URL https://doi.org/10.1007/s00125-011-2200-y.

Muhammad Bayu Sasongko, Tien Yin Wong, Kim C. Donaghue, Ning Cheung, Alicia J. Jenkins, Paul Benitez-Aguirre, and Jie Jin Wang. Retinal arteriolar tortuosity is associated with retinopathy and early kidney dysfunction in type 1 diabetes. *American Journal of Ophthalmology*, 153 (1):176–183.e1, January 2012. ISSN 1879-1891. doi: 10.1016/j.ajo.2011.06.005.

Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. *International Journal of Computer Vision*, 128(2):336–359, October 2019. ISSN 1573-1405. doi: 10.1007/s11263-019-01228-7. URL http://dx.doi.org/10.1007/s11263-019-01228-7.

Lukáš Semerád and Martin Drahanský. *Retinal Vascular Characteristics*, pages 309–354. Springer International Publishing, Cham, 2020. ISBN 978-3-030-27731-4. doi: 10.1007/978-3-030-27731-4_11. URL https://doi.org/10.1007/978-3-030-27731-4_11.

Sheng Shi, Xinfeng Zhang, and Wei Fan. A modified perturbed sampling method for local interpretable model-agnostic explanation, 2020. URL https://arxiv.org/abs/2002.07434.

O. Smedby, N. Högman, S. Nilsson, U. Erikson, A. G. Olsson, and G. Walldius. Two-dimensional tortuosity of the superficial femoral artery in early atherosclerosis. *Journal of Vascular Research*, 30(4):181–191, 1993. ISSN 1018-1172. doi: 10.1159/000158993.

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, 2019. ISSN 0030-3992. doi: https://doi.org/10.1016/j.

optlastec.2018.06.061. URL https://www.sciencedirect.com/science/article/pii/S0030399218306637. Special Issue: Optical Imaging for Extreme Environment.

Jost Tobias Springenberg, Alexey Dosovitskiy, Thomas Brox, and Martin Riedmiller. Striving for simplicity: The all convolutional net, 2015. URL https://arxiv.org/abs/1412.6806.

J. Staal, M.D. Abramoff, M. Niemeijer, M.A. Viergever, and B. van Ginneken. Ridge-based vessel segmentation in color images of the retina. *IEEE Transactions on Medical Imaging*, 23(4):501–509, 2004. doi: 10.1109/TMI.2004.825627.

Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions, 2014. URL https://arxiv.org/abs/1409.4842.

Emanuele Trucco, Hind Azegrouz, and Baljean Dhillon. Modeling the tortuosity of retinal vessels: Does caliber play a role? *IEEE Transactions on Biomedical Engineering*, 57(9):2239–2247, 2010. doi: 10.1109/TBME.2010.2050771.

Nicholas J. Tustison, Brian B. Avants, Philip A. Cook, Yuanjie Zheng, Alexander Egan, Paul A. Yushkevich, and James C. Gee. N4itk: Improved n3 bias correction. *IEEE Transactions on Medical Imaging*, 29(6):1310–1320, 2010. doi: 10.1109/TMI.2010.2046908.

UCL. The retina and retinal pigment epithelium (RPE), July 2020. URL https://www.ucl.ac.uk/ioo/research/research-labs-and-groups/carr-lab/bestrophinopathies-resource-pages/eye/retina-and-retinal.

Manuel AP Vilela, Carlos EV Amaral, and Maria Angélica T Ferreira. Retinal vascular tortuosity: Mechanisms and measurements. *European Journal of Ophthalmology*, 31(3):1497–1506, 2021. doi: 10.1177/1120672120979907. URL https://doi.org/10.1177/1120672120979907. PMID: 33307777.

Gengyuan Wang, Meng Li, Zhaoqiang Yun, Zhengyu Duan, Ke Ma, Zhongzhou Luo, Peng Xiao, and Jin Yuan. A novel multiple subdivision-based algorithm for quantitative assessment of retinal vascular tortuosity. *Experimental Biology and Medicine*, 246(20):2222–2229, 2021. doi: 10.1177/15353702211032898. URL https://doi.org/10.1177/15353702211032898. PMID: 34308658.

Jeffrey C Wigdahl, Carla Agurto, Sheila C Nemeth, Vinayak S Joshi, Wendall Bauman, Peter Soliz, and E Simon Barriga. Detection of plus disease in retinopathy of prematurity using automatic vessel tortuosity measurements. *Investigative Ophthalmology & Visual Science*, 58 (8):654, June 2017. ISSN 1552-5783.

Matthew D Zeiler and Rob Fergus. Visualizing and understanding convolutional networks. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part I 13*, pages 818–833. Springer, 2014.