Facultade de Informática
UNIVERSIDADE DA CORUÑA

TRABALLO FIN DE GRAO
GRAO EN ENXEÑARÍA INFORMÁTICA
MENCIÓN EN COMPUTACIÓN

Characterization of age-related ocular diseases in OCT images through deep learning techniques

Estudante: Iván García Fernández

Dirección: Marcos Ortega Hortas

Jorge Novo Buján

A Coruña, September de 2022.

To my family

Acknowledgements

I would like to begin by thanking my advisors, Jorge Novo Buján and Marcos Ortega Hortas, for their recommendations, support, and understanding when external factors challenged the original planning of this project. I would also like to express my gratitude towards researcher Plácido, whose theoretical discussions proved invaluable.

Finally, I acknowledge the work of all those that indirectly contributed to this work, from the researchers and medical practitioners on whose shoulders I stand, to my close ones who made it all possible.

Abstract

Age-Related Macular Degeneration (AMD) is the main cause of severe visual impairment and blindness in Europe, and its prevalence is expected to increase worldwide due to population aging. Optical Coherence Tomography (OCT) is a noninvasive retinal imaging technique that has become the standard of care in the diagnosis and monitoring of late AMD, where the great majority of severe symptoms are manifested. Neovascular late AMD, where new pathological blood vessels are formed that may leak fluid, often results in relatively rapid vision loss. Treatment exists for neovascular AMD, such that its detection and characterization plays a key role in patient outcomes.

This project applies deep learning techniques to the task of AMD characterization. To do so, a data set of OCT scans labeled as to the presence of fluid and neovascularisation is used to train deep convolutional networks. Analysis of this initial experiment produced two hypotheses of performance limiting factors: intra-expert variability and data scarcity. The former was addressed through the development of a machine-assisted review process based on the Class Activation Mapping (CAM) interpretability technique. A small blinded trial was favorable to the methodology. The latter resulted in the adaptation of a large public data set to explore domain-specific transfer learning.

Resumo

A Dexeneración Macular Asociada á Idade (DMAI) é a principal causa de discapacidade visual severa e cegueira en Europa, e espérase que a súa prevalencia aumente a nivel mundial debido ó envellecemento poboacional. A Tomografía de Coherencia Óptica (TCO) é un método non invasivo de imaxe retiniana que se converteu no estándar no diagnóstico e monitorización da DMAI tardía, onde se manifestan a maioría de síntomas graves. A DMAI tardía neovascular, onde se forman novos vasos sanguíneos patolóxicos que poden derramar fluído, a miúdo resulta en perda de visión de forma relativamente repentina. Existen tratamentos para a DMAI neovascular, de modo que a súa detección e caracterización xoga un papel crucial no prognóstico dos pacientes.

Este proxecto aplica técnicas de aprendizaxe profunda á tarefa de caracterización de DMAI. Con ese fin, un conxunto de datos de TCO anotado en base á presenza de fluído e neovascularización foi empregado para entrenar redes convolucionais profundas. A análise deste experimento inicial produciu dúas hipóteses sobre factores que limitan o rendemento: a variabilidade intra-experto e a escaseza de datos. O primeiro foi afrontado mediante o desenvolvemento dun proceso de revisión de anotacións asistido por computadora, baseado na técnica de interpretabilidade *Class Activation Mapping* (*CAM*). Un pequeno estudo cego foi favorable á metodoloxía. A segunda hipótese resultou na adaptación dun gran conxunto de datos público para a exploración de transferencia de aprendizaxe específica ó dominio.

Keywords:

- Deep learning
- Age-related macular degeneration
- Neovascularization
- Interpretability
- Medical imaging
- Optical coherence tomography

Palabras chave:

- Aprendizaxe profunda
- Dexeneración macular asociada á idade
- Neovascularización
- Interpretabilidade
- Imaxe médica
- Tomografía de coherencia óptica

Contents

List of Figures

List of Tables

Chapter 1
Introduction

This first chapter introduces the subject of research by presenting its background, motivation and significance. Subsequently, the scope and objectives of the project are described, along with the structure and division into chapters of this document.

1.1 Motivation

Age-Related Macular Degeneration (AMD) is the main cause of severe visual impairment and blindness in Europe [1], and its prevalence is expected to increase worldwide due to population aging [2]. AMD is a chronic disease of the central retina, with progression into its late stage being responsible for most vision loss symptomatology. Late AMD can be due to pathological new blood vessel formation, called *neovascular AMD* (also known as “wet” or exudative, as blood vessels may leak); or due to *geographic atrophy* (”late dry” AMD). While in geographic atrophy vision deteriorates progressively over many years, neovascular AMD often results in visual impairment over much shorter time frames [3]. In the last two decades, effective treatments for neovascular AMD have been introduced, based on the inhibition of the angiogenic (i.e. blood vessel growth inducing) protein VEGF [4].

AMD is diagnosed on the basis of retinal imaging, traditionally, color photographs of the eye fundus. Angiography is an invasive technique that allows for more precise detection of neovascularization. Through the intravenous injection of a fluorescent dye, an image of the blood vessels of the choroid, situated behind the retina, and possible leakage is obtained.

More recently, Optical Coherence Tomography (OCT) has emerged as an essential adjunct for diagnosis and monitoring, especially in light of anti-VEGF therapy [5]. It is a non-invasive technique that provides high resolution cross-sectional imaging of the retinal layers and underlying choroid. However, OCT image interpretation is a labor-intensive task that places great demands on ophthalmologists. As opposed to angiography or fundus examination, OCT generates a large amount of volumetric data. Furthermore, accurate monitoring of the progression of the condition of a patient requires consistency across time and between different experts.

To that end, the techniques of deep learning have recently been used to develop classification systems that support the diagnostic process [6]. Unlike human experts, machine learning systems do not suffer from fatigue when examining large amounts of data, and their predictions can be completely deterministic. Nevertheless, these systems often provide little justification for their decisions. The specific situation of a patient must be considered holistically, as clinical examination, current medication and complementary diagnostic techniques can alter the interpretation of otherwise ambiguous OCT images. Through the use of *interpretability* techniques, a deep learning model can help the ophthalmologist integrate the predictions of the machine into the wider context of the patient.

1.2 Problem statement

Given the desirability of accurate, consistent and interpretable models for OCT image classification, the problem consists of their construction and evaluation. Within the paradigm of supervised deep learning, a suitable neural network architecture is trained on a representative data set of OCT scans annotated by human experts.

Convolutional Neural Networks (CNNs) are highly general models for computer vision that have seen wide use in medical imaging, including ophthalmology. Their popularity across domains has encouraged the development of techniques to better understand their behavior. Concretely, Class Activation Maps (CAMs) provide human-understandable visualizations of which parts of an image contributed to each possible class. These characteristics make CNNs a good fit for our domain.

In general, the second part of the problem consists of the use of interpretability techniques to analyze the resulting model and the method that generated it. Together with the usual validation methods of machine learning, these techniques can help determine if the modeling capacity of the architecture is sufficient, as well as investigating the completeness and consistency of the data.

That second aspect, the feasibility of the training data, is recognized as crucial in what has recently been termed *data-centric AI*. Fundamental AI research benefits from consistent data sets in order to compare approaches, and data is sometimes deliberately restricted to investigate generalization capacity. Conversely, the application of AI focuses on absolute performance, which is heavily influenced by the data set and, in turn, by the procedure through which it is constructed.

In many important medical domains such as OCT characterization, performance is limited by our ability to obtain large, *accurate* data sets. Notice that the same factors, such as fatigue, that affect clinical diagnosis also apply to data set labeling. In fact, labeling accuracy has recently become a limiting factor even in foundational challenges like the Imagenet data set [7], where before it was not such a major factor [8]. This suggests that addressing label noise may be an important part of the problem of constructing an adequate OCT image classification model.

1.3 Research objectives

The main objective of this project is to research, implement and validate deep learning techniques for the construction of *computer-aided diagnosis* systems, concretely through the task of detection and characterization of exudative AMD in OCT images.

This objective is decomposed into iterative sub-goals that allow for the problem to be approached incrementally. The first of them is to establish the general viability of AMD characterization in a limited data set. The initial task will consist of binary classification, as in the presence or absence of signs of wet AMD. On the model side, proven architectures such as the *ResNet* and the *DenseNet* will be employed. Training itself will also follow an iterative approach, with the establishing of simple baselines to be superseded by more complex techniques, such as transfer learning or data augmentation.

If and when the evaluation of the performance of the model is satisfactory, the detection will be refined to the independent recognition of fluid and neovascularization. If this classification were also to produce adequate results, a finer distinction would be made between their sub-types according to their location with respect to retinal layers.

However, working with a very limited number of annotated examples is expected to pose challenges even to state-of-the-art models and training methodologies. Furthermore, it is often difficult to determine what inadequate performance is to be attributed to. This constitutes another sub-goal, where we will explore the detection of these inadequacies and their rectification through techniques such as *data augmentation*, domain-specific and general *transfer learning* and our proposed model-assisted intra-expert variability mitigation.

1.4 Outline

This report is structured according to the goals of the project and the steps and experiments that were performed to that effect. Therefore, chapters are a reflection of the trajectory of the project, from its introduction and contextualization to its final conclusion. The main chapter, Methodology, is simultaneously chronologically and logically ordered, following the progression of experiments. Such structure ultimately results in the following division into 5 chapters:

Chapter 1: Introduction The present chapter, which describes, motivates and provides this outline for the project.

Chapter 2: Contextualization A concise introduction to the domain of study and the techniques applied to it. In both cases it provides a broad overview, beginning from the anatomy of the human eye and machine learning respectively, and systematically narrows its focus towards the fields specific to the project.

Chapter 3: Planning Describes the organization of the project, including tasks and scheduling as well as the management of resources and estimated costs.

Chapter 4: Methodology The largest chapter in the document, it is divided according to the experiments performed throughout the project:

- **Detection of nAMD signs**
- **Model-assisted intra-expert variability mitigation**
- **Domain-specific pretraining**
- **nAMD characterization**

Chapter 5: Conclusion and future work Summarizes the approaches and results of the project. Includes a discussion of the general applicability of the methods and possible lines of future work.

Chapter 2

Contextualization

An adequate presentation of the techniques, results and challenges encountered in this project requires working knowledge of both the specifics of AMD diagnosis and the Deep Learning paradigm. To that end, a short introduction to the domain is provided, covering the basics of the relevant anatomy, imaging techniques, and how they can be used to characterize AMD. Following, there is a brief overview of the fundamentals of Deep Learning, with a strong focus on concepts relevant to the specific architectures used.

2.1 Domain description

This section provides an overview of the physiology of human vision, describing the alterations that characterize AMD. It also introduces OCT, the imaging technique on which this project is based.

2.1.1 The human eye

The human eye enables vision through photoreceptor cells, which detect visible light and generate corresponding electrical neural pulses. The rest of the eye supports this function in several ways, such as optically processing incoming light, supplying photoreceptors vascularly and propagating neural activations to the visual cortex of the brain.

Anatomically, the eye is an almost spherical organ, with a transparent protrusion on the front called the cornea (Figure 2.1). Its principle of operation is that of a *camera obscura*: an opaque chamber with a small opening, called the *aperture*, through which light enters. The aperture of the human eye is the pupil, the hole in the center of the iris, situated behind the cornea. The smaller the aperture, the more precisely light reaching each point of the interior is approximated by a single ray that passes through the center of the aperture. A correspondingly sharp inverted image of the outside scene is projected on the posterior of the eye.

However, at very small aperture sizes diffraction effects dominate, and a smaller aperture results in a blurrier image. Crucially, the reduction in the size of the aperture is also accompanied by a decrease in the intensity of the image, resulting in a trade-off between spatial resolution and sensitivity. To address it, the iris contracts and dilates the pupil in response to changes in light intensity.

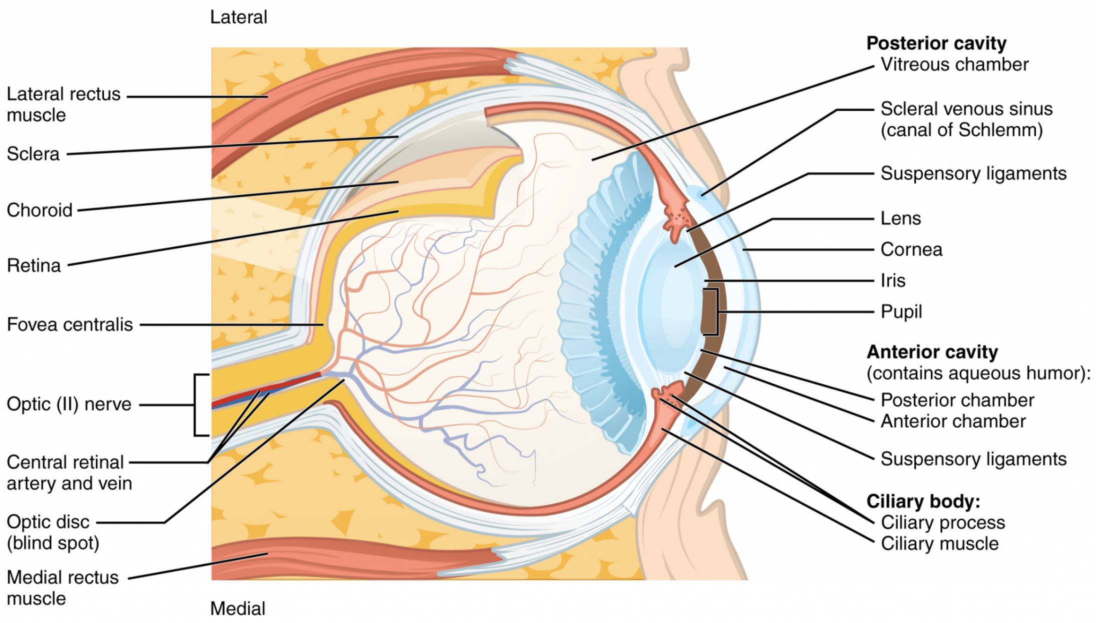


Figure 2.1: Structure of the eye [9].

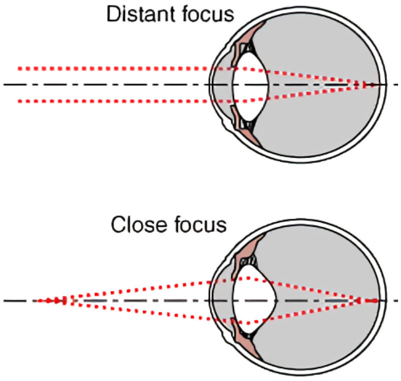


Figure 2.2: Accommodation. Adapted from [10].

Both man made cameras and the human eye employ lenses, which converge light to create a sharper image. The distance at which objects form the sharpest image depends on the focal length of the lens (a measure of how much it converges light) and its distance to the sensor. The latter can be adjusted mechanically in man made cameras. In the human eye the cornea acts as the main lens, providing most optical power, and focal length is adjusted through a process known as accommodation (Figure 2.2). The contraction of the ciliary muscle regulates the shape of the crystalline lens, situated behind the iris, adjusting the combined focal length of the cornea and the lens [11].

The rest of the ocular globe, the posterior segment, can be divided into three layers that enclose the vitreous body and are crossed by the optic nerve. The innermost tissue, the retina, contains the photoreceptor cells. These cells perform phototransduction, the detection of visible light and the corresponding generation of an action potential. The electrical potential then travels inwards through the retina towards the optic nerve and, eventually, to the visual cortex of the brain. The next layer, the choroid, surrounds the retina and supplies it vascularly. Finally, the choroid is enclosed by the sclera, connective tissue with a protective function that is primarily composed of collagen.

2.1.2 Anatomy of the retina

The anatomy of the retina can be presented through two complementary perspectives. One describes the back of the eye, the *eye fundus*, as viewed through an ophthalmoscope. The other corresponds to the layered structure of the retina, as can be seen in histologic sections of retinal tissue and, as will be discussed in the next section, through OCT.

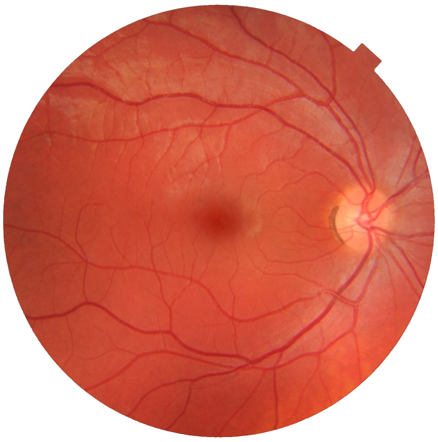


Figure 2.3: Fundus photograph of a normal right eye [12].

The fundus image in Figure 2.3 is centred on the *fovea*, the dark, avascular region which corresponds to the centre of the visual field. However, the fovea is not aligned with the approximate equivalent of an optical axis for the eye. It is often displaced temporally, on average 4 degrees, but with substantial variation [11].

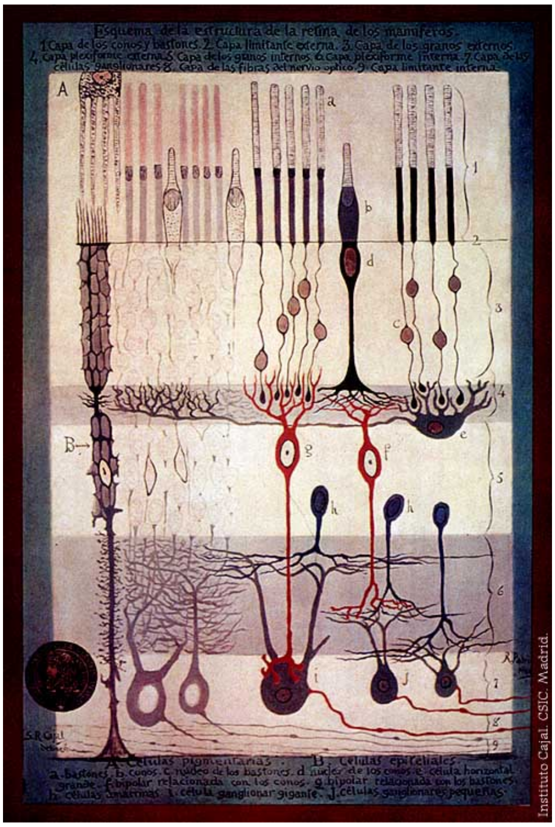
The *optic disc* can be seen as the bright spot situated approximately 17 degrees (4.5-5 mm) nasally to the fovea. Measuring around 2mm vertically and a little less horizontally, it constitutes the termination of the optic nerve in the retina and the source of blood vessels that supply it [13]. Its central location in the retina corresponds to the blind spot in the visual field, as it connects to the nerve fibers on top of the retina.

The *macula lutea*, after which AMD is named, is the pigmented area that surrounds and includes the fovea. Its characteristic yellow hue is due to the reflection of the xanthophyll carotenoids zeaxanthin and lutein. These pigments protect the central retina from short wavelength radiation such as the ultraviolet [13].

The macula is an example of how the structure of retinal tissue varies radially, from the center of the fovea to its edge, the *ora serrata*, close to the ciliary body (Figure 2.1). These changes are responsible for the varying visual acuity from central to peripheral vision, due to alterations of the structure below the surface.

The histological perspective of the retina reveals, as was noticed by Santiago Ramón y Cajal in the 1890s, the different kinds of cells that compose retinal tissue and their arrangement into three main layers of nerve cell bodies. The photoreceptors are situated in the back, behind, among others, bipolar neurons, and the ganglion cells whose axons connect to the optic nerve. This configuration is called an *inverted retina*, because light has to traverse several layers before reaching photoreceptors, and is responsible for the blind spot.

Figure 2.4 demonstrates the division into 9 layers of the neurosensory retina that is used today. The photoreceptors are separated by an outer limiting "membrane" into their nuclei and the segment layer. Inner segments contain mitochondria and other organelles, while outer segments contain opsins, light sensitive proteins.


Esquema de la estructura de la retina de los mamiferos.
1. Capa de los conos y bastones. 2. Capa limitante externa. 3. Capa de los granos externos. 4. Capa plexiforme externa. 5. Capa de los granos internos. 6. Capa plexiforme interna. 7. Capa de las células ganglionares. 8. Capa de las fibras del nervio optico. 9. Capa limitante interna.
A
a
b
c
d
e
f
g
h
i
J
B
1
2
3
4
5
6
7
8
9
S. R. Cajal
A. Células pigmentarias. B. Células epiteliales.
a. bastones. b. conos. c. núcleo de los bastones. d. nucleo de los conos. e. célula horizontal grande. f. bipolar relacionada con los conos. g. bipolar relacionada con los bastones. h. células amacrinas. i. célula ganglionar gigante. J. células ganglionares pequeñas.

Instituto Cajal. CSIC. Madrid.

Figure 2.4: Structure of the Mammalian Retina. Ramón y Cajal c. 1900.

There are two types of photoreceptors in these layers, *rods* and *cones*, named after their outer segment shapes. Three types of cones with opsins that respond to different wavelengths enable color vision. Despite their abundance, the output of several rods converges onto the same neuron in the retina, improving sensitivity but not resolving fine detail. The variation in the density of rods and cones on the retinal mosaic is a cause of visual acuity differences between central and peripheral vision (Figure 2.5).

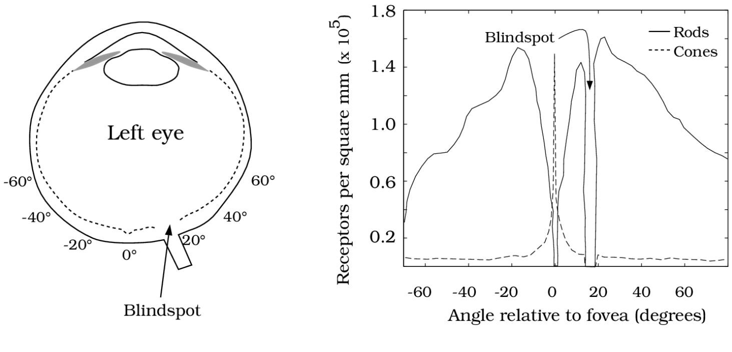


Figure 2.5: The distribution of rods and cones. Adapted from [14].

Figure 2.6 illustrates that between the photoreceptor and the inner nuclear layers, the latter containing horizontal, bipolar and amacrine neurons, and Müller glia; there is a neuropil, an area where they synaptically connect, called the outer plexiform layer. Similarly, between the inner and ganglion layers there is another neuropil, the inner plexiform layer. Finally, the axons of the ganglion cells form the nerve fiber layer that is separated from the vitreous by the inner limiting membrane.

More recently, a third type of photoreceptor has been discovered. A small subset of ganglion cells utilize another opsin based photopigment, melanopsin, and contribute to the regulation of circadian rhythms and the pupilary light response [15]. These intrinsically photosensitive retinal ganglion cells have also been found to contribute to conscious visual awareness [16].

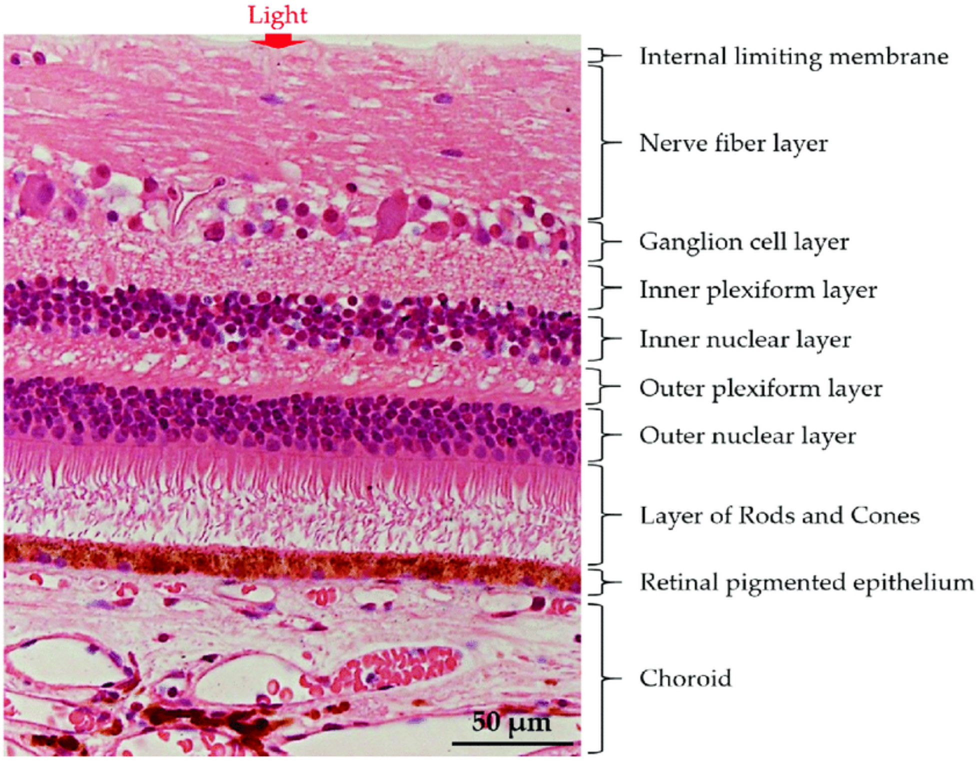


Figure 2.6: Histology of the retina [17].

The neural retina is separated from the choroid by the Retinal Pigment Epithelium (RPE), a single layer of pigmented epithelial cells. Along with providing immune privilege to the retina, supplying nutrients to photoreceptors and regulating their phototransduction chemically, the RPE absorbs the light that has traversed the retina. Several pigments are involved in this process, with melanin accounting for most of the light absorption [18].

2.1.3 OCT

First employed to obtain *in vivo* images of the retina in 1993, Optical Coherence Tomography (OCT), is a noninvasive imaging technique whose unique ability to provide near histologic resolution images of the retina has made it the standard of care in ophthalmology [19].

OCT is a technology analogous to ultrasound, in the sense that it provides images based on the magnitude and delay of reflected signals. The difference is that while ultrasound uses sound waves, OCT uses light. This poses several challenges, as light travels the depth of the retina in around 1 picosecond (one trillionth of a second).

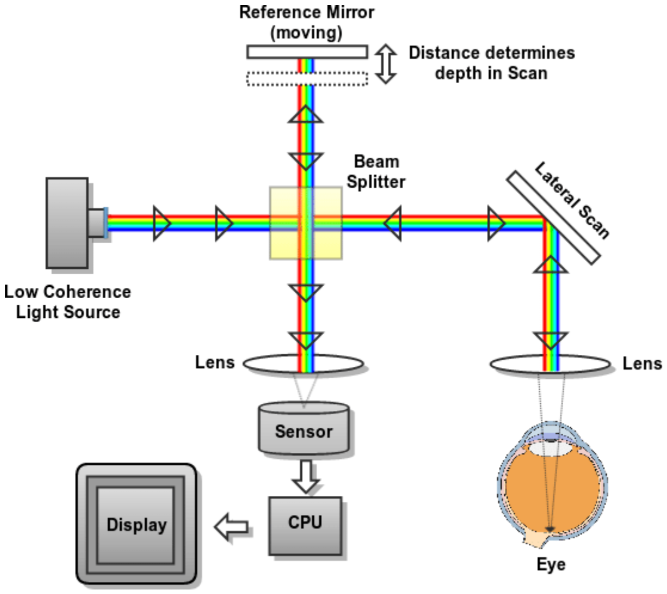


Figure 2.7: Schema of a basic Optical Coherence Tomography (OCT) acquisition system [20].

Initial attempts at ‘optical ultrasound’ utilized femtosecond lasers, but the key to the scalability of OCT is in its use of low coherence interferometry. Figure 2.7 demonstrates the Michelson interferometer configuration. In the case of OCT, low coherence light is split into two arms, one being the retina and the other a reference mirror that adjusts its distance to match that of the depths to be imaged. Some of the light reaching the retina is backscattered and interferes with the reference at the beam splitter, eventually reaching a sensor.

In the first OCT systems, the sensor is a photodetector, performing *time domain OCT*. For any given reference mirror position, only points that are situated around that distance in the retina, within the coherence length of the light, are the source of interference with the reference arm. By scanning the depth of the mirror, an *A-scan* can be obtained. Like in ultrasound, an *A-scan* is a plot of the intensity of backscattered light along a line. In OCT, the line is perpendicular to the retina. By scanning the eye arm laterally, a histologic-cut-like image can be obtained (Figure 2.8), called a *B-scan*.

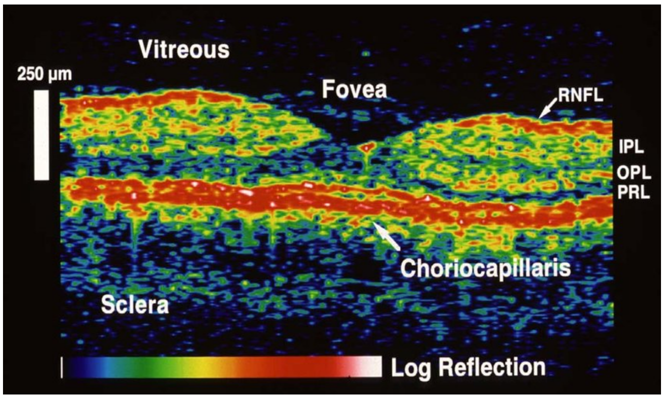


Figure 2.8: First *in vivo* OCT image of the normal retina in a human subject [19].

To improve imaging speeds and sensitivity, Fourier Domain OCT (FD-OCT) was developed. In Spectral Domain OCT (SD-OCT), a subtype, a spectrometer simultaneously measures all echoes of light, eliminating the need to scan the reference arm. This accelerates image acquisition substantially, which in turn improves motion artifacts and sensitivity. A second subtype of FD-OCT, Swept Source OCT (SS-OCT), trades the spectrometer for a frequency-swept laser, allowing the frequency-specific interference to be measured across time [19].

Another development of OCT is the use of several parallel *B-Scans* to form a *C-scan*, a volumetric reconstruction of the retina. The data can then be used to generate *en face* images of the retina like those of an ophthalmoscope. Furthermore, by measuring changes in the image that are not the result of patient or device movement, blood flow is detected. This allows for the creation of images similar to those of *fluorescein angiography*, in what is termed Optical Coherence Tomography Angiography (OCTA).

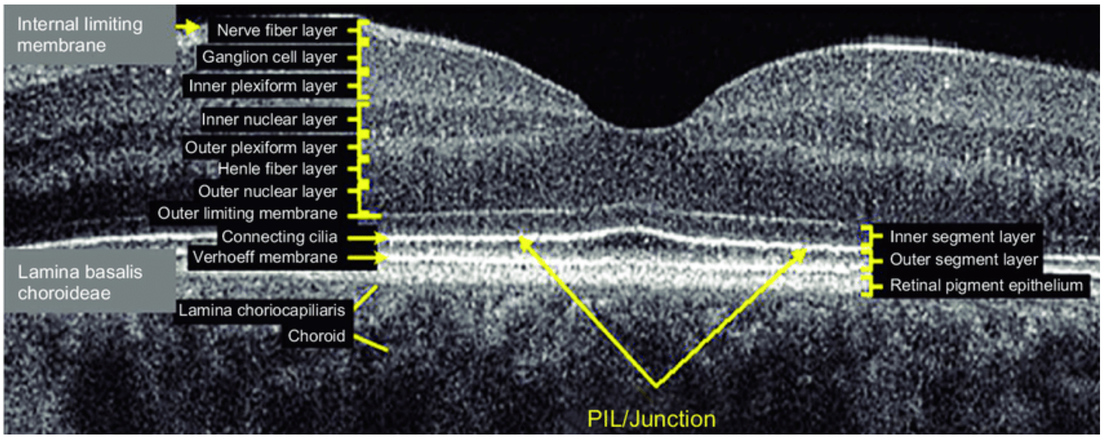


Figure 2.9: Layers of the retina as they appear on a SD-OCT B-scan [21].

As we can observe in Figure 2.9, the histological features of the retina (Figure 2.6) are well correlated to OCT B-scans images. Moreover, the morphological alterations to retinal histology that are associated with AMD also produce recognizable features in OCT images. Hence, they can be use to detect and characterize AMD.

2.1.4 Characterizing AMD

AMD is clinically classified into three stages , according to its progression [22]. When drusen appear of a size beyond what are considered normal aging changes, a patient is considered to have early AMD. The appearance of either large drusen or pigmentary abnormalities is indicative of the intermediate stage of AMD. These stages are designed to assess clinically relevant increases in the risk of the late form of AMD and RPE detachment, which are responsible for most severe symptomatollogy [22].

Late AMD itself can in turn be characterized into Geographic Atrophy (GA) or Neovascular Age-Related Macular Degeneration (nAMD). nAMD represents less than 10% of total AMD cases, but is responsible for the majority of cases of severe visual loss, and progresses more rapidly [23]. Within nAMD, it is fluid that causes symptoms and needs urgent treatment, however, a study found that neovascularization was present in 88% of eyes that would then go on to present exudation [24]. There exist nAMD-specific biomarkers in OCT images that can be of great utility for the monitoring of the progression and treatment of a patient [23]. In the following, the consensus characterization of nAMD is briefly presented, as well as its OCT imaging correlates.

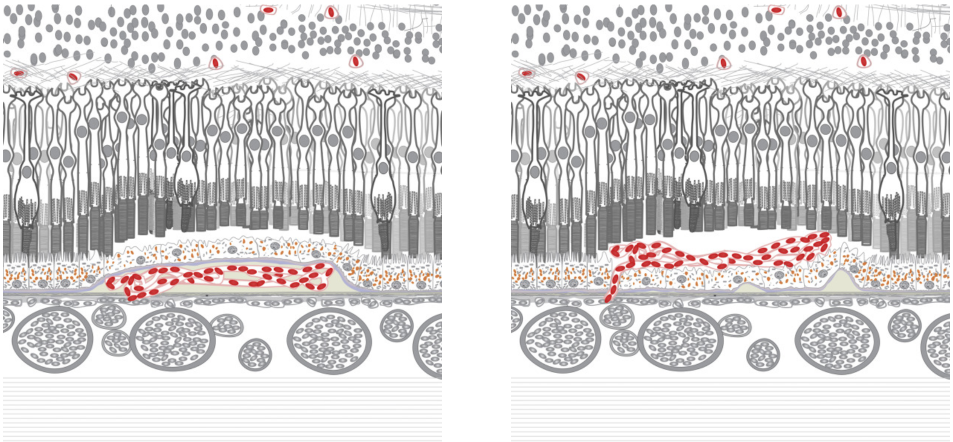
Neovascularization

In nAMD, neovascularization is the invasion by vascular and associated tissues into the outer retina, subretinal space or sub-RPE space in varying combinations [25]. Thus, three distinct types of neovascularization are recognized, according primarily to their location.

In type 1 neovascularization (Figure 2.10a), blood vessels from the choriocapillaris layer of the choroid grow into the sub-RPE space. The accumulation of fibrotic tissue may cause a detachment of the RPE, resulting in an elevated region in OCT B-scans.

Type 2 neovascularization (Figure 2.10b) implies the crossing of the RPE by the vessel growth and its development into the sub-retinal space. Nevertheless, the majority of fibrovascular proliferation occurs in the sub-rpe space from which growth is originated.

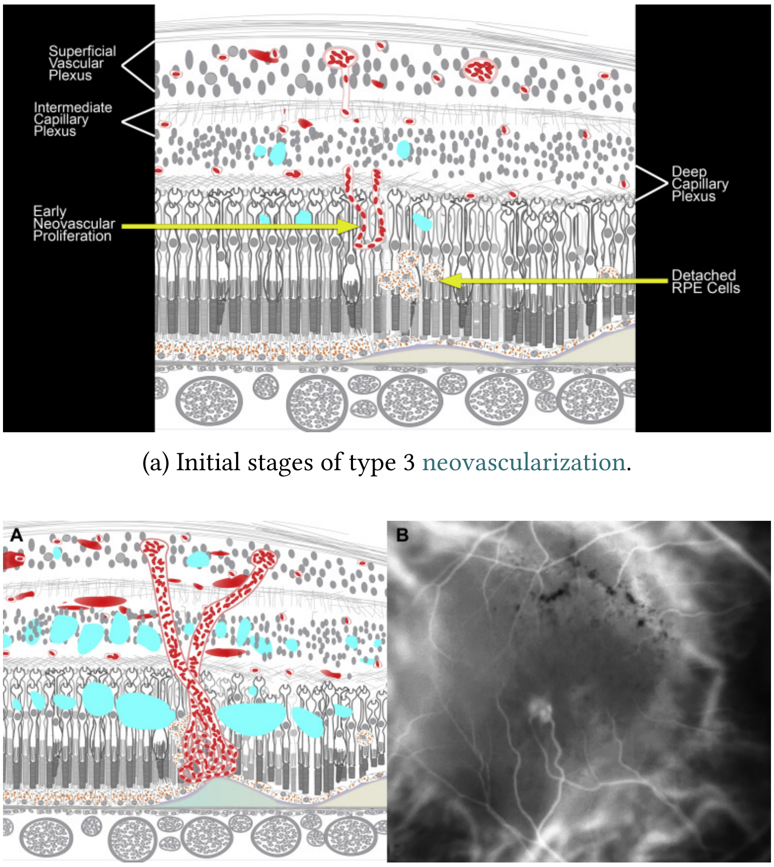
Finally, type 3 neovascularization (Figure 2.11) refers to vascular proliferation that originates from the superficial vascular plexus in the inner retina, as opposed to types 1 and 2 that originate from the choroid. However, the vascularization might progress towards and cross the RPE, resulting in an angiomatous lesion.



(a) Type 1 neovascularization.

(b) Type 2 neovascularization.

Figure 2.10: Diagrams comparing type 1 and 2 neovascularization, highlighting their location relative to the RPE [25].


Superficial Vascular Plexus
Intermediate Capillary Plexus
Early Neovascular Proliferation
Deep Capillary Plexus
Detached RPE Cells
A
B

(a) Initial stages of type 3 neovascularization.

(b) Angiomatous lesion resulting from type 3 neovascularization (A) and angiography of a patient with type 3 neovascularization (B).

Figure 2.11: Progression of type 3 neovascularization [25].

Fluid

Biomarkers have been researched as a means of predicting the long term outcome of AMD patients, e.g. fibrovascular RPE detachment as a marker of nAMD [26]. These markers can be divided into two distinct categories: structural features of the choroidal and retinal layers, and the distribution of fluid [23]. The appearance of fluid on OCT and fundus imaging is presented in Figure 2.12.

A distinction is made between Intraretinal Cystoid Fluid (IRC), Subretinal Fluid (SRF) and Sub-Retinal Pigment Epithelium Fluid (sub-RPE) [23]. Figure 2.13 demonstrates the different types and the association of sub-RPE with RPE detachment.

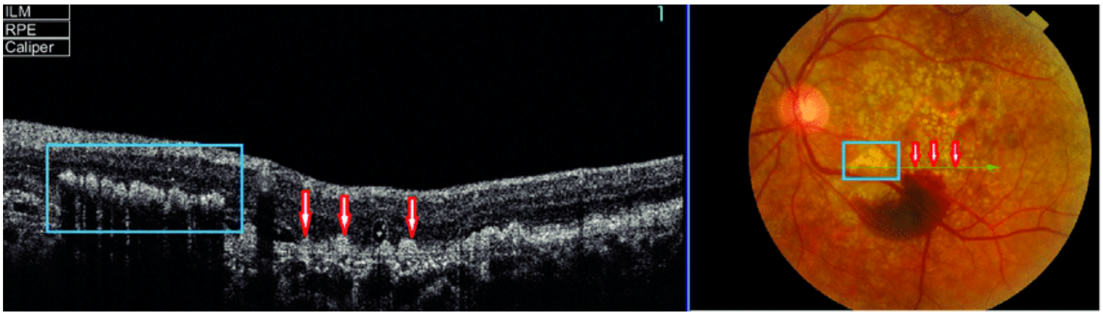


Figure 2.12: 'Fundus photograph (right) showing two different types of deposit: exudates (blue rectangle) and drusen (red arrow) in the left eye of a patient with wet ARMD; the B-scan line on the fundus photograph has the same width as the B-scan SD-OCT image (left)' [21].

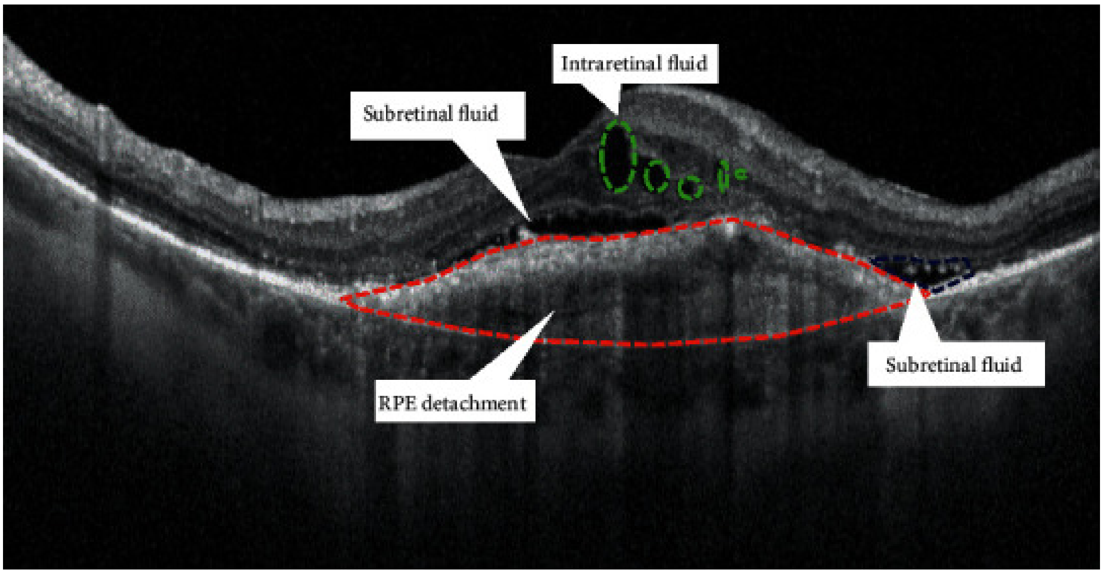


Figure 2.13: fluid distribution in nAMD, together with RPE detachment [23].

2.2 Machine Learning

Machine Learning (ML) is the study of computer algorithms that improve automatically through experience [27]. The field has its roots in artificial intelligence, where it has been considered a promising and increasingly necessary approach since its founding [28].

In the last decade, machine learning has experienced a surge in popularity within and beyond AI research. The availability of large data sets, together with the computing power and algorithms necessary to utilize them, has enabled machine learning approaches to surpass explicitly coded solutions in many commercially important domains, such as computer vision and natural language processing.

Such advances have had revolutionary consequences in many applications, of which medicine is no exception. Medical research can be seen as an extremely complex learning task, where data such as symptoms, diagnostic tests and available interventions are utilized to construct models of the body, its disease and treatment. Similarly, medical practice can be seen from a learning perspective, where diagnosis and treatment are predictions made from data.

The *deep learning* revolution is posed to rapidly increase the scope of the influence of ML on medicine. OCT constitutes a great example of this paradigm, as its scans generate great volumes of data, which pair perfectly with deep learning techniques that ingest it with minimal preprocessing.

2.2.1 Neural networks and backpropagation

Machine learning is a broad field with many techniques, however, a certain class of models has risen to prominence during the deep learning revolution: *neural networks*. Neural networks consist of many simple, connected processors called neurons [29]. Each neuron computes a real-valued output, called its activation, from one or more inputs. Those neurons whose inputs come from the environment are called input neurons; the input of the rest is the output of other neurons. The output of a subset of the neurons is the output of the network.

In practice, the output of a neuron is almost always computed by applying a nonlinear function to the sum of a parameterized linear combination of the inputs and a parameter called *bias*. Several functions are possible, with sigmoids being the historically popular option, and $f(x) := max(x, 0)$, known as ReLU in machine learning, being a faster modern alternative [30]. Inputs and activations are represented with floating point numbers, enabling efficient computation by dedicated hardware.

A neural network with undefined parameters can be seen as a higher order function, called a neural network architecture. Training is the process of determining the parameters, often the *weights* and *biases* of the linear combinations, that best fit some data. The most common training method is *supervised learning*, where a network architecture fits a function from sample input and output pairs. Parameters are set such that for each sample input the network produces an output close to that of the sample output.

Once the network is trained, inputs different from the samples used to train it can be fed into the network to obtain predictions, a process known as *inference*. Such predictions can then be compared with real samples to determine how well the network inferred, or *generalized*, the function that generated the samples. These samples constitute the *test set*, whose intersection with the *training set* of samples must be empty.

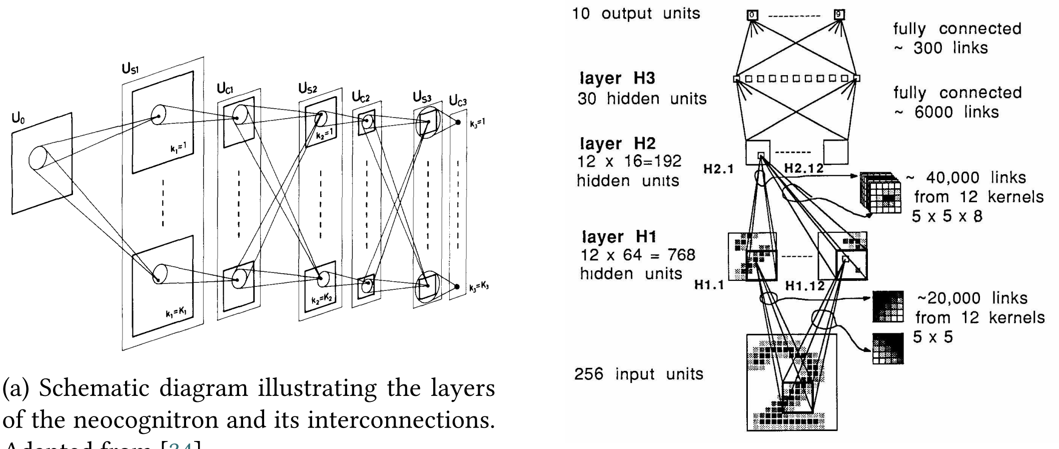
Training is often seen as a process of minimizing a *loss function*, a function that quantifies the difference between the network and sample outputs. By far the most common optimization method for neural networks is iterative *gradient descent*, i.e. to compute the gradient of the loss function with respect to the parameters and iteratively update the solution in the opposite direction. The gradient is multiplied by a number called the *learning rate*. It is a *hyperparameter*, because it parameterizes the training process that generates the parameters of the network. As is often the case with hyperparameters, there is no general theory to determine it *a priori*, therefore it is adjusted empirically.

In neural networks, computing the gradient is made feasible by an algorithm called *backpropagation* [31], which essentially efficiently computes the gradient of the loss function with respect to the parameters of the network by iteratively applying the chain rule for composite function derivation.

2.2.2 Convolutional neural networks

It has been long known in machine learning that an *inductive bias*, i.e. assumptions about the source of the training data not provided by the data itself, is necessary for generalization [32] [33]. Therefore, useful machine learning algorithms must contain *a priori* knowledge in order to make generalization possible. In the deep learning for neural networks paradigm, the architecture constitutes a significant part of the inductive bias. Thus novel architectures have been a fertile field of research in computer vision.

This is exemplified no better than by convolution, the backbone of the most popular neural network architectures in computer vision. Inspired by the pioneering neuroscientific work of Hubel and Wiesel on the visual cortex [36], Fukushima developed the precursor to modern convolutional networks [34]. Yann LeCun was the first to successfully train them through backpropagation, applying them to the domain of handwritten digit recognition [35] (Figure 2.14). Remarkably, the convolutional models that ignited the deep learning revolution are not too dissimilar architecturally from the first models [37]. On the contrary, the performance improvement is attributed to the general increase in computing power and the use of Graphics Processing Unit (GPU) to accelerate convolutional and fully connected layers.

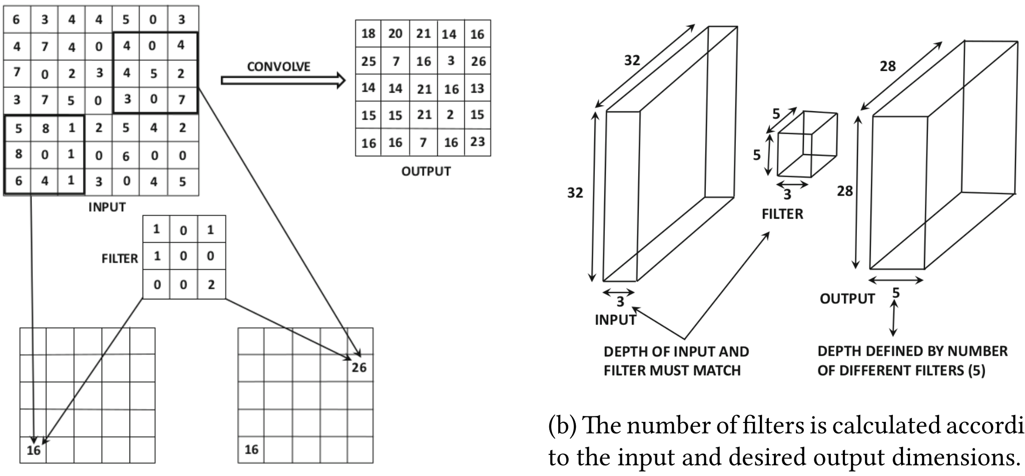


(a) Schematic diagram illustrating the layers of the neocognitron and its interconnections. Adapted from [34].

(b) First modern convolutional network. Adapted from [35].

Figure 2.14: Graphical comparison of the two pioneers of convolutional neural networks.

Indeed, the convolutional architecture of filtering and pooling is as effective as one could assume from its biological inspiration. 2D convolution is equivalent to sliding a rectangular filter across the image, multiplying the overlapping values and adding them to obtain a result (Figure 2.15). Because convolutional networks employ the same filter weights for all of the image, there is massive weight reuse compared to fully connected layers. This also provides the property of *shift invariance*, as filters respond equally to the same pattern in different parts of the image.


INPUT

6	3	4	4	5	0	3
4	7	4	0	4	0	4
7	0	2	3	4	5	2
3	7	5	0	3	0	7
5	8	1	2	5	4	2
8	0	1	0	6	0	0
6	4	1	3	0	4	5

FILTER

1	0	1
1	0	0
0	0	2

OUTPUT

18	20	21	14	16
25	7	16	3	26
14	14	21	16	13
15	15	21	2	15
16	16	7	16	23

CONVOLVE
16
26
16
32
32
3
INPUT
5
5
3
FILTER
28
28
5
OUTPUT
DEPTH OF INPUT AND FILTER MUST MATCH
DEPTH DEFINED BY NUMBER OF DIFFERENT FILTERS (5)

(a) Graphical depiction of convolution by a 3x3 filter without padding.

(b) The number of filters is calculated according to the input and desired output dimensions. In this example, 5 different filters would be necessary.

Figure 2.15: Schematic explanations of convolution [38].

The output of convolution is a slightly smaller image, according to the filter size, or the same size if padding is used. Furthermore, when applied to multichannel images such as RGB color, each input channel is convoluted with its own filter and the results are added, such that information is merged from all channels. In order to preserve or modify the number of output channels, the group of filters for each input channel is replicated with different weights for each of the output channels.

Traditional computer vision tasks, such as edge and corner detection or Gabor filtering, can be implemented as convolutional filters in a very direct manner. In trained convolutional networks, the initial layers actually learn this kinds of representations, that are then composed to form higher level features.

Furthermore, convolution can be strided, skipping those coordinates who are not 0 modulo the stride parameter. Strided convolution leads to smaller maps, which is usually combined with an increasing number of features to obtain more semantic features of lower spatial resolution. This process can continue until a vector of 1x1 maps is reached. This feature vector is often utilized as the input to a fully connected classifier. Pooling can be seen as a strided convolution with a nonlinear kernel, with the most common ones being the mean (average-pooling) or the maximum (max-pooling).

In summary, convolution is a very strong prior for images that takes advantage of the properties of locality and translation equivariance to greatly facilitate the training of deep neural networks.

2.2.3 Training deeper networks: the residual architecture

Perhaps the most important architectural innovation since convolutional networks themselves, Residual Networks or ResNets have allowed for the training of substantially deeper networks, with the corresponding gains in compositionality and performance [39]. ResNets simplify Highway Networks [40], which were among the first architectures to train effectively with depth exceeding 100 layers.

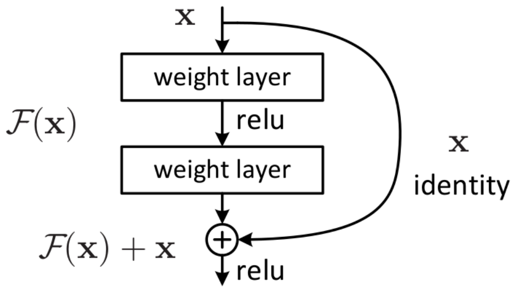


Figure 2.16: The residual learning block learns the difference between the desired and identity mappings [39].

At its core, the residual learning block (Figure 2.17) is a recognition that the training dynamics of a network are as important as its theoretical representation capacity. They are the product of analyzing an experiment where deeper networks underperform shallower ones. This must be due to difficulty of optimization, because there is a solution by construction of equal performance, namely setting the extra layers to the identity mapping. And indeed this recognition is the foundational prior of the residual block. By adding the input to the output, layers learn a residual function, i.e., the difference between the desired and the identity mapping.

As the original Residual Learning paper demonstrates [39], this recasting of the mapping greatly facilitates the optimization process and hence allows deeper networks to reach superior performance. The improvement in optimization is especially intuitive if we consider the effect of weight decay. Under 'plain' networks, weight decay incentivizes each layer to destroy all the information of the previous one, multiplying activations with weights close to zero. Under residual networks, weight decay incentivizes each layer to perfectly propagate the previous activations of the layer, and modifications to this layer by the weights are reasonably penalized as complexity.

The result is that Residual Networks are a theoretically sound architecture that demonstrates great performance in state-of-the-art benchmark tasks such as Imagenet [7], among many others. Residual Networks are very commonly used as baselines, a practice that has been vindicated by recent results applying modern training methodologies to the architecture [41].

2.2.4 Densely connected convolutional networks

Densely connected convolutional networks [42] take the idea of residual learning further by concatenating instead of summing feature maps. This can reduce the number of parameters needed, through feature reuse. Several residual blocks are hence substituted by a corresponding dense block (Figure 2.17).

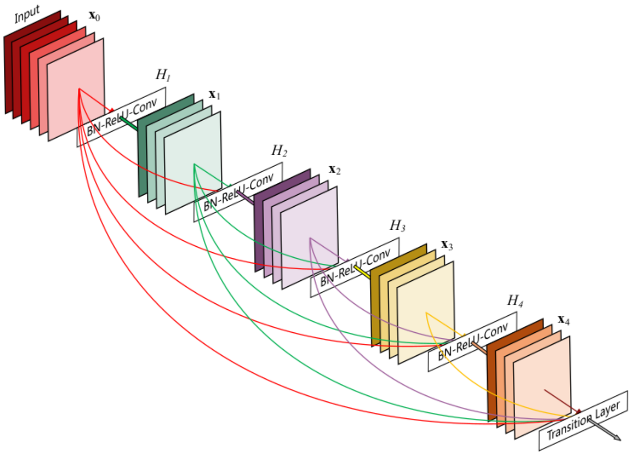


Figure 2.17: Each layer in the dense learning block receives the concatenation of all previous outputs as input [42]

While DenseNets do not seem to significantly outperform ResNets in the Imagenet benchmark [43], recent work has found them to be the only architecture invariant to translation among the ones tested [44]. This is very relevant to our domain because a limited data set might contain, for example, samples of fluid in one region that must be correctly recognized as such in other regions in the testing environment.

2.2.5 Transfer Learning: Imagenet pretraining

As was discussed in the introduction to convolutional networks, the first convolutional layers learn the classical filters of computer vision (Figure 2.18) and later layers compose these features to eventually perform a task such as classification, regression or segmentation. Because there is a great degree of statistical similarity between natural images, the weights of a classifier trained on Imagenet constitute a much better than random initialization for many different image tasks.

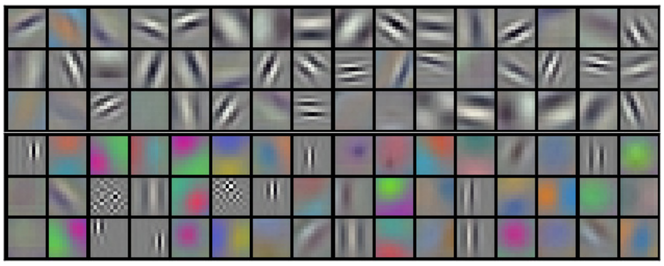


Figure 2.18: Features learned by the first layer of AlexNet [37].

The practice of using these weights as the initialization is known as *transfer learning*. It constitutes an effective way to accelerate training and especially to prevent overfitting in small data sets. A small classifier is often trained with the last convolutional layer as input, effectively only updating the latest layers and using the rest as a fixed feature extractor. Only when the classifier is trained are the rest of the layers updated, and then only slightly. This second process is known as *fine tuning*.

However, it is not so clear whether the benefits of transfer learning transfer to artificial images such as OCT scans. In principle the first layers should, because the notions of edge and texture are common. However, to what extent higher level features contribute to transfer is probably a matter of empirical investigation. This includes the research of closer, *domain-specific* transfer learning, where data sets are increasingly specialized.

2.2.6 Class Activation Mapping (CAM)

CAM [45] is an attribution technique for CNNs that leverages the spatial nature of convolution to determine which parts of the input image contributed to the output of the network. It can be used to obtain class-specific image regions from classifiers trained in whole-image labels; essentially constituting a very weakly supervised localization or segmentation model. However, it is most often used in *post hoc* analysis of classifiers for the sake of interpretability, including the medical domain. Figure 2.19 presents CAM schematically and illustrates its application to another ocular imaging domain where deep learning is utilized.

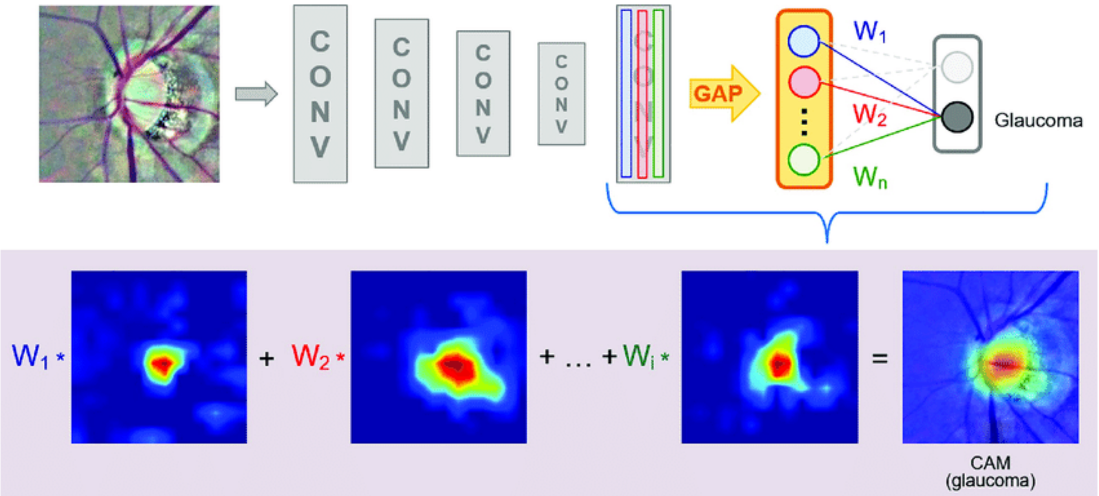


Figure 2.19: CAM applied to glaucomatous optic neuropathy [46].

CAM is only applicable to a particular kind of CNN architectures performing global average pooling [47] over convolutional maps, immediately before a last layer (e.g. softmax); Gradient-Weighted Class Activation Mapping (Grad-CAM) is a similar later technique that removes this limitation [48]. The latter has been used to generate the CAMs for this project, through the library listed in the next chapter.

Chapter 3
Planning

When it comes to developing a research project, planning is a critical aspect that can have a great influence on its outcome. Thus, this chapter describes the planning followed throughout the course of this project, from the development model chosen to resources and costs.

3.1 Development model

In order to best manage the characteristics of a research and development project, an iterative and incremental development model has been followed (Figure 3.1). Fundamentally, this model decomposes the project into discrete iterations, within which progress proceeds incrementally. In the case of this project, iterations map to the experiments that are performed. The greatest advantage of this approach is that every iteration produces working models and an analysis of metrics. This decoupling of progress allows for the review of partial results, whose adequacy is evaluated in a way that informs the development of subsequent ones.

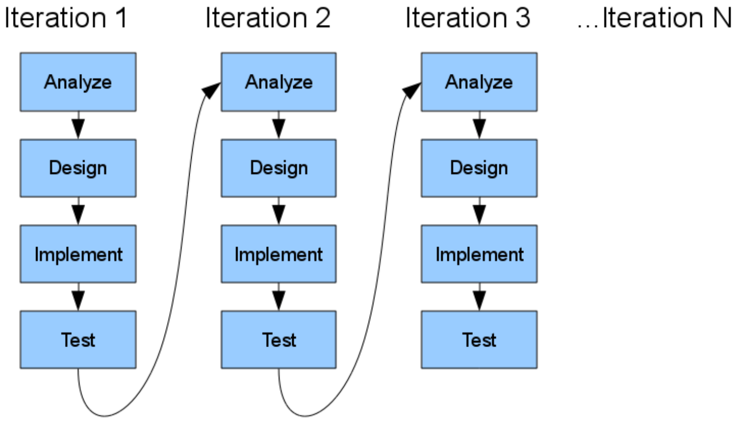


Figure 3.1: Iterative model of the software process [49].

3.2 Iterations and scheduling

Following the iterative model described above, the project has been scheduled through its decomposition into discrete iterations, detailed in the following paragraphs. These iterations were then used to elaborate the Gantt chart of Figure 3.2, that constitutes the baseline for the project.

Study of the domain and previous approaches The first iteration aims to establish a theoretical foundation of background knowledge that informs the development of the experiments. This domain-specific knowledge ranges from the anatomy of the eye to the specifics of AMD characterization. The main materials used will be scholarly sources such as research papers.

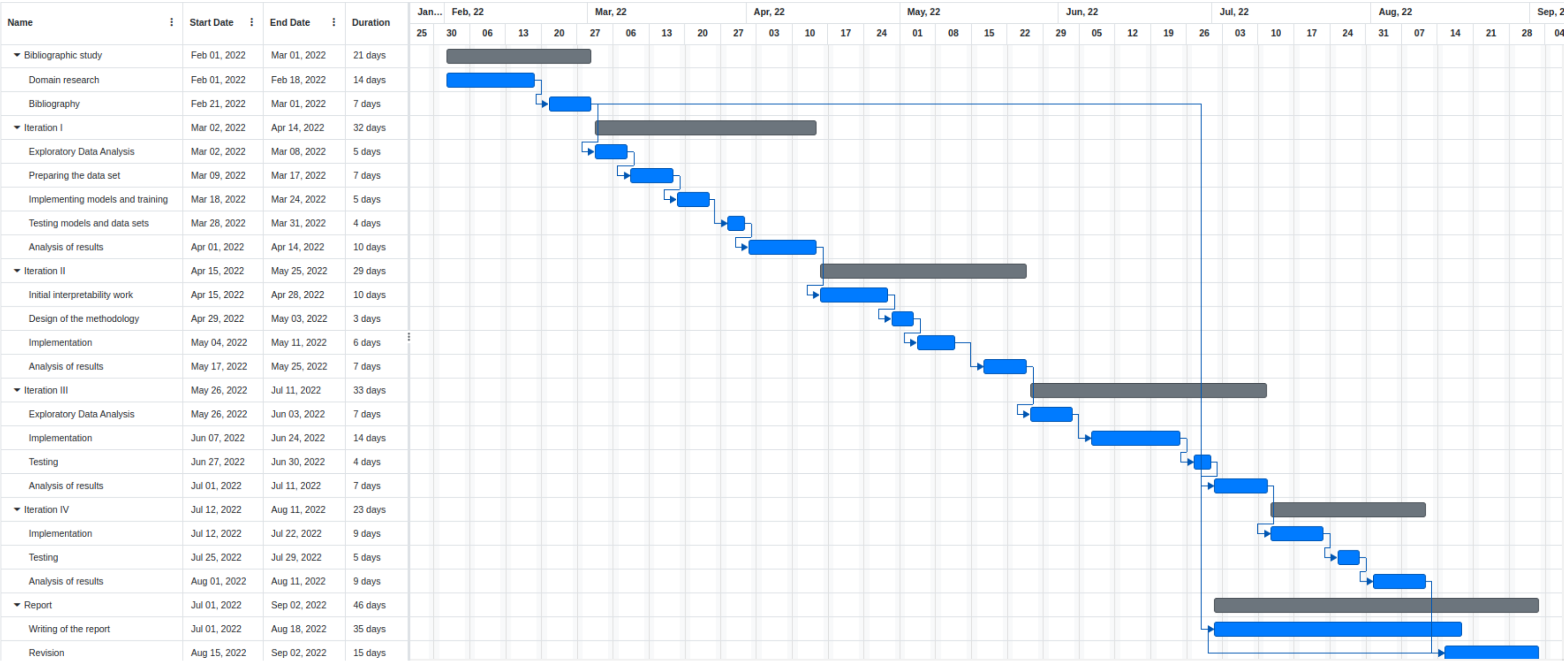
nAMD detection The next iteration includes the development of an initial experiment that tests the viability of detecting nAMD signs in OCT images. This first task will serve to establish baselines of performance, as well as to compare how different usual deep learning techniques compare. Furthermore, a comprehensive analysis of metrics will be performed to deeply understand the performance characteristics of the best approaches. In doing so, this iteration provides the grounds for the next technique and experiment.

CAM and review methodology This iteration analyses the results of the previous one through interpretability techniques. Concretely, it uses the CAM technique to analyze the factors that influence the outputs of models. This process involves the participation of domain experts, whose time to interpret results has to be accounted for in the scheduling.

Transfer learning This iteration considers the possibility of leveraging domain-specific transfer learning to develop better classification models in novel tasks. It relies on the results from the first experiment to evaluate the effects of transfer. Therefore it includes the development of compatible data loaders for both data sets and their use to fine tune models.

nAMD classification This iteration reproduces previous results with a finer classification. It analyses them in the more complex task of multi-label characterization of nAMD signs.

Writing of the report This last iteration includes everything related to the writing of the report. This is done incrementally from the saved results, discussing them from the perspective provided by the rest, contextualizing and highlighting their significance, and describing their relation to existing and future lines of work.



Name	Start Date	End Date	Duration
▾ Bibliographic study	Feb 01, 2022	Mar 01, 2022	21 days
Domain research	Feb 01, 2022	Feb 18, 2022	14 days
Bibliography	Feb 21, 2022	Mar 01, 2022	7 days
▾ Iteration I	Mar 02, 2022	Apr 14, 2022	32 days
Exploratory Data Analysis	Mar 02, 2022	Mar 08, 2022	5 days
Preparing the data set	Mar 09, 2022	Mar 17, 2022	7 days
Implementing models and training	Mar 18, 2022	Mar 24, 2022	5 days
Testing models and data sets	Mar 28, 2022	Mar 31, 2022	4 days
Analysis of results	Apr 01, 2022	Apr 14, 2022	10 days
▾ Iteration II	Apr 15, 2022	May 25, 2022	29 days
Initial interpretability work	Apr 15, 2022	Apr 28, 2022	10 days
Design of the methodology	Apr 29, 2022	May 03, 2022	3 days
Implementation	May 04, 2022	May 11, 2022	6 days
Analysis of results	May 17, 2022	May 25, 2022	7 days
▾ Iteration III	May 26, 2022	Jul 11, 2022	33 days
Exploratory Data Analysis	May 26, 2022	Jun 03, 2022	7 days
Implementation	Jun 07, 2022	Jun 24, 2022	14 days
Testing	Jun 27, 2022	Jun 30, 2022	4 days
Analysis of results	Jul 01, 2022	Jul 11, 2022	7 days
▾ Iteration IV	Jul 12, 2022	Aug 11, 2022	23 days
Implementation	Jul 12, 2022	Jul 22, 2022	9 days
Testing	Jul 25, 2022	Jul 29, 2022	5 days
Analysis of results	Aug 01, 2022	Aug 11, 2022	9 days
▾ Report	Jul 01, 2022	Sep 02, 2022	46 days
Writing of the report	Jul 01, 2022	Aug 18, 2022	35 days
Revision	Aug 15, 2022	Sep 02, 2022	15 days

Figure 3.2: Gantt chart of the baseline for the project.

3.3 Resources

This section describes the resources needed for the development of the project. They have been divided into human and material resources, the latter including the necessary software.

3.3.1 Material resources

The material resources necessary for this project can be divided into hardware and software. Both are listed and described in the following sections:

Hardware resources

- **Laptop**: A laptop computer has been used for bibliographical research, development and testing, and the writing of the report. Thanks to the use of the server and remote software listed below, there was no need for specialized hardware to train neural networks in the client. Rather, this task has been offloaded to the more appropriately specified workstation.
- **VARPA workstation**.
  - CPU: Intel(R) Core(TM) i9-9900K; 8 cores, 16 threads
  - GPU: Nvidia RTX 2080 Ti; 11 GiB VRAM
  - RAM: 32 GiB

Software resources

The following open source programs and libraries were used throughout the project:

- **PyTorch and Torchvision** (versions 1.12.0 and 0.13.0): PyTorch is the deep learning library which, together with its Torchvision component, has been used to implement the data sets, models and training scripts used throughout the project.
- **Pandas** (versions 1.4.3 and 0.11.2): This data science library was used for exploratory data analysis, the processing of labels and the creation of summaries from metrics.
- **Matplotlib and Seaborn** (versions 3.5.2 and 0.11.2): These data visualization libraries were used to generate the figures that appear on this report.
- **scikit-learn** (version 1.1.1): Used for calculating metrics such as ROC and its AUC.
- **grad-cam** (version 1.2.8): This Python library implements the Grad-CAM interpretability technique, for use with PyTorch models.

- **scipy** (version 1.8.1): The ability of this library to process MATLAB files was used to extract B-scans out of the public data set.
- **Visual Studio Code**: This extensible editor was used together with its Jupyter Notebook and SSH extensions to access the workstation remotely.

3.3.2 Human resources

The successful development of this project requires the participation of several distinct roles, of which four can be highlighted as the main ones: **project manager**, **analyst**, **designer** and **developer**. That role, as well as research advisor, has been performed by the project advisors. Their work is reflected in the periodic meetings where the progress and results of the project were discussed.

3.4 Cost estimation

The costs of this project can be broken down into two major categories, according to the associated resource: *material costs* and *remuneration of labor*:

Material costs: This first cost category is divided into the resource subcategories of hardware and software, with neither incurring any cost. The cost of the former is negligible because it utilizes resources, namely the laptop and workstation, that were already necessary for the student and the research group, and whose usage did not conflict.

Remuneration: This second source of costs has been calculated based on scheduling. From 30 20-hour workweeks for the roles of analyst, designer and developer, two hour weekly meetings by the two project managers and an estimation of their hourly rate, we obtain Table 3.1.

Role	Hourly rate	Person-hours	Investment
Project manager	30 €/h	120 h	3600 €
Analist, designer, developer	20 €/h	600 h	12000 €
		Total:	15600 €

Table 3.1: Estimation of costs dedicated to human resources.

Thus, given the negligible material costs, we determine the **estimated total cost of the project** to be 15600€.

Chapter 4

Methodology

This chapter describes the experiments performed during this project, as well as the design and implementation of methodologies and techniques aimed at addressing the problems that arise in the tasks considered and related ones.

4.1 Introduction to the data set

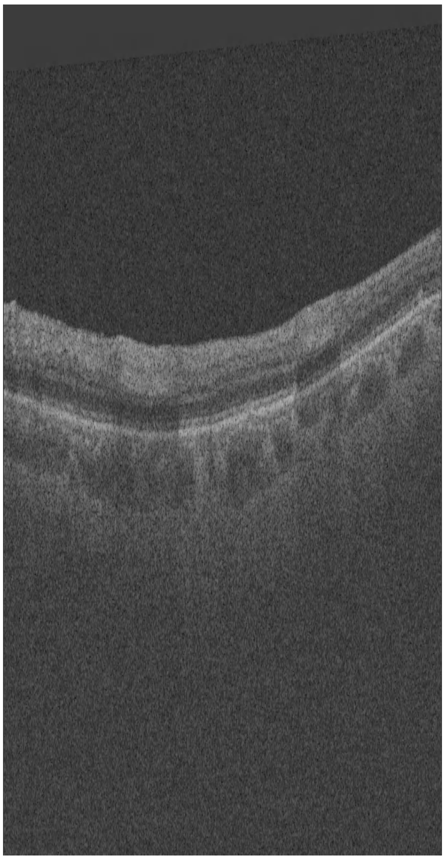


Figure 4.1: First image of the data set, presenting neither fluid nor neovascularization.

The experiments that follow have been performed on a data set of 1279 labeled scans obtained from an SS-DRI-Triton-OCTA device (Topcon Corp Inc, Tokyo, Japan). The labeling of the data set indicates the presence or absence of neovascularization and fluid. Of the 779 images that present either, 566 present the former and 575 the latter. Furthermore, both signs of AMD are decomposed into three subtypes each, according to their position in the retina.

The unprocessed images have a 512 by 992 resolution (Figure 4.2), which under usual settings would correspond to an area of 7 by 2.57 mm [50]. This would mean that the physical vertical separation of pixels is, at 13.67 µm, larger than the horizontal 2.59 µm, resulting in stretched images. However, the retina and the first layers of the choroid occupy only a fraction of the image vertically. In order to greatly reduce the size of the images while retaining all of the relevant information, a simple and robust algorithm was developed (Figure 4.2). A window is slid across the image, and its average intensity calculated for all positions. The sliding begins from a vertical offset of 150 pixels, which together with the window size of $262 = 512 - 150 - 100$, maximizes the average intensity of the retina-sized central region while ignoring the margins. This proved more robust than edge-based algorithms, especially for cases of posterior vitreous detachment or unusual morphology.

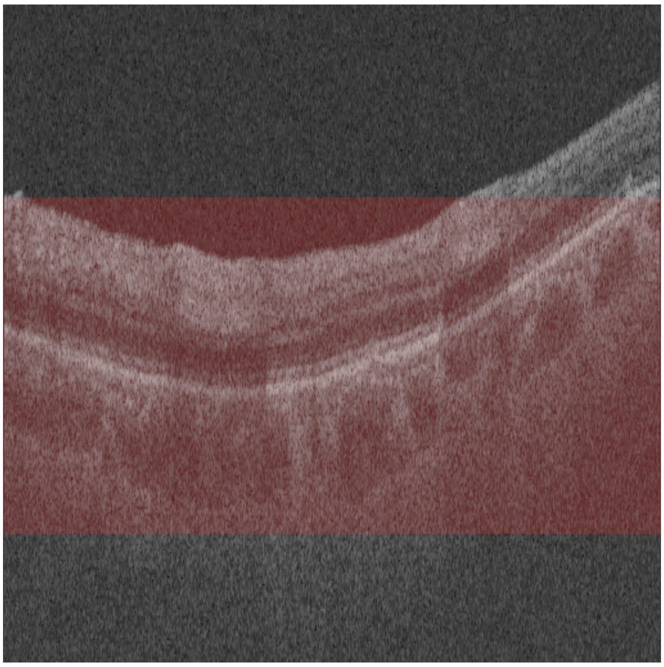
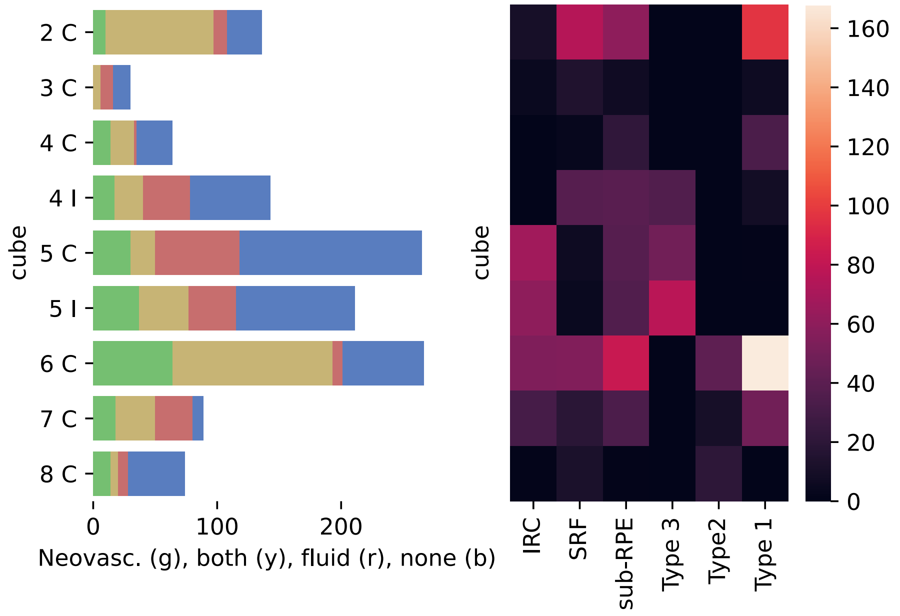


Figure 4.2: Figure 4.1 after being preprocessed. The red area indicates the window whose average brightness is maximized by the algorithm.

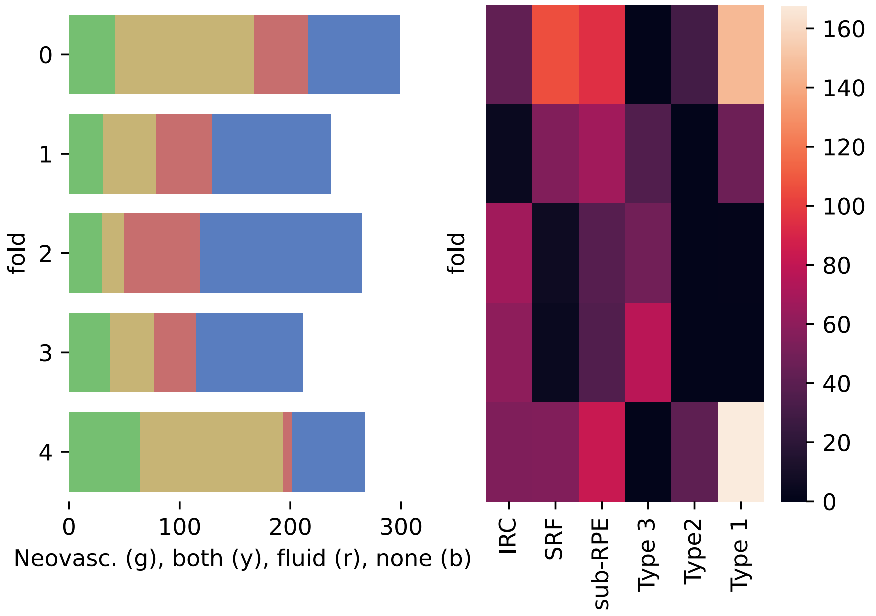
The scans are divided into 9 groups, corresponding to different eyes. For the purpose of correct performance assessment, it is important that images from the same eye are only utilized in one of training, validation or testing. This guarantees that intra-eye regularities are not contributing to the evaluation, being more representative of unseen images.

The decision to divide the groups according to the eye and not the more strict criterion of patient is based on the following: "Age-related macular degeneration can be asymmetrical; one eye may show manifestations, such as drusen, in the absence of fellow-eye abnormalities. The risk for progression in the eye without AMD stigmata is nearly zero, and accordingly, one should not diagnose AMD in an eye without visible abnormalities" [25]. In any case, only fold 1 contains the small matching cubes 4 C and 4 I, so any effect would be small.

Indeed, B-scans are neither equally spaced nor equinumerous across scans. This leads to a large imbalance, not only in the amount of samples, but in their class distribution. Nevertheless, by grouping scans according to their number of samples and their labels, a split into 5 groups can be achieved that greatly reduces the imbalance (Figure 4.3).



(a) Original data set distribution.



(b) Data set distribution after grouping.

Figure 4.3: Class distribution of C-scans before and after grouping.

4.2 First experiment: detection of nAMD signs

This first experiment will provide an initial evaluation of the data set, the selected architectures and their training methodology.

4.2.1 Data set considerations and their impact on the task

The *exploratory data analysis* described in the last section hints to the expected difficulty of the data set. The limited amount of samples, and especially of independent scans, motivates an initial approach that tests the viability of AMD characterization. Therefore, the first iteration will consist in the task of binary classification.

However, this approach is not equivalent to binary classification of nAMD. The particularities of the data set mean that there is no true control patient, as every C-scan contains some B-scans with the presence of either neovascularization or fluid. Indeed, the increased granularity of labeling implies that some samples that are considered negative, as in the lack of neovascularization and fluid, contain nevertheless other signs of AMD, namely drusen. This distinction will be especially apparent in the next experiments, where model performance will be assessed through interpretability techniques and the most problematic samples will be highlighted, e.g. Figure 4.4. Clearly pathological features, such as large drusen, are labeled as negative as long as fluid and neovascularization are not present.

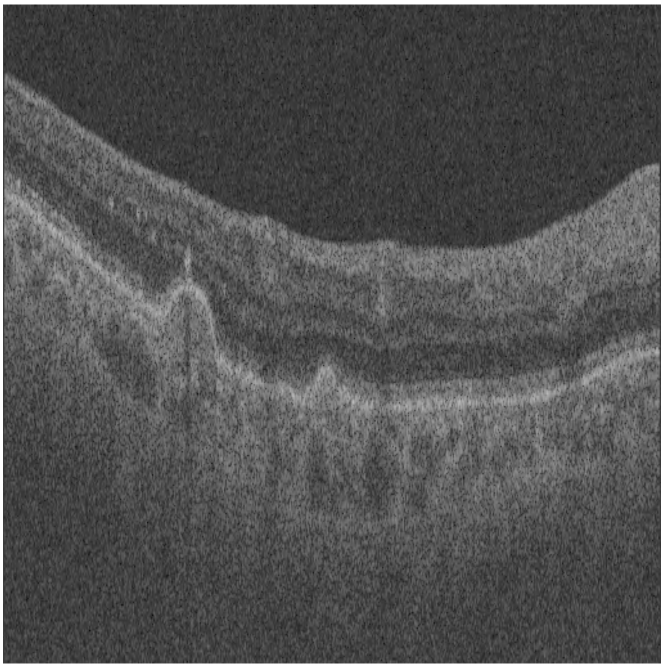


Figure 4.4: Sample from the data set demonstrating drusen.

As is best practice in machine learning, a test set will be reserved, not to be used in model selection nor fitting. In order to provide the most representative sample possible within the constraints of the data set, *group 0* (figure 4.3b), has been chosen for the purpose. Being comprised of three independent C-scans, it is the most numerous and well balanced in terms of class distribution. The remaining groups will be employed for *4-fold cross validation*.

In order to further accelerate development, initial experiments were performed with a down-sampled image, from the original 512 to 256 pixels squared. Subjectively, the image retains enough resolution not to impede classification, and the fourfold reduction in pixels entails a corresponding reduction in memory consumption and run time, through the diminution in convolutional activations. Furthermore, later experiments will demonstrate that the decreased resolution might be not an impediment, but a small boost to performance through a hypothesized regularizing effect, especially given the data-constrained regime.

4.2.2 Baseline architecture

The chosen neural network architecture is the ResNet-18. As was mentioned in the contextualization chapter, the Residual Network [39] is a comparatively conceptually simple architecture that proved to be a robust baseline in similar domains, especially in light of advances in training techniques [41]. The ResNet-18 is the smallest of the networks presented in the original paper [39], with comparable performance to “plain” networks. However, this relatively small size already shows faster convergence empirically, and is theoretically easier to train thanks to improved gradient flow during backpropagation. Moreover, the weights resulting from training it on Imagenet are widely available.

Therefore the first experiments were performed with such network. A fixed batch size of 4 was selected in order to enable direct comparison with other, more memory intensive architectures. In any case, increases in batch size were found to be detrimental to generalization performance, possibly through a decrease in the regularizing effect of stochastic gradient descent.

Metrics and baseline training procedure

The values in Table 4.1 represent the average cross-validated performance of ResNet-18 on the unaugmented data set. They provide an approximate baseline of performance against which further developments will be measured. The rationale for the chosen metrics is that the mild class imbalance, due to the initial binary task, obviates the need for base-rate-adjusted metrics such as precision. Binary cross entropy is used as the loss function, as it provides the most natural loss under the paradigm of maximum likelihood estimation. As a metric, binary cross entropy provides an evaluation of the confidence of the network, with lower values indicating more confident correct guesses and less confident mistakes. The Area Under the Receiver Operating Characteristic Curve (AUROC) indicates the probability that positive cases are ranked higher than negative ones, intuitively providing a measure of the quality of the ranking provided. This has clinical applicability through the prioritized screening of patients. Finally, accuracy provides a highly intuitive measure of classification performance.

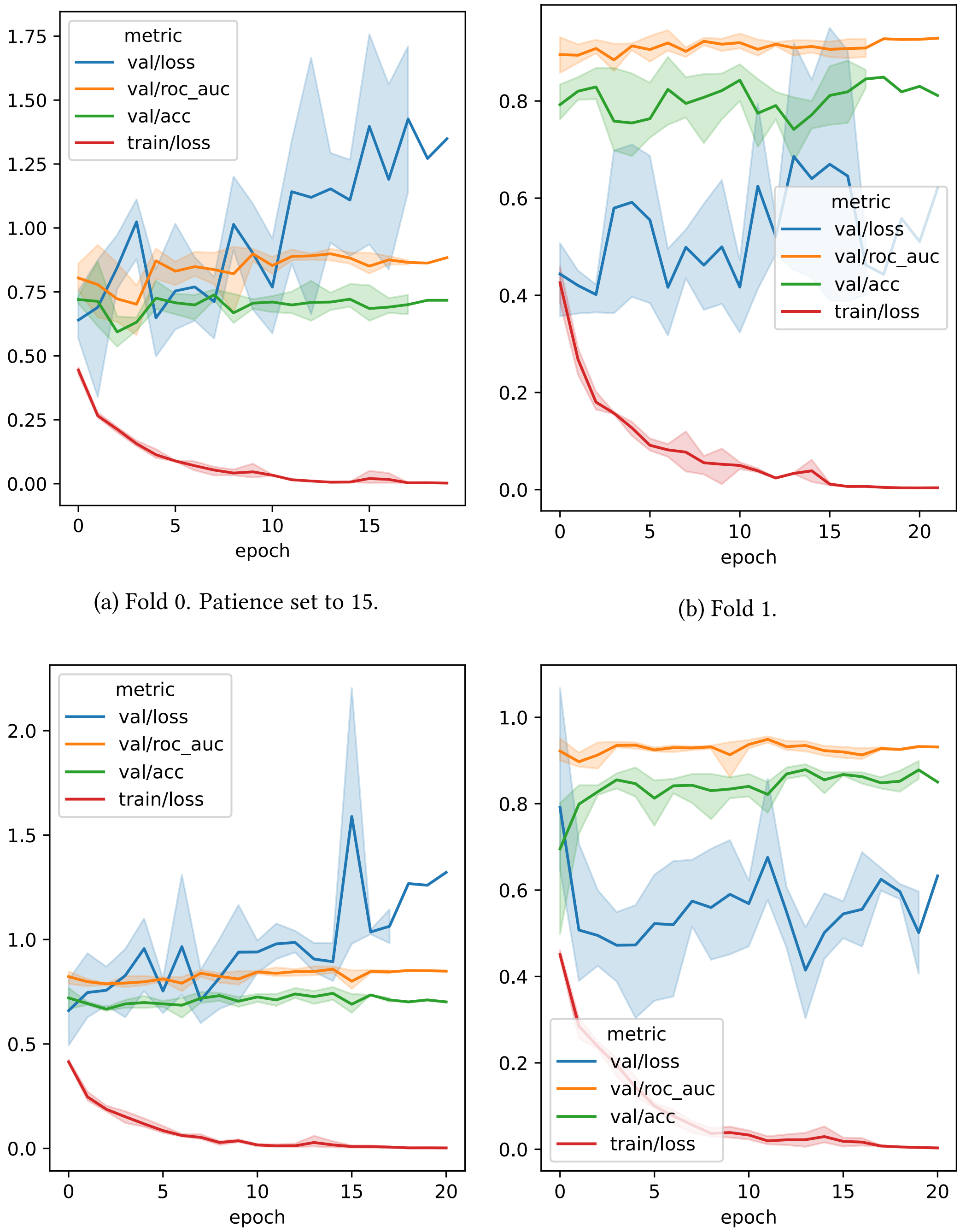
Pretrained	**LR**	Cross-entropy	Accuracy	AUROC
No	10^{-3}	0.4220	0.8029	0.8863
No	10^{-4}	0.5960	0.7252	0.8078
No	10^{-5}	0.6593	0.5765	0.5982
Yes	10^{-3}	0.4810	0.7851	0.8568
Yes	10^{-4}	0.4531	0.7943	0.9023
Yes	10^{-5}	0.4483	0.8090	0.8875
Yes	10^{-6}	0.4554	0.7990	0.8781

Table 4.1: Mean cross-validated performance of ResNet-18 on normalized but otherwise unprocessed data.

While useful for hyperparameter selection, these metrics might not be sufficient for model assessment. Therefore, full Receiver Operating Characteristic (ROC) analysis will be performed on key models. Unlike summary metrics like the F1 score, ROC analysis provides the complete range of trade-offs between *sensitivity* and *specificity*, extremely common metrics in the medical domain. It is for this reason that ROC analysis is widely considered the gold standard of diagnostic test evaluation. Moreover, as was evidenced by the lack of true control patients, there is reason to suspect that the prevalence of neovascularization and fluid in the data set is not representative of the patient distribution. Thus, Precision-Recall (PR) analysis, where precision implies conditioning ‘accuracy’ on prevalence, is deemed less adequate.

Using Imagenet as a fixed feature extractor yielded worse performance, albeit with no discernible overfitting and highly stable curves. Furthermore, fine tuning after training the classifier yielded no benefit over using Imagenet as initialization, thus the latter was used throughout the project.

As for the optimization process, the Adam stochastic optimizer [51] was used with a fixed learning rate schedule and *early stopping* with a patience of 10 epochs. This means that after 10 passes through the complete training data set without an improvement in validation loss training is stopped and the model with the lowest validation loss is returned.



(a) Fold 0. Patience set to 15.

(b) Fold 1.

(c) Fold 2.

(d) Fold 3.

Figure 4.5: Mean and 95% confidence interval for the training curves of the baseline model (ResNet-18 with a learning rate of 10^{-4})

4.2.3 Data augmentation

Closer inspection of the baseline results (Figure 4.5) reveals that the limiting factor of performance is not the capacity of the model, but rather its generalization. The limited amount of samples implies a large probability of the model finding spurious correlations in the training data, leading to overfitting. This may not always be solvable through generic regularization such as weight decay. Table 4.2 shows the results of applying the AdamW optimiser [52].

WD	**LR**	Cross entropy	Accuracy	AUROC
0	10^{-4}	0.4531	0.7943	0.9023
10^{-4}	10^{-4}	0.4665	0.7959	0.9009
10^{-3}	10^{-4}	0.4731	0.7952	0.8874
10^{-2}	10^{-4}	0.4613	0.8001	0.8684

Table 4.2: Mean performance after weight decay is applied to the baseline models.

Furthermore, the evolution in cross entropy and the ranking metric AUROC seem to decouple after the lowest validation loss in folds 0 and 2 (Figure 4.21a and Figure 4.5b). This leads to a hypothesis asserting the existence of outliers in those validation sets. The reasoning is that for a model to obtain a worse than chance cross entropy while simultaneously obtaining much better than chance AUROC and accuracy, it must be making very confident wrong predictions. Those could also be the result of the model always predicting the same class confidently, which would also result in high cross entropy; however the good accuracy at threshold 0.5 disproves it.

Moreover, in the initial epochs where training loss is still significantly high, the presence of similar correctly labeled samples provides a reasonable explanation as to why validation loss starts lower. Indeed, training for less epochs being a regularizer is the basis of early stopping. The outlier hypothesis fits the observation that if the model is to fit the training set almost perfectly, as it does, the presence of outliers would result in an increase in cross entropy without appreciable decreases in accuracy or AUROC, as is also the case.

Perhaps the most powerful form of regularization in computer vision, *data augmentation* generates artificial samples that are different from the training data yet retain the same labeling. Thus, *data augmentation* is domain-specific, as certain transformations of the image modify the labels of one task while leaving others unaffected. As an example, applying a Gaussian blur to the image of a dog does not turn it into a cat, but blurring the texture of neovascularization in an OCT scan may make the resulting sample negative. Therefore, data augmentation must be designed taking into account the particularities of the domain.

To that end, the transformations employed in the following experiments have been limited to “realistic” ones, i.e., those that not only maintain the label but that could plausibly belong to the data distribution. The first one explored, horizontal mirroring of the scans, takes advantage of bilateral symmetry of the eyes to effectively duplicate the number of samples. Similarly, applying a small, randomly placed crop to the image provides plausible samples, given that the scans are not aligned in the data set. Cropping from 256 to 248 pixels incurs only a small risk of excluding important features from samples.

Augmentations	Cross entropy	Accuracy	AUROC
None	0.4531	0.7943	0.9023
Flipping	0.3390	0.8600	0.9262
Flipping and cropping	0.3419	0.8589	0.9307

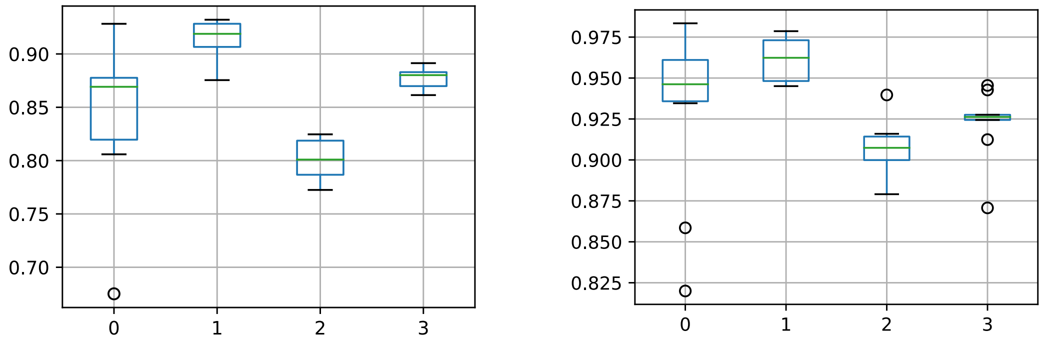
Table 4.3: Data augmentation results for the ResNet-18 model. Metrics are the mean of 3 times repeated 4-fold cross-validation.

Table 4.3 indicates that both of the plausible augmentations provide a boost to performance. However, the influence of cropping is mixed, as while it improves AUROC, it comes at the expense of the other metrics.

Variability across and within groups

As we refine the training methodology, the risk of overfitting hyperparameters, in this case the data augmentation, increases. Especially given that, in addition to the cross validation average being an imperfect estimator of performance on the real distribution, performance within folds is also subject to variation. While the weight initialization is kept consistent across folds due to the Imagenet pretraining, the shuffling of training samples is controlled through a *seed* parameter. Thus experiments may be repeated with different seeds to study the consistency of convergence.

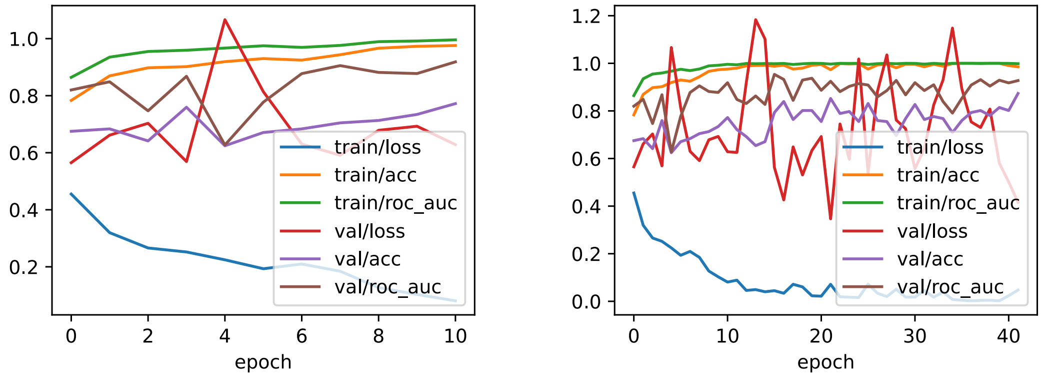
Figure 4.6 demonstrates the distribution of results of different shuffles for each fold. Inspecting the training and validation curves (Figure 4.7a) reveals that in the case of the accuracy outlier, the lowest validation loss was obtained in the first epoch. Judging by the similar loss during the next epochs and the upwards trend in accuracy and AUROC, it is likely that validation performance was limited by the early stopping. Indeed, Figure 4.7b shows that doubling the patience to 20 epochs allows the exact same model to continue to train to a validation performance consistent with the rest of the training shuffles.



(a) Variation of accuracy with shuffling.

(b) Variation of AUROC with shuffling.

Figure 4.6: The augmented baseline was trained with 10 different random shuffles of the training data. There is significant variation across folds and model instability within fold 0.



(a) Original training and validation curves of the outlier.

(b) Training repeated with 20 epochs of patience.

Figure 4.7: The outlier from Figure 4.6a obtained its lowest loss on the first epoch, compared to the next ten before it was early stopped.

Arguably, this could seem to indicate that the patience parameter should be increased to obtain a more accurate evaluation of hyperparameters. However, a longer patience not only gives up shorter training times for better assessment. On the contrary, an excessively long patience might provide misleading results, given the instability of the validation curves.

Validation performance is random with respect to true performance on the underlying distribution, and with respect to testing performance. Because the selection criterion of the early stopping is the lowest loss, it constitutes a biased estimate of true performance. When the model sufficiently fits the training data, as is the case in 4.7b, where it reaches 100% accuracy and AUROC; stochastic gradient descent provides random samplings of weights with similarly low loss. The longer this sampling process continues, the more biased the validation estimate becomes. Therefore, a shorter patience period could be beneficial in this unstable regime, implicitly promoting hyperparameters that demonstrate consistently increasing performance with increasing epochs, as opposed to those with high variance in outputs and therefore validation performance.

In summary, the modest size of the data set and its validation splits implies that validation performance is especially noisy. Therefore, the early stopping patience parameter must balance too eagerly cutting training short of convergence and biasing its estimate of performance. Given that too eager early stopping was the cause of misleading performance in only one out of forty training iterations, the choice of ten as the patience parameter is maintained. Nevertheless, individual validation curves will be monitored for this phenomenon.

4.2.4 Increasing model capacity

The previous experiments place the task in the data-constrained overfitting regime. Unlike the classical statistics view of overfitting, modern ML finds that heavily overparametrized models exhibit better generalization than those close to the complexity of the task, especially in the presence of label noise [53]. We explore whether an improvement in performance is possible through this phenomenon. To do so, we train the ResNet-50, which, as it name indicates, has more than twice as many layers as ResNet-18, and a similarly increased number of trainable parameters.

Model	**Augs.**	Cross entropy	Accuracy	AUROC
ResNet-18	*F*	0.3390 ± 0.0076	0.8600 ± 0.0014	0.9262 ± 0.0026
ResNet-18	*F & C*	0.3419 ± 0.0317	0.8589 ± 0.0199	0.9307 ± 0.0137
ResNet-50	*F*	0.3394 ± 0.0188	0.8638 ± 0.0091	0.9192 ± 0.0100
ResNet-50	*F & C*	0.3760 ± 0.0164	0.8414 ± 0.0056	0.9170 ± 0.0079
DenseNet-121	*F*	0.3032 ± 0.0287	0.8724 ± 0.0182	0.9439 ± 0.0008
DenseNet-121	*F & C*	0.2947 ± 0.0020	0.8759 ± 0.0072	0.9446 ± 0.0025
DenseNet-161	*F*	0.3085 ± 0.0399	0.8774 ± 0.0192	0.9411 ± 0.0155
DenseNet-161	*F & C*	0.3029 ± 0.0067	0.8732 ± 0.0157	0.9355 ± 0.0003

Table 4.4: Comparison of results for different architectures. *F* stands for horizontal flipping and *C* for randomly cropping to 248 pixels.

We selected DenseNets as the architectural alternative because their increased feature reuse and gradient propagation might make them easier to train, and their translation invariance [44] might help with detecting lesions in new regions of the scans. This turns to be the case as the DenseNet-121 shows results superior to those of ResNet-50 (Table 4.4), while taking approximately the same time to train and producing a smaller number of parameters.

Figure 4.8 demonstrates the ROC curve for all of the folds in the first repetition of training the DenseNet-121 with horizontal flipping and cropping, i.e., the best performing model. We can observe the great variability across folds once again, which goes beyond affecting the AUROC to modify the shape of the ROC curve.

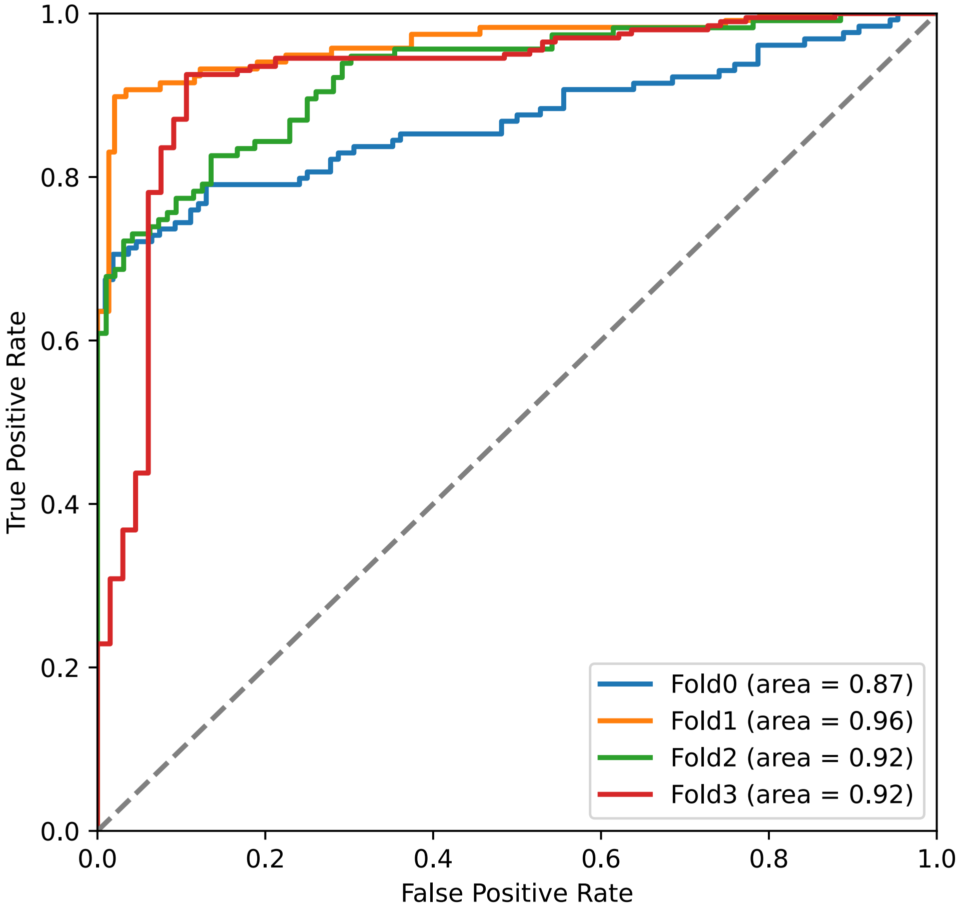


Figure 4.8: Validation ROC curve for the first repetition of the best model tested so far: DenseNet-121 with horizontal flipping and random cropping augmentations.

Furthermore, the shape of the curves can be used to establish hypotheses about the data of each fold. For example, the sudden drop-off in sensitivity (True Positive Rate) of fold 3 when increasing specificity (one minus the False Positive Rate) could be caused by samples mislabeled as negative, such that the threshold needs to be raised too much for them to result in negative classification.

One thing that might support the theory that validation variability is due to label noise in the validation sets is testing. If the patterns described previously do not hold in the testing set, then there is reason to suspect that validation is the culprit. And indeed Figure 4.9 confirms that the pattern of plummeting sensitivity in fold 3 validation does not translate to its testing results. Moreover, perhaps due to the special care taken to keep the testing set large and diverse, the ROC curves are much closer to the archetypal curve. Fold 0, however, is a testing outlier, maintaining its poor performance in the high sensitivity regime, to the point of dropping below chance.

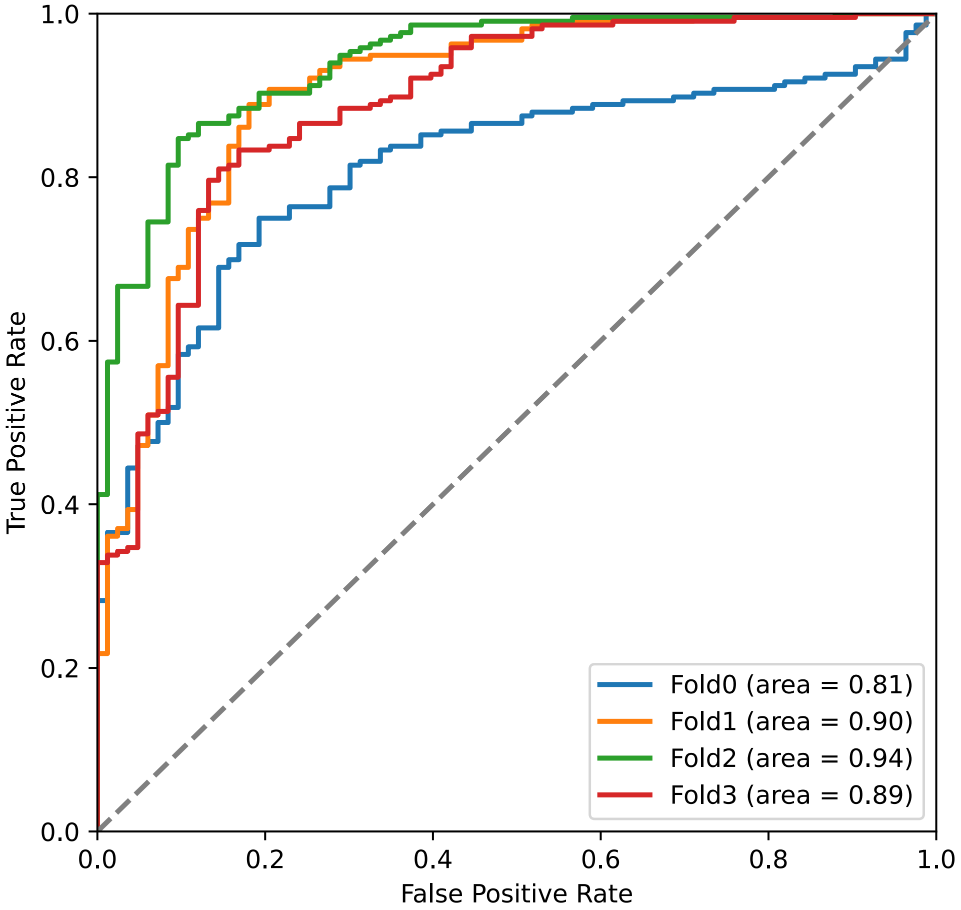


Figure 4.9: Testing ROC curve for the first repetition of the best model tested so far: DenseNet-121 with horizontal flipping and random cropping augmentations.

Overall, these results seem to support the theory that data considerations, such as scarcity and label noise, are the primary drivers behind the outcomes. Therefore, the next experiments will directly address these concerns, studying the data in more detail and attempting to provide a solution to its perceived limitations.

4.3 Second experiment: model-assisted intra-expert variability mitigation

While the difficulty of acquiring sufficient amounts of data is widely considered to be one of the greatest challenges of medical ML, expert variability constitutes an equally important problem. Due to the rapidly evolving nature of medical research and practice, together with the sheer difficulty of diagnosis, a certain amount of variability is to be expected. Therefore the development of techniques for the detection and correction of such variability are paramount to the advancement of ML in the medical domain.

4.3.1 Proposed methodology

We propose a methodology to perform intra-expert variability detection and mitigation with the assistance of deep learning models and interpretability techniques. As discussed in the introduction, OCT image classification is a labor-intensive task where human factors such as fatigue imply that even the most accurate experts demonstrate a certain level of inconsistency in their labels. The methodology herein proposed aims to address this problem by helping domain experts recognize samples that are inconsistent with the rest, according to some measure.

Given a graphical reasoning for the change, the expert either accepts or rejects the suggestion to reconsider the supposedly inconsistent label. However, were the expert to accept the change, there would now be two different labels for the same sample, at different times. It would not be prudent to give preference to the second one just because the model seems to agree with it. Therefore we do not. Instead, one or more independent experts are asked to review those cases where the original labeler admits that the label opposite to the original is reasonable. In this fashion, we leverage the power of expert committees to obtain more dependable labels. We do so *efficiently*, by asking the opinion of more than one expert only when we detect intra-expert variability.

It is important to highlight and restate that the mitigation of variability, i.e. the change of the labels, is **never** done on the basis of a neural network prediction. The outputs of the network and their associated interpretability data are only used to guide the expert to samples that are likely to receive different opinions different times; human experts always make the final decision of true labels. In the case of our project, the final verdict on labels is decided by a committee that includes both the medical imaging researchers with years of experience in the OCT domain and nAMD in particular that annotated the data set; and an independent clinician.

Several measures could be selected following the literature, from total retinal volume to thickness of the RPE [54]. However, in this project we capitalize on the ability of deep learning to extract the features necessary to build a predictor that maximizes the probability of a label directly. Concretely, for each of the five groups of the data set, we train a classifier on the other four, using one of them for validation and early stopping. The five models are tested on their independent group, and a ranking of samples is obtained according to the difference between the label and the output of the model.

Another benefit of the deep learning approach is that we can leverage the CAM technique to obtain a graphical reasoning for the disagreement. The CAMs of the highest ranked models highlight zones that are responsible for the prediction. Therefore, if the expert initially missed the need to closely inspect that part of the image, for example due to fatigue, an opportunity to reconsider the label based on it is provided.

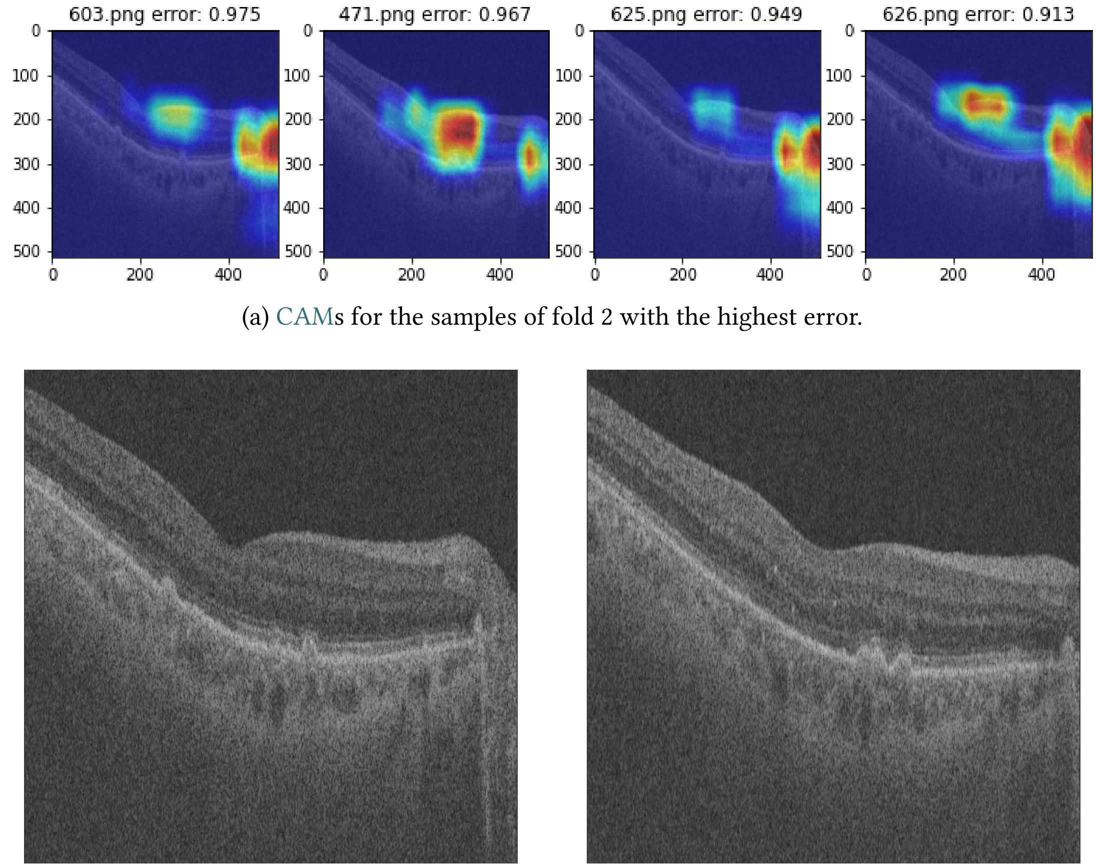
For the correct application of the methodology, it is important that the committee that reviews the samples that present intra-expert variability is blinded from the original and suggested labels. This impartiality also permits the development of a trial to investigate the impact of the model. Selecting images at random from the 1279 samples is unlikely to produce meaningful results, for the expert has a very low variability which we are trying to further reduce. Thus we develop an alternative strategy. The expert will select B-scans that are physically close to the ones suggested by the model and that have a similar rationale for changing, but that did not rank so highly on model disagreement. This guided search is likely to find additional outliers. It will also provide a test of whether the model is presenting samples that are more likely to change beyond what the expert can interpret from the CAM. While deep learning models fail in ways that humans do not, they sometimes also make correct predictions from factors that human experts cannot express formally in classification guidelines.

4.3.2 Results and discussion

A first inspection of the CAM results demonstrates how data set particularities affect model suggestion quality. Figure 4.10c is an example of how the presence of drusen is interpreted by the network as an indicator of a positive label. Indeed, from a clinical perspective, large drusen can be an indicator of AMD progression, and it is expected for the data set to show a high correlation between them and the other AMD signs that *are* being labeled.

However, Figure 4.10b also includes a less expected failure mode. In retrospect, fold 2 is the only one to contain such prominent depictions of the optic nerve. And the optic nerve is fundamentally a vertical disturbance of the layered structure of the retina, not unlike neovascularization. Thus, it is reasonable for a network that has not been exposed to samples that both contain the optic nerve and are labeled negative to classify such samples erroneously.

This level of interpretability is of great utility in the screening of potentially misclassified



(a) CAMs for the samples of fold 2 with the highest error.

(b) The sample with the highest error. The CAM clearly maps to the optic nerve.

(c) The next sample. The CAM maps to what the committee later confirmed as drusen.

Figure 4.10: Only fold 2 prominently includes the optic nerve. It is recognized as a depth-wise interruption of retinal structure and erroneously classified as pathological. Drusen are also highly correlated with other AMD signs, providing further false positives.

samples. There was no need to interpret an obscure statistical feature of the network activations or even solve a complicated dichotomy as to whether a B-scan situated between fluid and negative samples *is* fluid. Remarkably, an introductory understanding of the anatomy of the eye, as the one presented in the introduction to the present work, was sufficient to completely disregard some of the strongest model suggestions. This high level of confidence allows efforts to be focused on philosophically challenging samples.

Furthermore, this analysis has provided clear, actionable steps to improve the data set and the performance of subsequent models: obtain negative samples where the optic nerve is present. Even if such data collection is not possible, the analysis provides an understanding of which situations are expected to reduce the diagnostic efficacy of the model. Moreover, the detection of unfavorable situations can be combined to limit the situations where the model is recommended for use as a diagnostic aid.

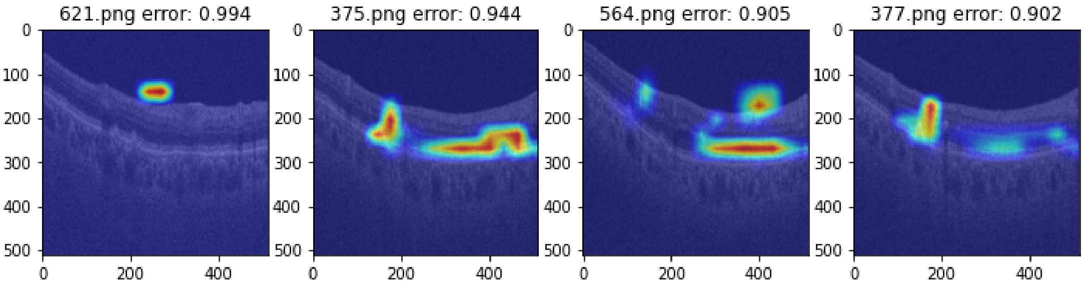


Figure 4.11: Images of fold 2 labeled as positive that the image considered negative.

Figure 4.11 shows the counterpart to the previous analysis. It is a representative example of the situations where the conservative approach in CAM acceptance prevents ambiguous CAMs from being reviewed. By selecting only those samples where the change is unambiguously justified by the CAM, we evaluate the methodology more fairly against those cases that the model did not rank highly. Due to this strict, conservative criterion, only one of the samples in Figure 4.11 was selected. It was accepted by the committee.

Finally, the quantitative analysis of agreement shows that, despite the inclusion of drusen samples that heavily penalized the model, those that the model ranked highly resulted in more changes to the data set by the committee. As a demonstration of the impact of drusen on the results, consider that all false positives in fold 2 were determined to be related to them.

Tables 4.5 and 4.6 contain two columns for the neovascularization agreement: possible and clear. This is due to the fact that the committee could not achieve consensus on some samples, rather a the third category of 'possible', as in possible neovascularization, was included. Therefore the agreement values had to be calculated in two ways: taking those possible cases to be positive or negative. For the purposes of changing labels, where the committee could not reach consensus the original label was kept.

Model ranking	Fluid	Neovasc. (positive)	Neovasc. (negative)
High	0.7273	0.7273	0.7500
Low	0.7778	1.000	0.8889

Table 4.5: Probability of agreement between the original labels and the clinician for the aided and unaided suggestions.

Lower values in the tables indicate that the review successfully detected samples of low agreement, i.e., those where the initial opinion of the expert and that of the committee differ. Due to the unaided group for fluid containing no originally positive samples, the Kappa measure between the low ranking group and the committee is 0. This is a statistical artifact not indicating perfect performance, for context, the probability of agreement is 77.78% (Table 4.5)

Model ranking	Fluid	Neovasc. (positive)	Neovasc. (negative)
High	0.4261	0.4371	0.4155
Low	0.0000*	1.0000	0.6087

Table 4.6: Cohen's Kappa measure of inter-expert agreement between the original labels and the clinician for the aided and unaided suggestions. *The zero value is a statistical artefact.

Even with the small amount of samples changed as a consequence of the revision (18 out of 1279, 1.4%) and the leaving aside of those used for testing (13 remain, 1%), the new labels produced a very significant improvement in performance across the board (Table 4.7). Notice how the performance increase from revision is superior to that obtained through architectural enhancements (Table 4.4).

Review	**Model**	Cross entropy	Accuracy	AUROC
Before	*ResNet-18 (F)*	0.3390 ± 0.0076	0.8600 ± 0.0014	0.9262 ± 0.0026
After	*ResNet-18 (F)*	0.2847 ± 0.0127	0.8827 ± 0.0088	0.9549 ± 0.0105
Before	*DenseNet-121 (F)*	0.3033 ± 0.0287	0.8724 ± 0.0182	0.9439 ± 0.0088
After	*DenseNet-121 (F)*	0.2623 ± 0.0016	0.8983 ± 0.0080	0.9529 ± 0.0049
Before	*DenseNet-121 (F & C)*	0.2947 ± 0.0020	0.8759 ± 0.0072	0.9446 ± 0.0025
After	*DenseNet-121 (F & C)*	0.2532 ± 0.0136	0.8973 ± 0.0100	0.9560 ± 0.0029

Table 4.7: Improvement in augmented baseline model performance due to label revision. In all cases results are the mean of three times repeated, 4-fold cross-validation. *F* stands for horizontal flipping and *C* for cropping to 248 pixels.

Aside from the success in improving data set quality and model performance, a subtle point can be extracted from the unprompted choice of "possible" as an evaluation. Perhaps whether a B-scan contains or not neovascularization can only be consistently predicted by experts in the context of surrounding scans and even other imaging modalities. This is testament to the difficulty of characterizing nAMD, especially with B-scan-level granularity.

One of the key aspects of the following experiment is that it employs per-eye labeling, circumventing this difficulty. Therefore any improvements in performance in the following experiment indirectly constitute evidence towards the hypothesis that the specification of precise B-scans is a major contributor to the difficulty of characterizing nAMD demonstrated in the last experiment.

4.4 Third experiment: domain-specific pretraining

The previous two experiments have been an exposition as to the difficulty of characterizing nAMD. However, their reasoning makes one assumption: model capacity is not the limiting factor. The following experiment will both explore the possibility of leveraging a larger data set to transfer features as well as attempt to demonstrate the adequacy of the models and techniques used for clsasification of AMD OCT images.

4.4.1 Data set and task considerations

We have performed binary classification on the "world's largest online annotated SD-OCT data set" [54], which is often used as a state-of-the-art benchmark in the domain. However in this case we are presented with true binary classification, as in the distinction between eyes with and without intermediate AMD. The main difference with the approach taken in the original paper is that for the purposes of our experiment, the network will have to classify an eye as having signs of AMD or not from a single B-scan. This is a much harder task, especially in the detection of the positive class, as not all B-scans may present strong signs of AMD. Some measures have been taken during preprocessing to address this, described within preprocessing in general.

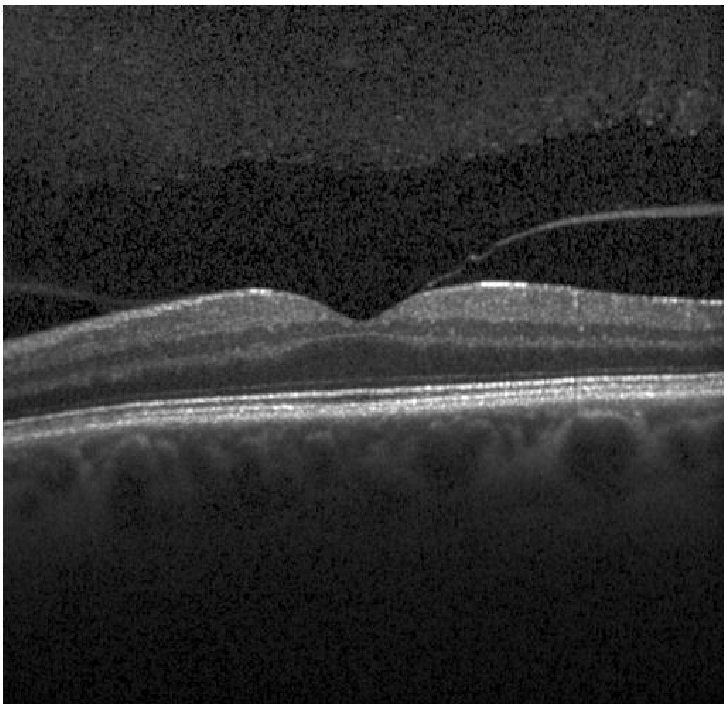


Figure 4.12: B-scan from a public data set [55] using a Spectralis OCT device.

After training a classifier in this new task, even if the scans do not match exactly, the closer domain should provide stronger pretraining than the natural images of Imagenet (i.e., images taken from a natural scene with a camera). Nevertheless, the most important consideration when selecting a public data set for transfer learning is naturally the similarity to the final task. Many OCT data sets were considered for the purpose, until the aforementioned prevailed. As we can observe in Figure 4.12, the appearance of samples, even within SD-OCT devices can vary substantially. In contrast, as soon as B-scans were extracted from the volumetric data of the selected data set, we could immediately appreciate the qualitative similarity (Figure 4.13). Still, the different device from Bioptigen, Inc, implies that the resolution and therefore the effective dimensions of retinal features differ.

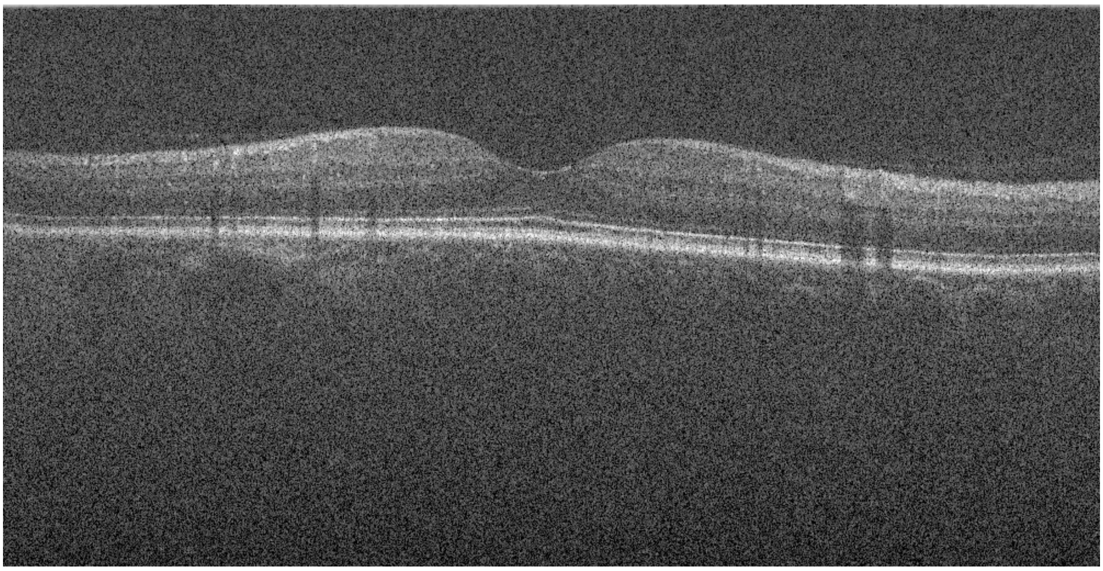
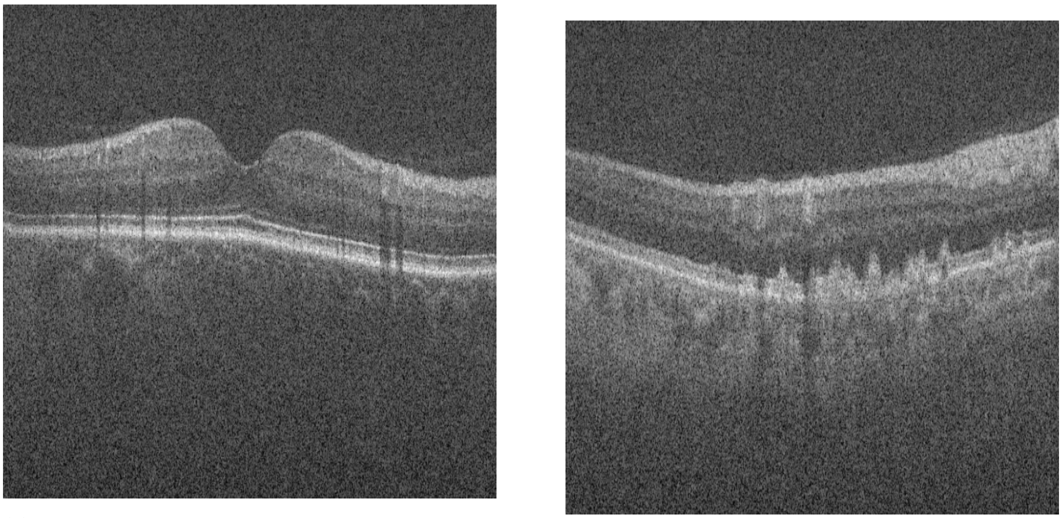


Figure 4.13: Images from the selected public data set [55] show great similarity to the one employed in the first two experiments.

However, this was also considered during the process of data set selection. The dimensions of the public data set are close to the common settings of the machine used in the first experiment, comprising an *en face* square of side 6.7 mm. [54]. The wider resolution can be easily equalized by down-sampling from 1000 to 512 horizontal pixels. And while the depth resolution is different, at 4.5 µm.[56], this can also be overcome.

Preprocessing

The depth dimension has been approximately adjusted by cropping the public data set to 416 vertical pixels, removing empty vertical space like in the previous data set, and interpolating back to 512. A vertical offset of 16 pixels is also applied to remove noise present at the top of some B-scans. Jointly, this greatly increases the similarity across data sets, both nominally and visually. In fact, it becomes nontrivial to distinguish between B-scans from the public and the nAMD data sets by inspection alone (Figure 4.14).



(a) The same B-scan as Figure 4.13 after preprocessing.

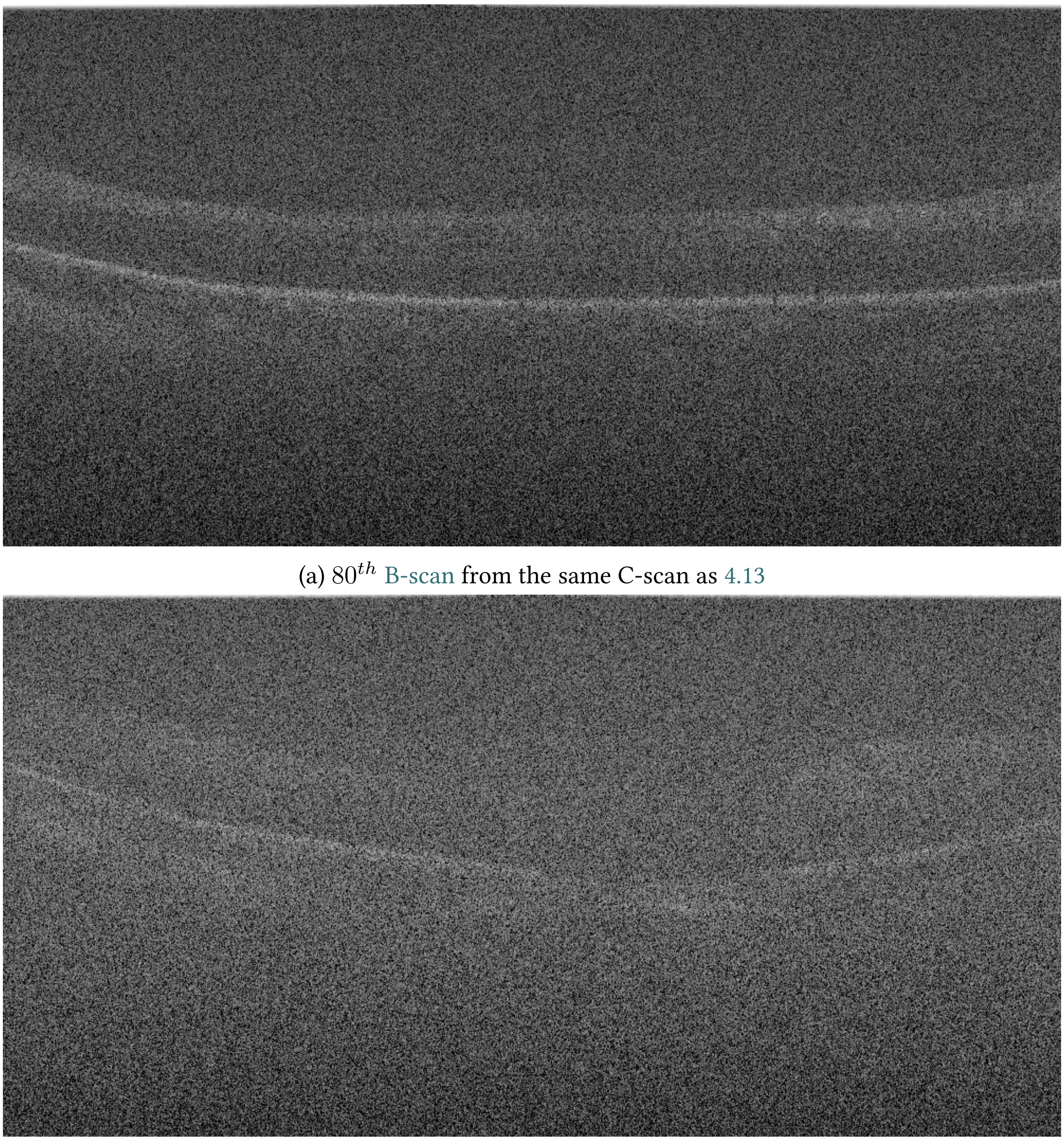
(b) Positive sample after preprocessing

Figure 4.14: Images from the public data set become hard to distinguish from the original ones after preprocessing.

Additional exploratory analysis of the public data set revealed some problems with the naïve approach of training a classifier on all samples. Unlike our original data set, as C-scans are to be considered globally, not all B-scans need to present the same image quality. Indeed the first C-scan already demonstrates a fading out in B-scans as they get farther away from the fovea (Figure 4.15). After further inspection of the data set, no samples were found to contain significant fading in the central area around the macula, but the problem of fading in the periphery was common.

To address it, the following procedure was performed: of the 100 B-scans in each C-scan only the middle half were utilized for training and evaluation. Interestingly, the authors of the data set had already considered the impact of restricting data to a limited distance from the fovea: They found that distances smaller than this half reduction resulted in negligible performance degradation [54].

Then, to accelerate training by reducing the vast amount of data, only every fifth B-scan of the middle half was selected. Unlike the original data set, where only an effectively random number of B-scans are provided, the public data set provides full C-scans. We therefore take advantage of the facts that contiguous samples are almost identical and that most clinically relevant morphological alterations are within the macula to at once eliminate the fading problem and reduce the amount of data by an order of magnitude.



(a) 80^{th} B-scan from the same C-scan as 4.13

(b) 95^{th} B-scan from the same C-scan as Figure 4.13

Figure 4.15: Some B-scans from the public data set fade out as they get farther away from the fovea.

4.4.2 Results

This heavy reduction in data resulted in outstanding performance both in terms of training time and testing performance. The first experiment on the data utilized a training recipe with hyperparameters (learning rate of 10^{-4}) and minimal data augmentation (horizontal flipping after resizing to 256) identical to those of the first experiment. The new data set is much more representative, as it has only 10 samples per eye and yet more samples overall. Exploiting that fact, cross validation was substituted for a single training-validation-testing split. The first 200 control samples and 400 AMD samples were separated for testing. The same amount of the next images were selected for validation, and the rest for model fitting. This scheme produced the results in Table 4.8 after training for just 10 minutes. Furthermore, the Imagenet pretraining enabled the network to reach great validation metrics on the first epoch (Figure 4.16). The first epoch showing higher training than validation error is a consequence of the training error being averaged from the training batches, and not calculated again passing through the complete data set.

Phase	Cross entropy	Accuracy	AUROC
Validation	0.0701 ± 0.0120	0.9767 ± 0.0050	0.9973 ± 0.0012
Testing	0.0837 ± 0.0228	0.9713 ± 0.0066	0.9966 ± 0.0010

Table 4.8: Performance of the augmented baseline configuration of the previous experiment on the pubic data set. Values represent the means and standard deviations of five repetitions.

To put these results into perspective, the preprocessing allowed a ResNet-18 to discriminate eyes with and without AMD from a single B-scan; with the lowest testing AUROC of all repetitions (99.58%) being above the best figure in the original paper's shallow learning, interpretable approaches [54]. Moreover, the cross entropy and accuracy figures are consistent with the AUROC value.

The object of this experiment is not, however, to compare the performance of the approach with others on this public data set; but rather to establish whether the combination of models, training techniques, hyperparameters and data augmentation are sufficient to obtain adequate performance in this benchmark domain.

4.4.3 Transfer learning performance and discussion

The second and main objective of the experiment is to determine whether training on this benchmark data set provides a superior transfer learning over the usual Imagenet. Initially in the restricted, binary task, and later in the full task of characterization. A unique feature of pretraining in such a similar task is the possibility of determining the "zero-shot" performance

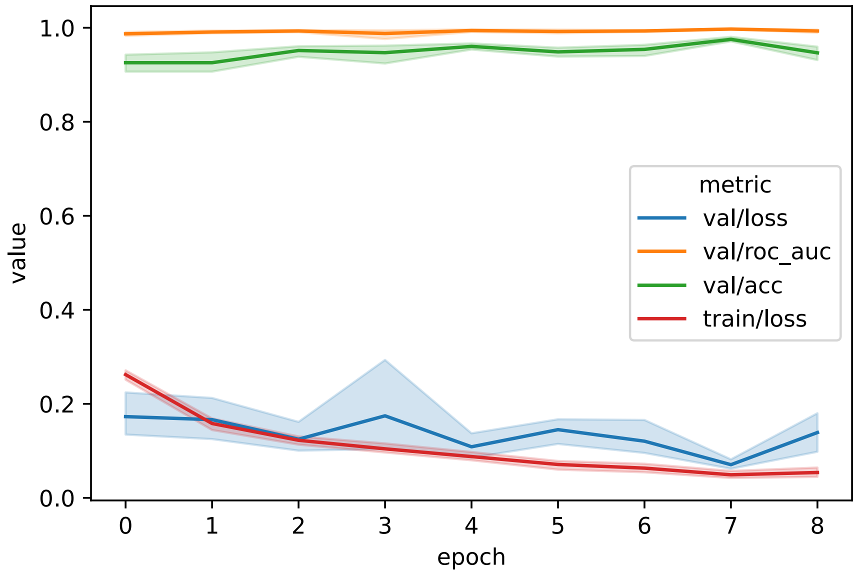


Figure 4.16: The training and validation curves from 5 repetitions, showing the 95% confidence interval.

of the network on the original data set, i.e., its performance without training on it at all. As demonstrated in Table 4.9, AUROC performance is significantly above chance globally, with high variability across folds. Moreover, Table 4.10 shows that the review process produced a modest but consistent improvement in zero-shot performance. Nevertheless, absolute performance remains barely above chance, which is reasonable given the task difference and the small extent of revision.

Fold	Cross entropy 0	Accuracy	AUROC
0	2.2417	0.7191	0.6423
1	4.4475	0.5443	0.4835
2	5.4415	0.4453	0.5518
3	4.1339	0.5450	0.6278
4	2.2414	0.7528	0.6311
Mean	3.7012 ± 1.4172	0.6013 ± 0.1300	0.5873 ± 0.0682

Table 4.9: Zero-shot performance on the 'binary' task.

While samples from the two data sets present a similar appearance visually, the different means of their pixel values, together with the consistency of the samples within data sets might indicate that the pixel intensity distribution is limiting zero-shot performance. Both

Fold	Cross entropy 0	Accuracy	AUROC
0	2.1702	0.7258	0.6625
1	4.4475	0.5443	0.4835
2	5.4769	0.4415	0.5520
3	4.1339	0.5450	0.6278
4	2.1319	0.7640	0.6364
Mean	3.6721 ± 1.4748	0.6041 ± 0.1359	0.5924 ± 0.0735

Table 4.10: Zero-shot performance on the original data set with revised labels.

data sets have small but significant differences in mean and standard deviation of their intensities (Table 4.11).

Data set	Mean	Standard deviation
Ours	0.3130	0.0997
Farsiu et al.	0.3484	0.1401

Table 4.11: Mean and standard deviations of pixel intensities for both data sets.

However, normalizing each data set with the calculated values instead of the ones used in PyTorch for Imagenet did not significantly improve performance. Even if the results improved, it would be preferable to develop a model robust to variation in intensity and contrast, as is the case with human experts. Thus, we performed an experiment with data augmentation specifically for this purpose: randomly changing the intensity and contrast of the images such that both data sets fall within the margin of variation. This is different from usual data augmentation in that the objective is not to improve validation measures on the data set, but rather to construct a model with the intention to generalize outside of the specific data set.

Table 4.12 indicates that zero-shot transfer performance improved slightly by jittering brightness and contrast 20%, which leaves the values of the original data set within the new training distribution. The improvement in performance from the original to the revised labels was also kept. While there was a substantial decrease in cross entropy, accuracy and AUROC increased very slightly. Even after the improvement, cross entropy remains worse than the one expected from a random classifier.

In summary, it is reasonable for these augmentations increase zero-shot performance only slightly, as the tasks are different and good performance is not to be expected even from a perfect classifier on the public data set. This is because according to the definition of the

Jitter	**Revised**	Cross entropy	Accuracy	AUROC
0	*No*	3.701	0.601	0.587
0	*Yes*	3.672	0.604	0.592
0.2	*No*	1.656	0.602	0.598
0.2	*Yes*	1.144	0.605	0.602

Table 4.12: Means of zero-shot performance on the original data set.

task on the public data set, all samples in the nAMD data set should be classified as positive, resulting in the expected negligible specificity. Even then, AUROC remains a valuable bias-independent data point that showed a consistent, if small, superiority over the random classifier.

What is important for the network is to extract features of broad applicability in order for the subsequent fine tuning process to be effective, and not so much the zero-shot performance, given the different tasks. Given the reasonable looking CAMs of Figure 4.17 and Figure 4.18, such features were probably extracted. In Figure 4.17, activations seem to follow drusen, morphological features (samples 1254 and 1250) and choriocapillary alterations. Of Figure 4.18 we can highlight the activations of sample 13, which surrounds an area of fluid, as well as sample 1164, whose activation focuses not on the area of elevated retina, but rather on the RPE, which we know to be the possible causal factor. Testing without training revealed that the performance of a model completely independent from the data set also improves with label revision, further solidifying the results from the previous experiment.

Finally, fine tuning the best model on the original data set produced no significant improvement over the first experiment. This seems to indicate that label scarcity and to a much lesser extent variability are the main limiting factors in the original experiment, given the great similarity between the two data sets and the dissimilar performance.

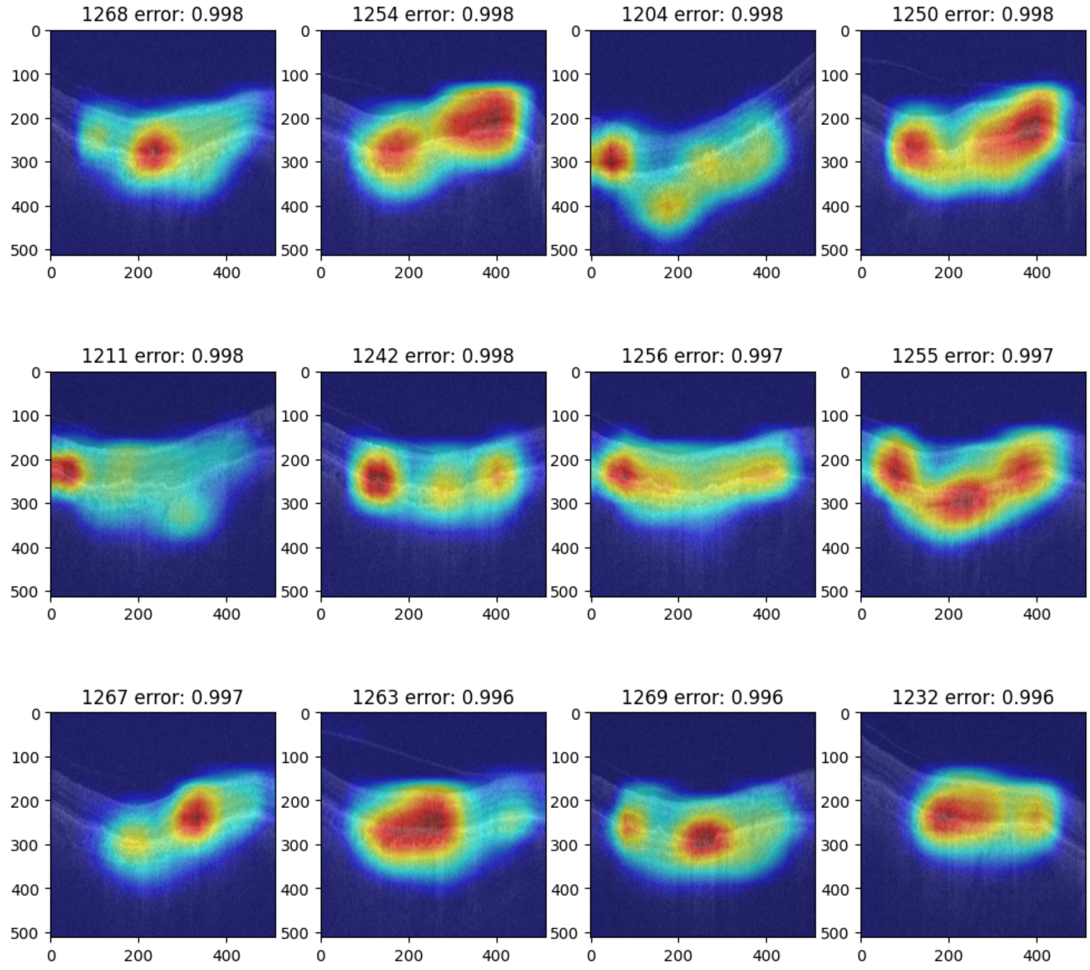


Figure 4.17: CAM of the model trained only on the public data set for some negative samples of fold 0 of the nAMD data set.

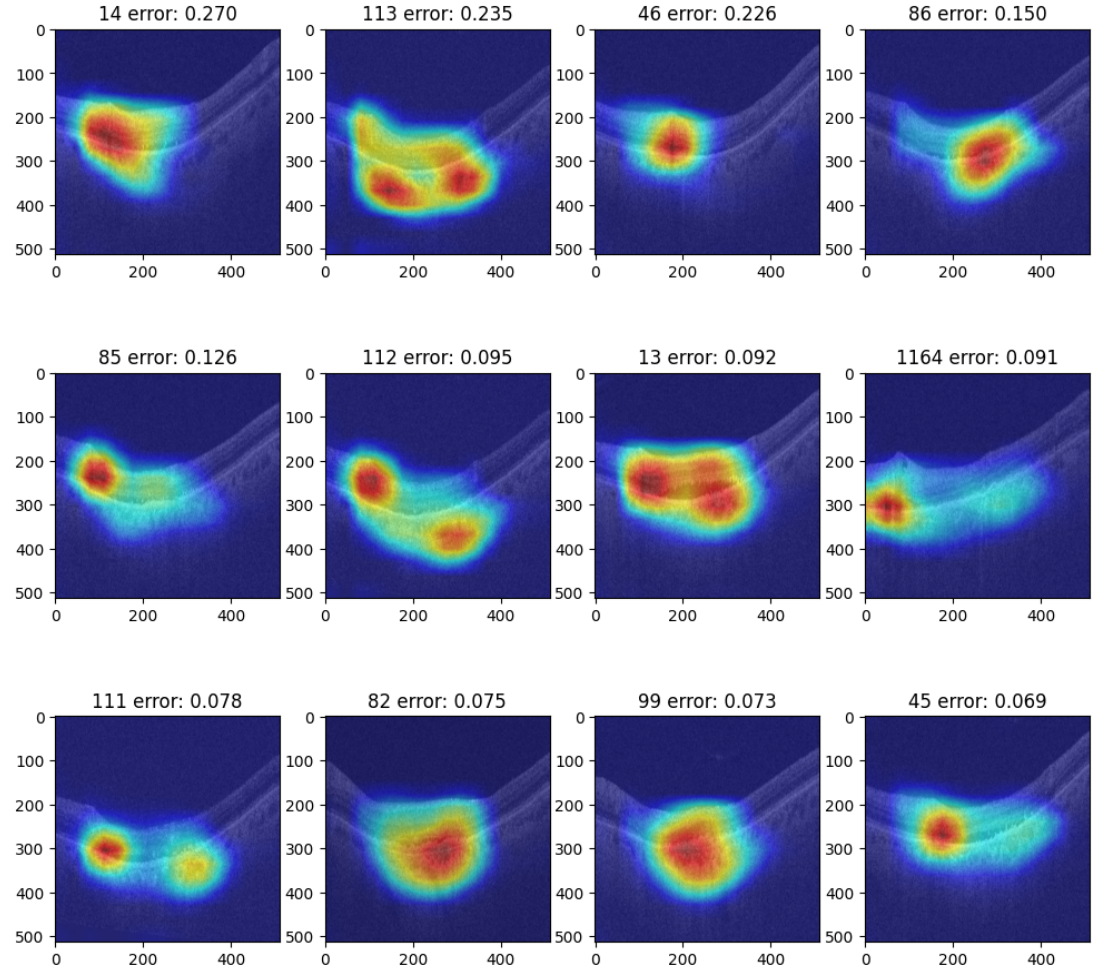


Figure 4.18: CAM of the model trained only on the public data set for some positive samples of fold 0 of the nAMD data set.

4.5 Fourth experiment: nAMD characterization

Finally, after determining data set and model feasibility through an initial detection approach, addressing intra-expert variability and data scarcity through a relabeling pipeline and pre-training, we will evaluate the original task of multi-label classification.

Metrics

The rationale for the chosen metrics is similar to that of the first experiment. As opposed to multi-class classification, where classes are exclusive, multi-label classification is essentially performed with multiple binary classifiers that share features. Therefore, the same metrics utilized in the first experiment will be maintained, but decomposed according to the class. In the case of the ROC, a curve will be calculated for each class.

4.5.1 Reproduction of the baseline

Model	**Augs.**	Fluid accuracy	Neovasc. accuracy
ResNet-18	*F*	0.8576 ± 0.0933	0.7881 ± 0.0944
ResNet-18	*F & C*	0.8602 ± 0.1042	0.7755 ± 0.0953
DenseNet-121	*F*	0.8688 ± 0.0770	0.8264 ± 0.0795
DenseNet-121	*F & C*	0.8910 ± 0.0506	0.8176 ± 0.0637

(a) Multilabel accuracy results.

Model	**Augs.**	Fluid AUROC	Neovasc. AUROC
ResNet-18	*F*	0.9183 ± 0.0756	0.8264 ± 0.1137
ResNet-18	*F & C*	0.9157 ± 0.0951	0.8294 ± 0.0915
DenseNet-121	*F*	0.9317 ± 0.0530	0.8804 ± 0.0604
DenseNet-121	*F & C*	0.9499 ± 0.0341	0.8741 ± 0.0605

(b) Multi-label AUROC results.

Table 4.13: Means and standard deviations of the baseline multi-label results.

We begin the experiment by quantifying the performance of the augmented baseline models, ResNet-18 and DenseNet-121, established during binary classification. Table 4.13 contains the results of this experiment. Reproducing the procedure of the initial experiment, horizontal flipping alone is compared to flipping and cropping the image to a size of 228 pixels. We can observe how the 3 times repeated, 4-fold cross-validation means differ between the two classes, with fluid always resulting in superior metrics. This was expected from the domain knowledge presented in the contextualization chapter, as fluid has a more marked appearance on OCT images, thus making it easier to recognize for domain experts and models alike.

On the other hand, the standard deviation calculated within and across folds shows a slightly different performance profile. While for the AUROC there is an improvement in consistency from neovascularization to fluid, this trend does not hold for the standard deviation of accuracy. Another difference is that with the introduction of cropping, the consistency of the accuracy results of ResNet-18 seems to regress a little, with standard deviation increasing by 1%. This is one of the only cases where more augmentation hurt performance.

The other one being neovascularization in DenseNet-121. There, the introduction of cropping decreased both accuracy and AUROC. This is in stark contrast to performance in the classification of fluid, which markedly increased from an AUROC of 93.17% to essentially 95%, all while almost cutting deviation in half. Accuracy reported a similarly large increase from 86.88% to 89.10%, and again, almost halving the variability of results. Finally, even if neovascularization accuracy decreased, it also experienced a reduction in variability. Therefore adding cropping improved the consistency of convergence globally.

4.5.2 Analysis of revision

Model	**Augs.**	Fluid	Neovascularization
ResNet-18	*F*	0.8560 ± 0.0906	0.7903 ± 0.969
ResNet-18	*F & C*	0.8603 ± 0.1033	0.7802 ± 0.0988
DenseNet-121	*F*	0.8688 ± 0.0753	0.8311 ± 0.0836
DenseNet-121	*F & C*	0.8898 ± 0.0511	0.8223 ± 0.0681

(a) Multi-label accuracy results.

Model	**Augs.**	Fluid	Neovascularization
ResNet-18	*F*	0.9178 ± 0.0770	0.8323 ± 0.1191
ResNet-18	*F & C*	0.9163 ± 0.0914	0.8377 ± 0.0995
DenseNet-121	*F*	0.9313 ± 0.0506	0.8883 ± 0.0676
DenseNet-121	*F & C*	0.9471 ± 0.0372	0.8822 ± 0.0678

(b) Multi-label AUROC results.

Table 4.14: Means and standard deviations of the multi-label results after revision.

After investigating the specific performance characteristics of the baseline experiment, we turn to the revised labels to deepen the analysis. Table 4.14 replicates the layout of the baseline, showing the results for the new labels. While improvements outnumber regressions and have a higher magnitude, they all lie well within one standard deviation.

The best model, DenseNet-121 with flipping and cropping, shows improvements in neovascularization accuracy and AUROC, with a regression in fluid performance of half the magnitude. We can not conclude from these results alone that the improvement in nAMD detection produced a significant improvement in nAMD characterization.

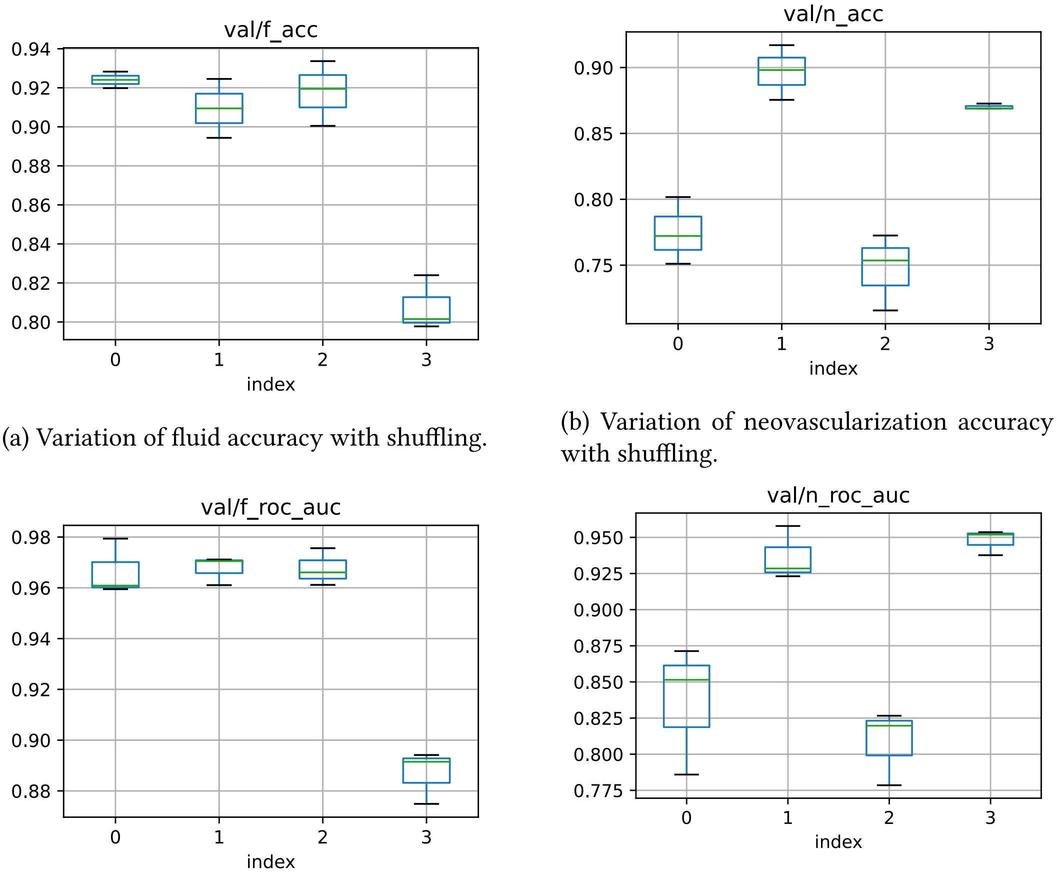
However, validation loss, which aggregates fluid and neovascularization and is much more sensitive to outliers, does show a consistent improvement (Table 4.15) in both mean and standard deviation. There is a general decrease of around 2% with the revised labels. This might not seem large relative to the standard deviation, but there are two factors to consider that explain why such a change is relevant.

Model	**Augs.**	Before	After	Difference
ResNet-18	*F*	0.4339 ± 0.1119	0.4240 ± 0.1086	-0.0099 (2.28%)
ResNet-18	*F & C*	0.4219 ± 0.1244	0.4152 ± 0.1232	-0.0067 (1.59%)
DenseNet-121	*F*	0.3670 ± 0.0955	0.3590 ± 0.0918	-0.0080 (2.18%)
DenseNet-121	*F & C*	0.3750 ± 0.0985	0.3669 ± 0.0951	-0.0081 (2.16%)

Table 4.15: Mean multi-label cross entropy results for the revision.. The difference is mean is indicated in absolute amount and relative to the original loss (in parentheses).

First, we must recall the absolute amount of change in the data set produced by revision. The labels of only 13 out of approximately 1300 samples were changed in the training and validation set. Thus that 1% of labels produced a 2 % decrease in cross validation, meaning that they were above average contributors to it.

Second, we must also consider that the standard deviations presented for both tables are calculated from three times repeated cross validation, and that experiment 1 demonstrated consistent variation in performance between folds. As Figure 4.19 demonstrates, the variation in performance remains consistent with experiment 1. It is much smaller within that between folds, whose creation is not random. Therefore, the impact of chance in the comparative power of mean results is smaller than the standard deviation might suggest. To illustrate this, we take the standard deviations for the repeated folds, as only those are random. In the case of DenseNet-121, that implies at least a 5 times reduction in standard deviation for the worst cases. In fact, the standard deviation of validation loss for the repetitions of fold 3 is smaller than the 2.16% improvement in mean cross entropy.



(a) Variation of fluid accuracy with shuffling.

(b) Variation of neovascularization accuracy with shuffling.

(c) Variation of fluid AUROC with shuffling.

(d) Variation of neovascularization AUROC with shuffling.

Figure 4.19: Multilabel performance across folds.

Considering these two factors, as well as the fact that there was improvement for all model and augmentation configurations, we conclude that the revision process produced a small but consistently measurable improvement in cross entropy in the nAMD characterization task.

4.5.3 Further developments

Like in the analysis of experiment 1, we turn to the training curves for more information about training dynamics. Figure 4.21 illustrates the training curves of the best performing model, DenseNet-121 with flipping and cropping augmentations trained on the revised labels. Validation cross entropy does not seem to significantly decouple from accuracy and AUROC except for the single outlier in fold 2 that managed to evade early stopping through a last improvement in epoch 18. A possible explanation is that in the more difficult task of nAMD characterization, it takes more epochs for the model to converge, making the nominally equal early stopping patience of 10 epochs more aggressive in relative terms.

Furthermore, validation cross entropy remains below that of the random classifier for the large majority of the time, which was not the case in the binary task. This hints to a reduced importance of outliers as a result of the initial revision process and mostly due to the increased granularity of labeling. Which could mean that a most effective way to continue developing models for this task would be to collect more labeled data or even unlabeled data from the same machine and configuration.

Nevertheless, cross entropy is still close to chance performance, much higher than would be expected from the accuracy and AUROC curves, implying that outliers still play a role. Indeed, a relatively small number of samples produce disproportionately high erroneous confidence in the model, with reasonable CAMs that warrant investigation (Figure 4.20). Of course, the great majority of those mistakes are expected to be due to generalization error, however, it remains an open question whether repeating the variability mitigation with more detailed labels could provide additional improvements.

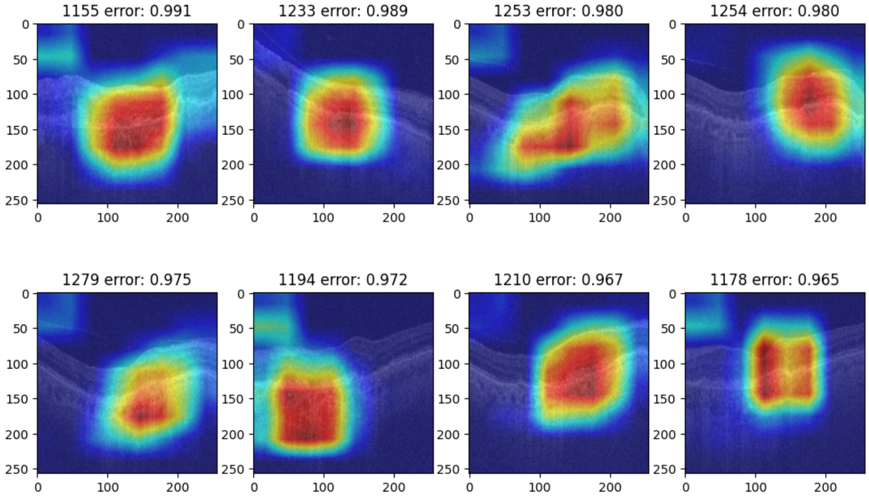
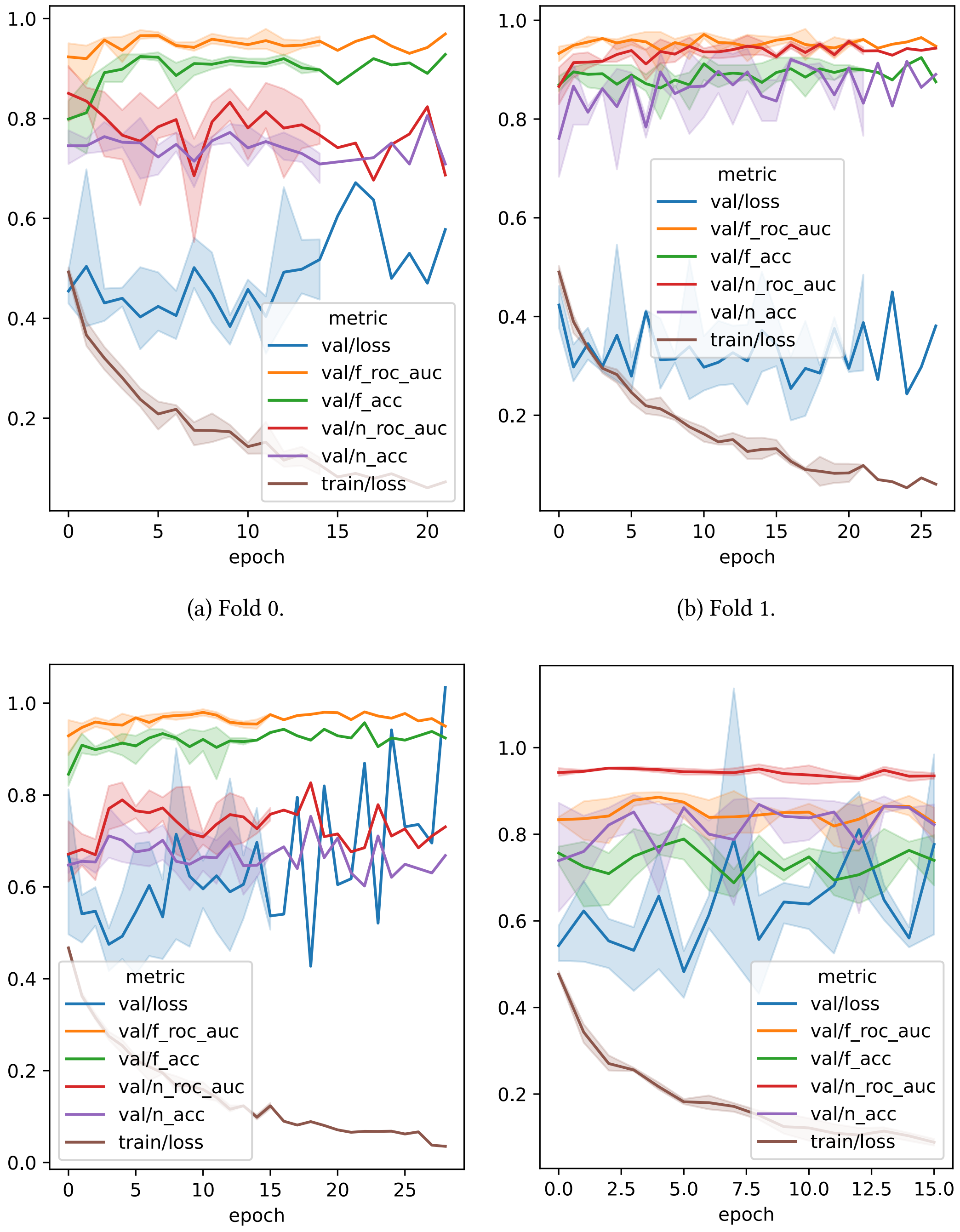


Figure 4.20: Class-specific CAMs depicting fluid candidates.

In conclusion, the variability mitigation from experiment 2 produced measurable improvements in characterization performance. However, due to the limited extent of the changes, performance improvement was not as dramatic as in the case of binary detection. The analysis of the training curves seems to imply that label noise is no longer one of the primary limitations to the generalization performance of the model. Nevertheless, some class-specific CAMs produced by this new task seem promising for another cycle of intra-expert variability mitigations. However, such repetition of the methodology lies outside the scope of the present project and within the possible lines of future work.



(a) Fold 0.

(b) Fold 1.

(c) Fold 2.

(d) Fold 3.

Figure 4.21: Multilabel training curves for DenseNet-121.

Chapter 5

Conclusion and future work

This chapter constitutes the closure of the project. It includes a review of its paradigms, methods and results, while also providing a view to related lines of research and possible continuations of the work herein.

5.1 Characterization of nAMD in OCT images

This project has fundamentally consisted in an investigation of the task of AMD characterization, which, as presented in the introduction, informs the effective therapy of one of the leading causes of severe visual impairment. Beginning with the initial task of binary classification, development has been a data-centric exploration into the characteristics of neovascularization and fluid, two of the most prominent signs of advanced AMD.

The initial task is a demonstration as to the difficulty of characterizing AMD in the real-world regime of data scarcity and expert variability uncertainty. These problems were discovered and addressed through the usual tools of machine learning: statistical analysis of the training and validation metrics resulting from cross validation. Together with the domain expertise resulting from the bibliographic research and hands on experience from analyzing the individual samples, the statistical results could be interpreted to diagnose both models and data.

These results led to the formulation of hypotheses about possible avenues for improvement. Following the discussion of label noise in state of the art research in computer vision, it was considered whether a model trained on part of the data could not only detect, but help address the presence of label noise. Because AMD characterization is a complex domain with subtle labeling, careful manual review could not be utilized to correct the noise introduced by the usual semi-automated methods.

Instead, a methodology for machine-assisted intra-expert variability mitigation was developed. For each of the five independent groups in which the data set was divided, a model was trained on the rest of the data. Then, the predictions of the model were used to rank samples according to the disagreement between the model and the original labeler. Naturally, even if the classifier had developed perfect generalization, some of the samples of high disagreement could be outside the training distribution and hence simply misclassified. The solution resides in the application of interpretability techniques, concretely Class Activation Mapping (CAM), to assist the reviewer through understanding the rationale for model outputs.

This proved highly effective for several reasons. First, spatial information was a boon for the recognition of positive samples that were not correctly labeled as such. For example, one sample was reviewed where the model presented highly concentrated activations around a hyporeflective zone. With the benefit of context, namely adjacent scans and their labels, the expert was able to confidently identify as fluid an otherwise difficult sample. Second, it allowed for the immediate rejection of unpromising candidates, promoting a better allocation of expert time towards challenging samples. The paramount example of this was the prominent presence of the optic nerve in one of the five groups. The CAM enabled the correct discovery of a large amount of false positives due to the optic nerve being a depth-wise disruption to the layered structure of the retina. And third, the interpretable review process provided a new perspective into the data set and the model. Concretely, through the last example, it provided actionable insight for its improvement. If examples of patients without signs of AMD where the optic nerve was prominently depicted were added to the groups, the model would become robust to it. Alternatively, if such data could not be obtained, it would bring valuable information as to the applicability of the model to different situations.

The methodology was evaluated through a blinded, controlled trial. For each sample where the expert was convinced by the CAM as to the opposite label, similar samples were found that were not detected by the model. The review committee was not informed about the proposed labels, nor as to whether the review suggestions were proposed with or without assistance. Instead, a set of samples were presented to be classified, to prevent the introduction of bias. The results were compared to the original labels and the suggestions, and suggestions by the model were found to consistently outperform similar images, both for fluid and neovascularization.

The other major hypothesis examined data quantity as opposed to quality. Similarly to how pretraining on natural images, such as the ubiquitous Imagenet, improved generalization performance on the initial domain, perhaps transfer learning from a closer domain could provide additional gains. Instead of the inductive bias of images in general, a classifier trained on another OCT data set could possibly contain more adequate feature extractors. If the data set contained signs of AMD, even higher level features could be relevant.

To test it, a similar but much larger data set had to be found. Through bespoke preprocessing, the scans of a large public data set became close to indistinguishable from the nAMD data. This was achieved through careful consideration of resolution and its corresponding physical dimensions, as well as the basic distinction between OCT modalities. Then a strategy was devised to, ironically, reduce the amount of training data. Through equally spaced sampling of the middle half of the 3D scans, the number of 2D scans was reduced by an order of magnitude, while theoretically remaining representative due to the high correlation between contiguous folds. This meant that the time to train a model was measured in minutes, and not hours.

Porting the training recipe from the original experiment immediately produced great results. As a reference, it beat the already impressive AUROC of 0.9917 of the original paper. This constituted great evidence for the capability of the exact configuration used in the initial experiment to classify, if not characterize, AMD. However, the zero-shot results on the original data set were less impressive. Because the original data set only contained eyes with AMD, the model trained to detect AMD from a single B-scan had negligible specificity. Nevertheless, its ranking capability, measured through AUROC, showed a small but significantly above chance zero-shot performance.

The attempt to fine tune this pretrained model did not result in an improvement in performance, however. Given the rest of the experiments, a reasonable explanation is that the task is in the data-constrained regime, and as such, better representations for the model are unlikely to provide sizeable benefits. Nonetheless, as soon as the data is extended, model capacity could start to play a bigger role on performance. Therefore, a model pretrained on an extremely close domain could be a useful asset during the process of assisted review or even labeling, as will be discussed in the section on future work.

Finally, once the hypotheses had been tested, their impact on the multi-label task was evaluated, including CAMs. The conclusion is that, while performance is perhaps short of what would ideally be clinically applicable, this research project gained insight into the requirements for building a AMD characterization system, as well as the specific but generally applicable methodologies for its construction.

5.2 General methods

While the project has been focused on and guided by the task of AMD characterization in OCT images, it is important to point out that the methods herein developed could be of interest for machine learning in medical imaging in general. Quality of labels is of paramount importance in the development of such systems and, given that medicine is not an exact science, some amount of expert variability is to be expected. However, simply accepting that labels should ideally be reviewed by a committee does not make it cost effective, especially in light of the data and labor-intense nature of medical imaging.

The application of model-based ranking with CAM analysis demonstrated in this work can scale with compute to serve much larger data sets, allowing for domain expert time to be intelligently allocated to challenging samples. And it can in principle do so for any domain where spatial activation is relevant, including imaging modalities and scales ranging from histology to radiography.

In the modern ML regime of heavily overparametrized models, training error converges to essentially zero, and generalization ability depends on the probability of reasonable interpolations. Bigger models together with the implicit bias of SGD empirically demonstrate better performance, but this new paradigm beyond classical machine learning has no correspondingly extensive theoretical basis. Transfer learning is empirically found to be a powerful method for biasing large models toward reasonable interpolations. Therefore pretraining in an adjacent task acts synergistically with the proposed review methodology, raising the probability of outputs being the result of reasonable CAMs.

In closing, this project has explored how the changing paradigms of machine learning translate to medical imaging. Taking the recognition that even in benchmark-grade, state-of-the-art data sets label noise is a major limit to model performance, it considers whether the models themselves could serve as a driving factor behind the solution. And through interpretability, how more nuanced suggestions for data gathering can be provided. Finally, with the advent of ever larger models, domain-specific transfer learning was leveraged for research into increasingly specific tasks.

5.3 Future work

Taking this project as a starting point, there are many possible avenues for continuation that could provide interesting new insights and answer some of the open questions. The following sections provide a view into some of the most promising ones.

5.3.1 Interpretability techniques

Interpretability techniques have proven their usefulness throughout this work as a means of demystifying what used to be black box models. Many in the ML branch of medical imaging vouch for the urgent implementation of explainable and interpretable techniques in the field, given the crucial need for justification in medicine.

While Grad-CAM is one of the most popular techniques, it is by no means the only one. Further work could explore the application of other useful methods, such as occlusion tests, to complement the perspective offered by gradient-based methods. The greater variety in class attribution techniques could make their results more robust, benefiting all the applications of them seen throughout this project.

This includes the workflow of intra-expert and inter-expert variability mitigation, which in itself constitutes another line of research. Bigger trials could shed light into what are the contributing factors to variability among and within experts from a data set standpoint. Deep learning models can help select the toughest examples to classify, greatly increasing the efficiency of inter-expert agreement trials by removing the large majority of samples where overwhelming agreement is expected.

Another promising approach follows the fields of semi-supervised learning and human-in-the-loop ML, leveraging the great asymmetry present between samples and labels. During usual clinical practice, a large number of OCT C-scans is collected; nevertheless, to the best of our knowledge, it is extremely rare for them to be labeled B-scan-wise with the level of detail presented in this work. Self-supervised learning capitalizes on this asymmetry by extracting relevant features without the need for labeling, greatly widening the potential for data ingestion. Semi-supervised learning elaborates on this approach by providing a relatively small number of labels, which have an outsized impact on performance thanks to self-supervision. Several techniques exist that exploit model knowledge to select samples for labeling with the most expected impact, not unlike the review methodology presented in this project.

While labeling and testing is to be determined exclusively by human experts, the use of human-in-the-loop techniques that help construct large training sets is a promising way to improve testing performance in data sets that have not been influenced by any model.

Applying human-in-the-loop principles to this work would result in the iterative application of the review methodology, investigating whether better labels produce better review, which in turn produces better labels, compounding the effects until only the most ambiguous samples remain.

5.3.2 Segmentation

Another interesting open question is whether the increse in efficiency provided by these methods could be applied to the task of segmentation of neovascularization and fluid. Given the correlation between reasonable CAMs and classification performance, a simple bounding box labeling pipeline could be sufficient for improved results. Even if there was only an interest in classification, the bounding box or full segmentation labels allow for sophisticated data augmentation such as obstructions or deformations to the image that leave the crucial region unaffected.

5.3.3 Data augmentation

As discussed in the main work, this project has limited itself to a conservative policy of augmentation where only those transformations that produce images that could plausibly belong to the data set were considered. This strict approach has the benefit of almost guaranteeing that augmentation is not negatively influencing results, which proved very useful for the purposes of this project.

However, transformations that produce implausible samples might nonetheless improve classification performance in unaugmented samples. The most salient example is the application of elastic transformations, that are capable of smoothly deforming the large scale morphology of the image while retaining texture and smaller features. This would negatively impact the interpretability of aspects such as CAM activations following retinal morphological changes due RPE detachment. Nevertheless, it would be interesting to consider the consequences of its application on performance and attribution.

5.3.4 Conclusion

In closing, this project has demonstrated how the application of machine learning, interpretability and domain specific knowledge can help improve the performance of nAMD characterization in OCT images and better inform as to its limitations. The development of the further advances outlined in the sections above could prove to be a crucial factor in the introduction of machine learning assistance to clinical practice.

List of Acronyms

AMD Age-Related Macular Degeneration. 1, 3, 5, 7, 12, 14, 23, 27, 30, 41, 42, 45, 49, 61–63

AUC Area Under the Curve. 25

AUROC Area Under the Receiver Operating Characteristic Curve. 31, 32, 34–36, 38, 49–52, 55–59, 63

CAM Class Activation Mapping. i, iv, 1, 2, 21, 23, 41–43, 52–54, 59, 62–66

CAMs Class Activation Maps. 2

CNNs Convolutional Neural Networks. 2, 21

CPU Central Processing Unit. 25

FD-OCT Fourier Domain OCT. 11

GA Geographic Atrophy. 12

GPU Graphics Processing Unit. 16, 25

Grad-CAM Gradient-Weighted Class Activation Mapping. 21, 25, 65

IRC Intraretinal Cystoid Fluid. 14

ML Machine Learning. 15, 37, 40, 64, 65

nAMD Neovascular Age-Related Macular Degeneration. ii, iii, 4, 12, 14, 23, 30, 40, 44, 45, 47, 52–55, 57, 58, 61, 63, 66, 69

OCT Optical Coherence Tomography. ii–iv, 1–3, 5, 7, 10–12, 14, 15, 20, 23, 34, 40, 45, 46, 56, 61–63, 65, 66, 69

OCTA Optical Coherence Tomography Angiography. 11

PR Precision-Recall. 32

RAM Random Access Memory. 25

ROC Receiver Operating Characteristic. 25, 32, 38, 39, 55

RPE Retinal Pigment Epithelium. 9, 12–14, 41, 52, 66, 69

SD-OCT Spectral Domain OCT. iii, 11, 45, 46

SGD Stochastic Gradient Descent. 64

SRF Subretinal Fluid. 14

SS-OCT Swept Source OCT. 11

sub-RPE Sub-Retinal Pigment Epithelium Fluid. 14

VRAM Video Random Access Memory. 25

Glossary

A-scan Amplitude scan: Unidimensional OCT signal, perpendicular to the retina.. 10

B-scan Bidimensional OCT signal providing a cross-sectional image of the retina. Often constructed through the grouping of A-scans. iii, 10–12, 26, 28, 30, 41, 42, 44–49, 63, 65

C-scan Three-dimensional OCT signal providing a volumetric image of the retina. Often constructed through the grouping of B-scans. iii, 11, 29–31, 47, 48, 65

choroid The choroid of the eye is primarily a vascular structure supplying the outer retina [57]. 7, 9, 12, 14, 27

convolution A convolution is an integral that expresses the amount of overlap of one function g as it is shifted over another function f [58]. 16, 17

druse Accumulation of extracellular, polymorphous material between the RPE and the inner collagenous zone of Bruch's membrane, that presents a yellow appearance in fundus imaging [59]. 12

exudate In the context of nAMD characterization, the accumulation of serum components in excess of the local capability of removal due to the breakdown of the blood-retinal barrier [25]. Synonym of fluid. iii, 14

fine tuning The practice of adjusting the some or all layers of a pretrained model, often after fully training the last layers. 20

fluid In the context of nAMD characterization, the accumulation of serum components in excess of the local capability of removal due to the breakdown of the blood-retinal barrier [25]. Synonym of exudate. iii, 3, 12, 14, 20, 27, 30, 32, 42, 43, 52, 56, 57, 61, 62, 65

neovascularization Invasion by vascular and associated tissues into the outer retina, subretinal space or sub-RPE space in varying combinations [25]. iii, 1, 3, 12, 13, 27, 30, 32, 34, 41, 43, 44, 56, 57, 61, 62, 65

overfitting Refers to the process of learning specific idiosyncrasies from a training set such as spurious artifacts or random noise, which results in an over-adaption to the training set and therefore in a degradation of the ability to generalize to new, unseen data [60]. 20, 32, 34, 35, 37

transfer learning The practice of using the parameters of a network trained in one data set for another task. 20

Bibliography

[1] J. Q. Li, T. Welchowski, M. Schmid, M. M. Mauschitz, F. G. Holz, and R. P. Finger, "Prevalence and incidence of age-related macular degeneration in europe: a systematic review and meta-analysis," *British Journal of Ophthalmology*, vol. 104, no. 8, pp. 1077–1084, 2020.

[2] W. L. Wong, X. Su, X. Li, C. M. G. Cheung, R. Klein, C.-Y. Cheng, and T. Y. Wong, "Global prevalence of age-related macular degeneration and disease burden projection for 2020 and 2040: a systematic review and meta-analysis," *The Lancet Global Health*, vol. 2, no. 2, pp. e106–e116, 2014.

[3] L. S. Lim, P. Mitchell, J. M. Seddon, F. G. Holz, and T. Y. Wong, "Age-related macular degeneration," *The Lancet*, vol. 379, no. 9827, pp. 1728–1738, 2012.

[4] P. Mitchell, G. Liew, B. Gopinath, and T. Y. Wong, "Age-related macular degeneration," *The Lancet*, vol. 392, no. 10153, pp. 1147–1159, 2018.

[5] P. A. Keane, P. J. Patel, S. Liakopoulos, F. M. Heussen, S. R. Sadda, and A. Tufail, "Evaluation of age-related macular degeneration with optical coherence tomography," *Survey of ophthalmology*, vol. 57, no. 5, pp. 389–414, 2012.

[6] C. S. Lee, D. M. Baughman, and A. Y. Lee, "Deep learning is effective for classifying normal versus age-related macular degeneration oct images," *Ophthalmology Retina*, vol. 1, no. 4, pp. 322–327, 2017.

[7] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE conference on computer vision and pattern recognition*. Ieee, 2009, pp. 248–255.

[8] L. Beyer, O. J. Hénaff, A. Kolesnikov, X. Zhai, and A. v. d. Oord, "Are we done with imagenet?" *arXiv preprint arXiv:2006.07159*, 2020.

[9] L. M. Biga, S. Dawson, A. Harwell, R. Hopkins, J. Kaufmann, M. LeMaster, P. Matern, K. Morrison-Graham, D. Quick, and J. Runyeon, *Anatomy & physiology.* OpenStax/Oregon State University, 2020.

[10] B. Sweet and M. Kaiser, "Depth perception, cueing, and control," in *AIAA Modeling and Simulation Technologies Conference*, 2011, p. 6424.

[11] P. Artal, "Optics of the eye and its impact in vision: a tutorial," *Advances in Optics and Photonics*, vol. 6, no. 3, pp. 340–367, 2014.

[12] M. Häggström, "Medical gallery of mikael häggström," 2014.

[13] H. Kolb. (2011) Simple anatomy of the retina. Accessed: Mar. 26, 2022. [Online]. Available: https://webvision.med.utah.edu/book/part-i-foundations/simple-anatomy-of-the-retina

[14] B. A. Wandell, *Foundations of vision.* Sinauer Associates, 1995.

[15] M. W. Hankins, S. N. Peirson, and R. G. Foster, "Melanopsin: an exciting photopigment," *Trends in neurosciences*, vol. 31, no. 1, pp. 27–36, 2008.

[16] F. H. Zaidi, J. T. Hull, S. N. Peirson, K. Wulff, D. Aeschbach, J. J. Gooley, G. C. Brainard, K. Gregory-Evans, J. F. Rizzo III, C. A. Czeisler *et al.*, "Short-wavelength light sensitivity of circadian, pupillary, and visual awareness in humans lacking an outer retina," *Current biology*, vol. 17, no. 24, pp. 2122–2128, 2007.

[17] S. L. S. Ding, S. Kumar, and P. L. Mok, "Cellular reparative mechanisms of mesenchymal stem cells for retinal diseases," *International Journal of Molecular Sciences*, vol. 18, no. 8, p. 1406, 2017.

[18] O. Strauss, "The retinal pigment epithelium in visual function," *Physiological reviews*, vol. 85, no. 3, pp. 845–881, 2005.

[19] J. Fujimoto and E. Swanson, "The development, commercialization, and impact of optical coherence tomography," *Investigative ophthalmology & visual science*, vol. 57, no. 9, pp. OCT1–OCT13, 2016.

[20] P. Cabaleiro, J. de Moura, J. Novo, P. Charlón, and M. Ortega, "Automatic identification and representation of the cornea–contact lens relationship using as-oct images," *Sensors*, vol. 19, no. 23, p. 5087, 2019.

[21] E. Zaharova and J. Sherman, "The use of sd-oct in the differential diagnosis of dots, spots and other white retinal lesions," *Eye and Brain*, vol. 3, p. 69, 2011.

[22] F. L. Ferris III, C. Wilkinson, A. Bird, U. Chakravarthy, E. Chew, K. Csaky, S. R. Sadda, B. I. for Macular Research Classification Committee *et al.*, "Clinical classification of age-related macular degeneration," *Ophthalmology*, vol. 120, no. 4, pp. 844–851, 2013.

[23] C. Metrangolo, S. Donati, M. Mazzola, L. Fontanel, W. Messina, G. D'alterio, M. Rubino, P. Radice, E. Premi, and C. Azzolini, "Oct biomarkers in neovascular age-related macular degeneration: A narrative review," *Journal of Ophthalmology*, vol. 2021, 2021.

[24] K. Amissah-Arthur, S. Panneerselvam, N. Narendran, and Y. Yang, "Optical coherence tomography changes before the development of choroidal neovascularization in second eyes of patients with bilateral wet macular degeneration," *Eye*, vol. 26, no. 3, pp. 394–399, 2012.

[25] R. F. Spaide, G. J. Jaffe, D. Sarraf, K. B. Freund, S. R. Sadda, G. Staurenghi, N. K. Waheed, U. Chakravarthy, P. J. Rosenfeld, F. G. Holz *et al.*, "Consensus nomenclature for reporting neovascular age-related macular degeneration data: consensus on neovascular age-related macular degeneration nomenclature study group," *Ophthalmology*, vol. 127, no. 5, pp. 616–636, 2020.

[26] K. T. Kim, H. Lee, J. Y. Kim, S. Lee, J. B. Chae, and D. Y. Kim, "Long-term visual/anatomic outcome in patients with fovea-involving fibrovascular pigment epithelium detachment presenting choroidal neovascularization on optical coherence tomography angiography," *Journal of Clinical Medicine*, vol. 9, no. 6, p. 1863, 2020.

[27] T. M. Mitchell, *Machine Learning*, ser. McGraw-Hill international editions - computer science series. McGraw-Hill Education, 1997. [Online]. Available: https://books.google.es/books?id=xOGAngEACAAJ

[28] A. M. Turing, "Computing machinery and intelligence," *Mind*, vol. 59, no. 236, pp. 433–460, 1950. [Online]. Available: http://www.jstor.org/stable/2251299

[29] J. Schmidhuber, "Deep learning in neural networks: An overview," *Neural networks*, vol. 61, pp. 85–117, 2015.

[30] V. Nair and G. E. Hinton, "Rectified linear units improve restricted boltzmann machines," in *Icml*, 2010.

[31] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, "Learning representations by back-propagating errors," *nature*, vol. 323, no. 6088, pp. 533–536, 1986.

[32] T. M. Mitchell, *The need for biases in learning generalizations*. Department of Computer Science, Laboratory for Computer Science Research ..., 1980.

[33] D. H. Wolpert, “The lack of a priori distinctions between learning algorithms,” *Neural computation*, vol. 8, no. 7, pp. 1341–1390, 1996.

[34] K. Fukushima and S. Miyake, “Neocognitron: A self-organizing neural network model for a mechanism of visual pattern recognition,” in *Competition and cooperation in neural nets*. Springer, 1982, pp. 267–285.

[35] Y. LeCun, B. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard, and L. D. Jackel, “Backpropagation applied to handwritten zip code recognition,” *Neural computation*, vol. 1, no. 4, pp. 541–551, 1989.

[36] D. H. Hubel and T. N. Wiesel, “Receptive fields of single neurones in the cat’s striate cortex,” *The Journal of physiology*, vol. 148, no. 3, p. 574, 1959.

[37] A. Krizhevsky, I. Sutskever, and G. E. Hinton, “Imagenet classification with deep convolutional neural networks,” *Communications of the ACM*, vol. 60, no. 6, pp. 84–90, 2017.

[38] C. C. Aggarwal *et al.*, “Neural networks and deep learning,” *Springer*, vol. 10, pp. 978–3, 2018.

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

[40] R. K. Srivastava, K. Greff, and J. Schmidhuber, “Training very deep networks,” *Advances in neural information processing systems*, vol. 28, 2015.

[41] R. Wightman, H. Touvron, and H. Jégou, “Resnet strikes back: An improved training procedure in timm,” *arXiv preprint arXiv:2110.00476*, 2021.

[42] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, “Densely connected convolutional networks,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 4700–4708.

[43] R. Wightman, “Pytorch image models,” https://github.com/rwightman/pytorch-image-models, 2019.

[44] V. Biscione and J. S. Bowers, “Convolutional neural networks are not invariant to translation, but they can learn to be,” *arXiv preprint arXiv:2110.05861*, 2021.

[45] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, “Learning deep features for discriminative localization,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 2921–2929.

[46] Y.-C. Ko, S.-Y. Wey, W.-T. Chen, Y.-F. Chang, M.-J. Chen, S.-H. Chiou, C. J.-L. Liu, and C.-Y. Lee, "Deep learning assisted detection of glaucomatous optic neuropathy and potential designs for a generalizable model," *PLoS One*, vol. 15, no. 5, p. e0233079, 2020.

[47] M. Lin, Q. Chen, and S. Yan, "Network in network," *arXiv preprint arXiv:1312.4400*, 2013.

[48] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-cam: Visual explanations from deep networks via gradient-based localization," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 618–626.

[49] "Software process models," http://users.csc.calpoly.edu/~jdalbey/308/Lectures/SoftwareProcessModels.html, accessed: 2022-08-30.

[50] D. Lachinov, P. Seeböck, J. Mai, F. Goldbach, U. Schmidt-Erfurth, and H. Bogunovic, "Projective skip-connections for segmentation along a subset of dimensions in retinal oct," in *International Conference on Medical Image Computing and Computer-Assisted Intervention.* Springer, 2021, pp. 431–441.

[51] D. Kingma and J. Ba, "Adam: A method for stochastic optimization," *International Conference on Learning Representations*, 12 2014.

[52] I. Loshchilov and F. Hutter, "Fixing weight decay regularization in adam," 2018.

[53] P. Nakkiran, G. Kaplun, Y. Bansal, T. Yang, B. Barak, and I. Sutskever, "Deep double descent: Where bigger models and more data hurt," *Journal of Statistical Mechanics: Theory and Experiment*, vol. 2021, no. 12, p. 124003, 2021.

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

[55] D. S. Kermany, M. Goldbaum, W. Cai, C. C. Valentim, H. Liang, S. L. Baxter, A. McKeown, G. Yang, X. Wu, F. Yan *et al.*, "Identifying medical diagnoses and treatable diseases by image-based deep learning," *Cell*, vol. 172, no. 5, pp. 1122–1131, 2018.

[56] S. J. Chiu, J. A. Izatt, R. V. O'Connell, K. P. Winter, C. A. Toth, and S. Farsiu, "Validated automatic segmentation of amd pathology including drusen and geographic atrophy in sd-oct images," *Investigative ophthalmology & visual science*, vol. 53, no. 1, pp. 53–61, 2012.

[57] D. L. Nickla and J. Wallman, "The multifunctional choroid," *Progress in retinal and eye research*, vol. 29, no. 2, pp. 144–168, 2010.

[58] E. W. Weisstein, “Convolution. From MathWorld—A Wolfram Web Resource,” https://mathworld.wolfram.com/Convolution.html, accessed: 27/7/2022.

[59] W. M. Al-Zamil and S. A. Yassin, “Recent developments in age-related macular degeneration: a review,” *Clinical interventions in aging*, vol. 12, p. 1313, 2017.

[60] R. O. Duda, P. E. Hart *et al.*, *Pattern classification.* John Wiley & Sons, 2006.