UNIVERSIDADE D
COIMBRA

Rita Caetano Trindade

Identification of Changes in Retinal Images of Animal Models of Alzheimer's Disease Using a Deep Learning Approach

Thesis submitted to the Faculty of Science and Technology of the University of Coimbra for the degree of Master in Biomedical Engineering with specialization in Clinical Informatics and Bioinformatics, supervised by Prof. Dr. Rui Bernardes and Prof. Dr. Pedro Serranho.

September, 2021

1 2  9 0
FACULDADE DE CIÊNCIAS E TECNOLOGIA
UNIVERSIDADE D COIMBRA

Rita Caetano Trindade

Identification of Changes in Retinal Images of Animal Models of Alzheimer's Disease Using a Deep Learning Approach

Thesis submitted to the
University of Coimbra for the degree of
Master in Biomedical Engineering

Supervisors:
Prof. Dr. Rui Bernardes
Prof. Dr. Pedro Serranho

Coimbra, 2021

Agradecimentos

Gostaria de agradecer, em primeiro lugar, aos meus orientadores, o Professor Rui Bernardes e o Professor Pedro Serranho, por terem sido sempre tão prestáveis e atenciosos. Graças ao seu profissionalismo, disponibilidade e orientação adquiri muitos conhecimentos valiosos.

Aos meus amigos da universidade, em especial à Beatriz e à Patrícia, com quem partilhei os últimos 5 anos. Deste percurso ficaram muitas vivências e histórias para contar. Obrigada por tudo!

Finalmente, um agradecimento muito especial à minha mãe, à minha irmã e à minha avó, que me deram forças para continuar, mesmo quando a minha vontade era de desistir. Aos meus primos e tios, que são sinónimo de uma lufada de ar fresco e de tardes bem passadas. E, por fim, ao meu anjo da guarda, o meu pai, que está lá em cima, orgulhoso e feliz, a olhar por mim.

Funding

This work was developed under the framework of a project (OCT4BRAIN) funded by the Portuguese Foundation for Science and Technology (FCT) through PTDC/EMD-EMD/28039/2017 and by FEDER-COMPETE through POCI-01-0145-FEDER-02803.

"If I have seen further than others, it is by standing upon the shoulders of giants."

ISAAC NEWTON

Resumo

A doença de Alzheimer é um transtorno neurodegenerativo progressivo cujo diagnóstico permanece um desafio, pois este só é possível quando já ocorreram danos neurológicos significativos. Dado que a retina e o cérebro têm a mesma origem embrionária, cada vez mais se tem vindo a utilizar a retina como uma janela para o cérebro.

Este estudo teve como objetivo identificar alterações características em imagens da retina de modelos animais da doença de Alzheimer, imagens reconstruídas a partir de dados de tomografia de coerência ótica nos estádios iniciais da doença.

Para alcançar este objetivo, foi utilizada uma abordagem de aprendizagem profunda. Em primeiro lugar, foram criadas seis redes neuronais de convolução (CNNs) para determinar se as imagens de cada camada ou conjunto de camadas agregadas da retina eram distintas entre ratos do tipo selvagem e ratos transgénicos (modelos da doença de Alzheimer). Em seguida, o método *Gradient-weighted Class Activation Mapping* (Grad-CAM) foi utilizado para avaliar em que áreas (regiões) das imagens se encontrava a informação conducente à distinção entre grupos. Adicionalmente, foi também realizada uma experiência para verificar se os olhos direito e esquerdo partilham as mesmas características. Por último, uma rede neuronal foi desenvolvida para avaliar se a combinação das classificações resultantes das seis CNNs melhoraria o desempenho na distinção entre grupos. Com base nesta última tarefa, duas técnicas (algoritmo de conexão de pesos e método de perturbação) foram utilizadas para descobrir que camadas/conjunto de camadas mais contribuíram para a classificação nos grupos de controlo e transgénico.

As CNNs provaram conseguir classificar corretamente as imagens usadas considerando apenas uma camada/conjunto de camadas, com uma exatidão entre 79,0% e 89,2% no grupo de teste, indicando assim que todas as camadas contêm informação suficiente para discriminar ratos do tipo selvagem de ratos transgénicos. Em geral, os mapas de calor sugeriram que as características essenciais estão presentes numa área

mais extensa em imagens classificadas como transgénicas do que em imagens classificadas como do tipo selvagem, e que essas áreas não estão localizadas nas mesmas regiões entre as várias camadas/conjunto de camadas da retina. Além disso, as imagens do olho direito e esquerdo não partilham as mesmas características. Por fim, a combinação dos dados provenientes das seis camadas/conjunto de camadas não melhorou o desempenho da classificação, atingindo uma exatidão de 85,4%. Nesta tarefa, a camada nuclear interna (INL) foi a que mais contribuiu, ao contrário da camada plexiforme interna (IPL) e das camadas das fibras nervosas e das células ganglionares (RNFL-GCL) que tiveram uma contribuição residual.

Palavras-chave: Doença de Alzheimer, Redes Neuronais, Tomografia por Coerência Ótica

Abstract

Alzheimer's disease (AD) is a progressive neurodegenerative disorder whose diagnosis remains a considerable challenge as it is only possible when significant neurological damage has occurred. Given that the retina and the brain have the same embryonic origin, increasing attention has focused on using the retina as a window into the brain.

This study aimed to identify the characteristic changes in computed optical coherence tomography ocular fundus images of animal models of AD in the early stages of the disease.

To achieve this goal, a deep learning approach was used. Firstly, six Convolutional Neural Networks (CNNs) were created to determine if the computed fundus images of each retinal layer/layer-aggregate were distinct between wild-type and transgenic mice. Afterwards, the Gradient-weighted Class Activation Mapping (Grad-CAM) method was applied to assess which image areas were decisive to distinguish groups. An additional experiment was also made to ascertain if the right and left eyes share the same characteristics. Lastly, a neural network was developed to assess if combining the six CNNs predictions would improve the performance. Based on this last task, two techniques (connection weights algorithm and perturbation method) were used to discover which layers/layer-aggregates had a higher contribution to classification.

The CNNs proved able to classify correctly the fundus images belonging to only one layer/layer-aggregate, with accuracies between 79.0% and 89.2% on the test set, implying that all layers have helpful information to discriminate wild-type from transgenic mice. In general, the heatmaps suggested that the meaningful characteristics in images classified as transgenic were present in a more extensive area than the wild-type, and those areas were not located in the same places between retinal layers/layer-aggregates. Moreover, the right- and left-eye images did not convey the same information. Finally, combining data from the six layers/layer-aggregates did

not improve further the classification performance, reaching an accuracy of 85.4%. The Inner Nuclear Layer (INL) was the most contributing layer in this task, and the Inner Plexiform Layer (IPL) and Retinal Nerve Fibre and Ganglion Cell Layers (RNFL-GCL) had a residual contribution.

Keywords: Alzheimer's Disease, Neural Networks, Optical Coherence Tomography

Contents

List of Figures

List of Tables

List of Abbreviations

^{18}**F-FDG** Fluorine 18 fluorodeoxyglucose.

Aβ amyloid-beta.
AD Alzheimer's Disease.
Adam Adaptive Moment Estimation.
APOE4 Apolipoprotein E 4.
APPswe Swedish Amyloid Precursor Protein.

BOLD Blood-Oxygen-level-dependent.

CAM Class Activation Mapping.
CLAHE Contrast-Limited Adaptive Histogram Equalisation.
CNN Convolutional Neural Network.
CNS Central Nervous System.
CSF Cerebrospinal Fluid.

ELM Extreme Learning Machine.

FFNN Feed-Forward Neural Network.
fMRI Functional Magnetic Resonance Imaging.

GCIPL Ganglion-Cell/Inner-Plexiform Layers.
GCL Ganglion Cell Layer.
GPR Gaussian Process Regression.
Grad-CAM Gradient-weighted Class Activation Mapping.

HIC High-Income Countries.

iCBR Coimbra Institute for Clinical and Biomedical Research.
ILSVRC ImageNet Large Scale Visual Recognition Challenge.
INL Inner Nuclear Layer.
IPL Inner Plexiform Layer.

LMIC Low- and Middle-Income Countries.

MAPT Microtubule-Associated Protein Tau.
MCI Mild Cognitive Impairment.
MD-DTI Mean Diffusivity-Diffusion Tensor Imaging.
MRI Magnetic Resonance Imaging.
MVF Mean Value Fundus.

NFT Neurofibrillary Tangles.
NINCDS-ADRDA National Institute of Neurological and Communicative Disorders and Stroke and Alzheimer's Disease and Related Disorders Association.
NN Neural Network.

OCT Optical Coherence Tomography.
ONL Outer Nuclear Layer.
OPL Outer Plexiform Layer.

PLS Partial Least Squares.
PSEN1 Presenilin 1.
p-Tau phosphorylate-Tau.
PET Positron Emission Tomography.

ReLU Rectified Linear Unit.
RMSProp Root Mean Square Propagation.
RNFL Retinal Nerve Fibre Layer.
RNFL-GCL Retinal Nerve Fibre and Ganglion Cell Layers.
ROI Region of Interest.

SGD Stochastic Gradient Descent.
sMRI Structural Magnetic Resonance Imaging.
SVM Support Vector Machine.

t-Tau total-Tau.
TR Total Retina.

VBM Voxel-based Morphometry.

1

Introduction

1.1 Context and Motivation

Alzheimer's Disease (AD) is the leading cause of dementia in the elderly, and it is characterised by a progressive decline in cognitive functions [1]. It begins with the difficulty in forming recent memories, as well as in completing simple daily tasks. Patients have mood and personality changes, and they also complain about decreased colour vision and contrast sensitivity [2]. In the late state, AD leads to total dependence of caregivers to assist in activities of daily living [3].

The current process of AD diagnosis falls short of the ideal standards. Patients are only diagnosed with AD when they are already showing signs of cognitive decline, which means irreversible damage has occurred in the brain. Therefore, it is crucial to perform an accurate diagnosis in the preclinical stage before permanent brain damage occurs. This is the optimal time window to intervene with therapies to stop or slow AD progression [4].

Up to the present, the current techniques used to support the AD diagnosis are limited by high cost and low availability (Magnetic Resonance Imaging (MRI) and Positron Emission Tomography (PET)) [5]. In addition, Cerebrospinal Fluid (CSF) biomarkers have the invasiveness drawback. Furthermore, these modalities have suboptimal sensitivity and specificity, especially in distinguishing between AD and other dementia disorders.

Studies suggest that the neurodegeneration process observed in the brain may also occur in the retina [6]. Indeed, histological evidence shows the accumulation of amyloid-beta (Aβ) peptides and tau proteins in the retina - these being the brain hallmarks of AD. Therefore, the retina can be considered a window to the brain since retinal abnormalities may give valuable insights into pathological brain features. For example, Aβ plaques were observed in the retina of mice 2.5 months earlier than

detected in the brain, suggesting that Aβ deposition in the retina precede the brain [7].

Within the scope of an ongoing project, the OCT4BRAIN project, the research group gathered Optical Coherence Tomography (OCT) data from wild-type and triple-transgenic mice model of AD. The retinal volume data was then segmented into six different ocular fundus images for individual retinal layers or layer-aggregates (Retinal Nerve Fibre and Ganglion Cell Layers (RNFL-GCL), Inner Plexiform Layer (IPL), Inner Nuclear Layer (INL), Outer Plexiform Layer (OPL), Outer Nuclear Layer (ONL), and Total Retina (TR)).

Previous studies conducted by the research group demonstrated that changes in the individual layers/layer-aggregates were present. Moreover, these differences were sufficient to distinguish both groups in the early stages of disease using Support Vector Machine (SVM) with radial basis function kernel [8]. Besides, texture analysis was applied to the computed fundus images and a substantial distinction between groups at one- and two-months-old was found across all layers/layer-aggregates [9].

The study herein intends to embrace a deep learning approach to confirm previous findings stating that the computed fundus images of each particular layer/layer-aggregate are distinct between groups. In addition, it aims to determine which areas in the images convey relevant information for each layer/layer-aggregate. Furthermore, it expects to ascertain if combining the information from all layers/layer-aggregates results in higher classification performance and which layers had higher importance in this classification task.

Using OCT to aid the AD diagnosis and to monitor AD progression would be very beneficial. In addition to being non-invasive, it is fast, inexpensive, and has a great potential to provide biomarkers of alterations at the early stages of AD.

1.2 Goals and Contributions

As aforementioned, the main goal of this study is to identify the characteristic changes in retinal images of animal models in the early stages of AD. These findings can then be sought in humans to determine whether there is a parallel between retinal changes in the animal models and humans diagnosed with AD. In that case, this study will subsequently contribute to an earlier diagnosis, hopefully, before irreversible brain damage occurs.

There are six objectives to fulfil in this thesis:

- exploit the dataset and develop a preprocessing phase to reduce variability between images with no relevance for the task at hand;
- build six Convolutional Neural Networks (CNNs) to assess their ability in distinguishing between wild-type and transgenic mice using images from a particular layer;
- gain insights into the visual changes in the images between both groups by exploring how the CNNs made the decisions using a back-propagation method;
- investigate whether both right and left eyes share the same information - this will be achieved by developing CNN models that learn with images from one eye and are tested on images from the other eye;
- build a Feed-Forward Neural Network (FFNN) to assess whether joining the predictions of all six CNNs improves classification performance, and;
- determine which retinal layers or layer-aggregates have contributed the most to the classification by applying algorithms that assess the input features importance.

1.3 Document Structure

This document is structured in seven chapters.

Chapter 2 provides the basic knowledge about AD. It begins with a brief historical context. Then, the epidemiology in the world and Portugal is presented. Moreover, the pathophysiology and the current diagnosis are described.

Chapter 3 reviews the existing literature about the current techniques in study to aid the AD diagnosis. These include CSF biomarkers and three imaging techniques: PET, MRI, and OCT.

Chapter 4 describes deep learning models (FFNN and CNN), comprising their learning process, regularisation methods, transfer learning and explainability methods.

Chapter 5 presents a description of the dataset, describing how the data was collected and preprocessed.

In Chapter 6, the obtained results are reported and discussed.

Finally, Chapter 7 concludes the thesis and presents some possibilities for future work.

2

Alzheimer's Disease

2.1 Brief Historical Context

The first case of Alzheimer's disease traces back to the beginning of the 20th century, when Alois Alzheimer, a German psychiatrist and neuropathologist, presented his patient's clinical case at the 37th Meeting of the Southwest German Psychiatrists in Tübingen in November 1906 [10].

Alois Alzheimer was born in 1864 in Markbreit am Main, Germany, and showed enormous enthusiasm for Natural Sciences from an early age [11]. He graduated in medicine, and while working as a resident at the *Frankfurt Institution of the Mentally Ill and Epileptic*, Ms Auguste Deter was admitted to the institution on 25th November 1901 [12].

Auguste Deter was a 51-year-old patient from Frankfurt, and when observed by Alzheimer, she revealed decreased memory and comprehension, disorientation, progressive aphasia, and significant psychosocial impairment [12]. Alzheimer's case presentation stated that Ms Deter symptoms were so distinct from all the other described diseases that it was impossible to label her condition based on known mental illnesses.

Alzheimer moved to Munich in 1903 by invitation from a German psychiatrist known today as the founder of modern scientific psychiatry, psychopharmacology, and psychiatric genetic, Emil Kraepelin [10]. Indeed, it was Kraepelin who named this disease after Alois Alzheimer.

Afar, Alzheimer kept following Auguste's disease state. When she died in 1906, her brain was sent to Munich so Alzheimer could perform an autopsy to discover the histological changes responsible for her clinical symptoms [11]. This examination showed a massive loss of cells throughout the brain, deposits of an unidentified substance in the form of plaques within the cerebral cortex, and peculiar thick and

strongly staining fibrils in the remaining neurons [13].

Nowadays, it is known these are, in fact, the hallmarks of AD: loss of neurons, and accumulation of Neurofibrillary Tangles (NFTs) and Aβ plaques [13].

Later on, between 1907 and 1908, Alzheimer took care of another three patients experiencing similar symptoms to Ms Deter. Their brains' examination revealed that they shared the same histopathological changes: deposition of plaques and NFTs within the cerebral cortex [11]. The report Alzheimer wrote about the second patient, Johann F., was the one that drew the scientists' attention, in contrast to his former presentation in the 37th Meeting in Tübingen, which was even considered "unsuitable for publication in the meeting proceedings by the organisers" [11].

2.2 Epidemiology

The World Alzheimer Report of 2015 estimated that 46.8 million people worldwide lived with dementia that year, and these numbers are expected to increase to 74.7 million in 2030 and 131.5 million in 2050 [14]. These figures are approximately 13% higher than the estimations made in the 2009 report [15] due to the increasing prevalence of dementia among those aged over 60 in Asia (from 3.9% in 2010 to 4.7% in 2015) and Africa (from 2.6% to 4.6%). On the other hand, it was noticed a slight decrease in Europe and America (6.2% to 5.9% and 6.5% to 6.4%, respectively) [14, 15]. Although these figures account for all types of dementia, AD is estimated to be responsible for 70% of the cases [16].

In what concerns the incidence, studies suggest that, since there is an exponential increase with age, dementia is an imminent outcome of ageing [17]. The incidence duplicates every 6.3 years among those aged 60 and over, increasing from 3.9 per 1000 person-years to 104.8 per 1000 person-years at 90+ years of age [14].

Moreover, while High-Income Countries (HIC) appear to have greater incidence, doubling in age every 5.8 years from 3.5/1000 per year to 124.9/1000 per year, in Low- and Middle-Income Countries (LMIC), there is a double increase every 8.6 years from 5.2/1000 per year to 58.0/1000 per year. Nevertheless, it was determined that the overall incidence of dementia in HIC is only 10% higher than in LMIC, which is not statistically significant [14].

In Portugal, 160 287 people had dementia in 2013, corresponding to 5.9% of the population residing in the country and aged over 60 years. These values suggest an increase in the prevalence over the past decade because, in 1991, it was ascertained

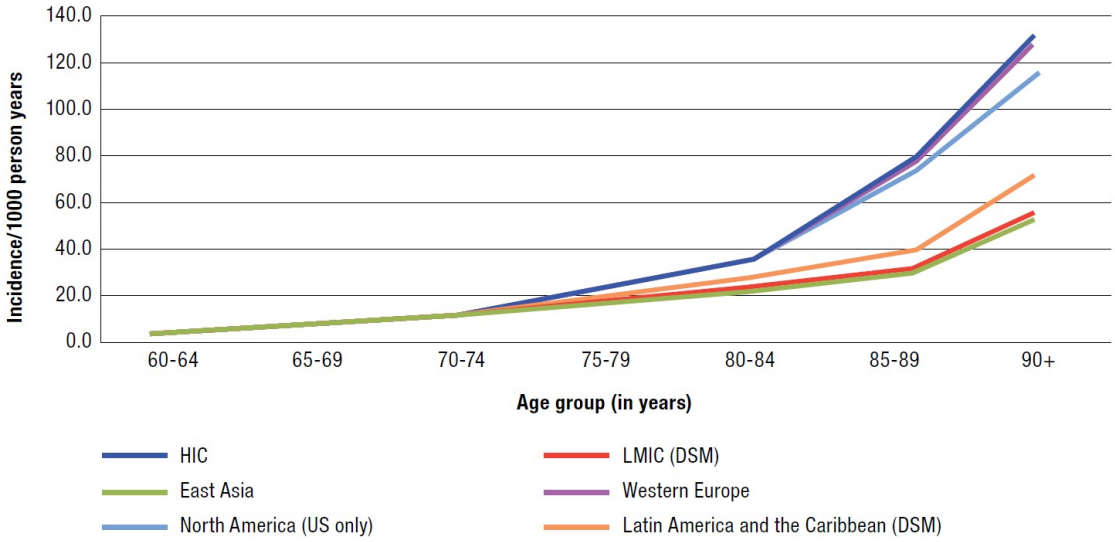


Figure 2.1: Estimated age-specific annual incidence of dementia. Image from "World Alzheimer Report 2015 The Global Impact of Dementia An analysis of prevalence, incidence, cost and trends" [14].

at a 4.6% rate. Because Portugal is facing increasing demographic ageing, it is expected that the number of dementia cases will grow in the coming years [18].

2.3 Pathophysiology

2.3.1 Brain

Pathologically, AD is characterised by extracellular deposition of fibrils of Aβ in diffuse and neuritic plaques [16]. These are consistently distributed within the cerebral cortex. It is thought that amyloid plaques contribute to the disruption of cellular activities and communication in the brain resulting in neuroinflammation and neuronal cell death [7].

The second hallmark of AD is the intracellular aggregation of hyperphosphorylated tau into NFTs [19]. NFTs, responsible for reducing the cytoskeleton integrity, leading to neuronal dysfunction [20], are mainly present in the medial and lateral temporal lobes and, sparingly, in the parietal, occipital, and frontal ones [21].

The third pathological feature of AD is the progressive decline in dendrites, synapses, and neurons, that is, neurodegeneration, a consequence of the abnormal Aβ and tau accumulation in the brain [22].

2.3.2 Retina

The retina is an integral part of the Central Nervous System (CNS) and shares an embryonic origin with the brain [23, 24]. Consequently, histological and clinical research suggests that the brain's neurodegenerative process also affects the retina [25]. Despite being smaller or having different shapes, Aβ plaques and NFTs were identified in the retina of rodents and humans diagnosed with AD [4].

Several studies reported pathological changes in the retina, such as thinning and degeneration of the macular layers, neuroinflammation, retinal ganglion cell death and axonal loss [2, 4, 6]. It was observed a considerable thickness reduction of the Retinal Nerve Fibre Layer (RNFL), the Ganglion Cell Layer (GCL) and the OPL, compared to healthy subjects [25]. By contrast, the ONL revealed a significant thickening [25]. These findings may result directly from Aβ and tau accumulation and, secondarily, by the visual cortex degeneration [25].

These pathological changes were detected in the early stages of this disease, even before causing damage in the hippocampal area, associated with memory formation [26]. However, the mentioned retinal abnormalities also occur in different neurodegenerative diseases, such as glaucoma, ocular hypertension, multiple sclerosis, and Parkinson's [25].

2.4 Diagnosis

The Alzheimer's disease diagnosis has been facing frequent revisions for the last decades since its breakthrough in 1906 [1].

Nowadays, doctors and researchers commonly use the National Institute of Neurological and Communicative Disorders and Stroke and Alzheimer's Disease and Related Disorders Association (NINCDS-ADRDA) criteria proposed in 1984 [27]. It categorises AD according to three labels: possible, probable, or definite, by linking the clinical examination to the neuropathological patterns [16].

It is required that the patients meet the conditions for probable AD and show supporting histopathologic evidence via autopsy or biopsy to classify AD as definite [27]. When the typical symptoms are fulfilled, and no other disorder might be the cause, it is characterised as probable [27]. Finally, it is diagnosed as possible when the cognitive decline is unusual or no other conditions contribute to dementia [3].

Alzheimer's disease is only diagnosed when the patient exhibits significant cog-

nitive signs of Alzheimer's dementia, which affects the therapy success since it has been shown to be more effective in earlier stages of the disease [28]. Besides, the criteria accuracy is reasonably low, with a sensitivity of 80% and specificity of 70% [28].

With the substantial progress in biochemical and genetic understanding of AD, neuropsychological assessment, and brain imaging, the NINCDS-ADRDA criteria were recently updated to include imaging biomarkers, such as PET, Structural Magnetic Resonance Imaging (sMRI), Functional Magnetic Resonance Imaging (fMRI), and CSF biomarkers [3, 28]. They are vital to exclude other causes of dementia such as brain tumour or subdural haematoma [28], therefore enhancing the diagnosis specificity.

Moreover, another diagnostic has been included: Mild Cognitive Impairment (MCI). It represents a transitional state between normal cognition and dementia [16]. Although MCI patients are not demented yet, they show a higher risk of progression to dementia [21].

3

Alzheimer's Disease State of the Art Diagnosing Tools

The characteristic pathophysiological abnormalities of Alzheimer's disease, such as Aβ plaques or tau-related neurodegeneration, precede the clinical symptoms by many years [29], being detectable by several biomarkers[1].

Different biomarkers are currently being sought, as they are precious in Alzheimer's disease diagnosis, progression, and treatment. Firstly, AD biomarkers establish a more reliable and accurate diagnosis since they are independent of patients' responses. They may facilitate an early diagnosis, which is of the utmost importance to accomplish more promising results in AD therapies. They may also allow identifying patients who are already developing the pathologic hallmarks of the disease but do not yet show any signs of dementia. Finally, they help following the disease progression and the treatment effectiveness, and improve the current understandings of AD neurophysiology [31].

Below, AD biomarkers widely used in clinical trials to support the updated diagnostic criteria will be presented. These belong to two categories: body fluids (CSF biomarkers) and imaging techniques (PET, MRI, and OCT).

3.1 Cerebrospinal Fluid Biomarkers

CSF is a clear liquid that surrounds the brain and the spinal cord. Among other functions, it provides mechanical protection against shock – it is a shock absorber – and transports active metabolic substances and waste products released by the brain [32]. The tau proteins - total-Tau (t-Tau) and phosphorylate-Tau

[1] A biomarker is a "characteristic that is objectively measured and evaluated as an indicator of normal biological processes, pathogenic processes, or pharmacologic responses to a therapeutic intervention" [30].

(p-Tau) - and Aβ peptides (Aβ_{42}) levels in CSF give valuable insights into the metabolic processes occurring within the brain. Thus, comparing CSF reference and patients' values may contribute to a more flawless AD diagnosis at the early stages. However, the correlation between CSF biomarkers and concentrations of Aβ plaques and neurofibrillary tangles in the brain remains unclear [16].

A CSF sample is collected through a procedure called "lumbar puncture". This is an invasive test with some associated risks: discomfort during and after the procedure, headache, allergic reaction to the anaesthetic, and infection at the puncture area [33].

The analysis of CSF samples belonging to healthy individuals and demented patients led to the conclusion that the latter have lower levels of Aβ_{1-42} and higher levels of t-Tau and p-Tau [16]. It was pointed out that the quantity of t-Tau in CSF tends to increase with AD progression [1]. Moreover, studies suggest that Aβ_{42} CSF levels become abnormal earlier than detected on amyloid PET scans and before neurodegeneration occurs in the brain [34].

Regarding the CSF diagnostic efficiency, it was determined that AD discrimination from healthy elderly individuals is highly accurate [35]. Additionally, essential findings show that Parkinson's disease and depression present normal CSF levels and p-Tau aids to distinguish AD from frontotemporal dementia and Lewy Body dementia [35]. Furthermore, the combination of the three biomarkers, t-Tau, p-Tau, and Aβ_{42}, seems to improve further the capability of prediction of whether an MCI diagnosis will progress to AD or other distinct conditions [35].

3.2 Positron Emission Tomography

PET is a non-invasive functional imaging technique that analyses the distribution of a radiotracer in the brain to assess the molecular abnormalities in enzyme activity, receptor distribution, and brain metabolism. There are three different types of PET scans that have been extensively used for *in vivo* imaging of AD pathology: radioisotopes that trace the path of Aβ within the brain, isotopes that bind to tau-proteins, and glucose metabolic tracers.

3.2.1 Amyloid PET

C-11 Pittsburgh compound B, F-18 florbetapir, and F-18 florbetaben tracers bind preferentially to Aβ peptides. Therefore, the higher the Aβ deposition, the

greater the activity observed in the brain cortex [36].

A negative amyloid PET scan implies that AD is quite unlikely. Since Aβ plaques in the brain are hallmarks of AD, these tracers are very useful in excluding AD diagnosis. However, it cannot be used as the only tool to confirm AD because it overlaps with healthy individuals of the same age. Furthermore, it cannot differentiate between different symptomatic stages [37].

3.2.2 Tau PET

Developing tau-specific tracers has been particularly challenging since it is more difficult to access tau proteins than Aβ because the former aggregates intra- and extra-cellularly. The tracers THK5317, THK5351, AV-1451, and PBB3 have revealed promising tools to monitor tau-protein proliferation [37].

Concerning AD diagnosis, tau PET might not be as discriminating as needed because tau deposition is also observed in other neurodegenerative diseases involving the aggregation of tau proteins into neurofibrillary tangles (tauopathies), such as chronic traumatic encephalography and progressive supranuclear palsy [36]. Nonetheless, it might be crucial to evaluate the treatments' efficacy in reducing the tau burden in the brain.

3.2.3 Fluorine 18 fluorodeoxyglucose (^{18}F-FDG) PET

^{18}F-FDG PET is used to analyse brain activity, vascular deficits, and blood-brain impairments by measuring the brain's glucose consumption. It was observed that individuals diagnosed with MCI or AD have a reduced glucose metabolism rate in the temporoparietal regions. [34]

Minoshima et al. (2001) [38] and Albin et al. (1996) [39] suggested that ^{18}F-FDG-PET markers can discriminate between AD and dementia with Lewy Bodies. Moreover, Santens et al. (2001) [40] made an ^{18}F-FDG-PET visual analysis of three different brain regions, confirming asymmetric patterns between AD and frontotemporal dementia, suggesting, therefore, that it is possible to distinguish between these two diseases based on ^{18}F-FDG-PET. Indeed, the US Centre for Medicare Service approved this imaging test to differentiate the two conditions (AD and frontotemporal dementia), making this imaging technique the first to make a positive diagnosis of dementia [41].

3.2.4 Final remarks about PET

Overall, PET biomarkers are a valuable asset in the diagnosis and progression analyses of AD. Since they assess the brain's metabolic functions and biochemical processes, PET scans effectively detect the disease in its early stages or before the first clinical symptoms.

On the downside, PET imaging is quite expensive. An amyloid PET scan costs approximately 4000 US dollars [34], which is not affordable for all countries. Moreover, even though radioactive isotopes have short half-lives, exposure to radioactivity is inevitable and, for that reason, it is not recommended in pregnant patients [42].

3.3 Magnetic Resonance Imaging

3.3.1 Structural Magnetic Resonance Imaging

MRI is a non-invasive medical imaging tool that produces three-dimensional images. This technology is beneficial for the brain's examination, for example, the shape, the volume of the total brain and its different regions, and the structural integrity of the grey and white matter.

The most common findings in MRI of AD patients are medial temporal lobe atrophy, including the amygdala, hippocampus, parahippocampal gyrus, entorhinal cortex and ventricular hypertrophy, and total brain volume reduction [6].

A significant part of earlier studies was concentrated on volumetric approaches, such as Voxel-based Morphometry (VBM), an automatic volumetric method for comparing regional grey matter concentration between two groups of subjects [43]. Zhang et al. (2019) [44] developed an Extreme Learning Machine (ELM) model to distinguish AD patients from healthy controls. To build the ELM diagnostic model, the authors used VBM obtained from the hippocampal region in MRI (three-dimensional features), patient information, and texture parameters (two-dimensional features). Then, this ELM model was compared with SVM, Gaussian Process Regression (GPR), and Partial Least Squares (PLS) models, showing to be effective in distinguishing AD patients from healthy subjects.

As an alternative to volumetric methods, Khvostikov et al. (2018) [43] developed the design of a multimodal 3D Inception-based CNN with hippocampal sMRI and Mean Diffusivity-Diffusion Tensor Imaging (MD-DTI) data for the diagnosis of AD. The comparison with the conventional AlexNet-based network revealed a fewer

number of weights and a higher performance level.

Predicting which patients diagnosed with MCI will progress to AD is a challenging task. Along this line, Li et al. (2019) [45] developed a deep learning model of CNNs with residual connections using hippocampal MRI data to estimate the risk of progression of MCI patients' to AD. This risk was then combined with clinical information, such as age, sex, education, and Apolipoprotein E 4 (APOE4), a gene associated with an increased risk for developing AD, to build a second prognostic model that provides a cost-effective and accurate approach for prognosis and eases the enrolment of patients in clinical trials likely to progress to AD.

Rebbah et al. (2018) [46] performed a study on the conversion of MCI to AD based on regions of interest (ROIs) using a combination of two sMRI biomarkers: cortical thickness and cortical curvature measures. The histogram analysis of the sMRI biomarkers resulted in eight parameters that went through a feature selection process. Afterwards, the classification was performed through SVM using the optimal feature subset.

Feng et al. (2021) [47] proposed an ROI-based contourlet sub-band energy feature to represent sMRI images in the frequency domain for AD classification since spatial analysis methods weaken the discrimination ability of spatial features. Thus, each preprocessed sMRI image was segmented into 90 ROIs from which energy sub-bands were obtained. Finally, an SVM classifier used the concatenation of the sub-bands energy feature vectors of the 90 ROIs to classify subjects. This study demonstrated that the ROIs' energy information could detect differences between AD and healthy controls.

3.3.2 Functional Magnetic Resonance Imaging

While sMRI is used to analyse anatomical structures, fMRI is a helpful tool for studying brain activity [16]. Commonly, fMRI relies on the Blood-Oxygen-level-dependent (BOLD) method, which measures changes in blood flow by analysing variations in the concentration of deoxyhemoglobin [1]. Active neurons consume more energy, released by the blood in the form of oxygen and sugar, than inactive neurons. Therefore, active brain regions have an increased blood flow [48].

Dickerson et al. (2005) [49] used fMRI to investigate differences between MCI, AD, and controls in hippocampal and entorhinal activation during learning. In comparison to healthy ageing, AD patients showed a reduced BOLD signal in hippocampal and entorhinal regions. The authors also suggested an increased medial

temporal lobe activation in an early phase of AD, followed by a subsequent decrease as the condition progresses.

Similar findings were described by Celone et al. (2006) [43]. Less impaired MCI subjects showed hyperactivation in the hippocampus compared to healthy controls, while more impaired MCI revealed substantial hypoactivation.

Notwithstanding these conclusions, the BOLD signal is an indirect approach to measuring brain activity [48]. It depends on physiological, anatomical, and imaging factors, limiting its effectiveness as a differential diagnosis because of the BOLD signal variation among individuals [1].

Because within the neuroimaging field the number of features is frequently numerous, Bi et al. (2018) [50] proposed a new random SVM cluster to analyse fMRI data and distinguish between AD and healthy controls. This method combines several SVMs into a random SVM cluster. Each SVM is set up with random samples and features, leading to dimensionality reduction and increased generalisation performance.

Functional MRI generates 4D data containing both spatial and time-varying information of the brain. However, classification models usually transform the data into 2D or 3D images, which might cause the neglect of crucial information. Thus, Li et al. (2020) [51] presented a 4D deep learning model, named C3d-LSTM, for AD detection. This model can deal with 4D fMRI data directly; therefore, no spatial or time-varying information is lost. Results showed that spatial and temporal information preserved in 4D fMRI data are, in fact, significant for AD discrimination, leading to an increased classifier's performance when compared to a 2D or 3D fMRI dataset.

Wang et al. (2018) [52] analysed brain network connectivity patterns. They presented a classification method for AD, MCI, and healthy subjects, robust to a limited number of fMRI data samples. Firstly, they selected the regions of interest – the hippocampus and the isthmus of the cingulate cortex – and calculated the Pearson coefficients for every ROI pair to form a feature vector for each subject. To reduce the noise effect caused by the limited dataset, a linear discriminant analysis (LDA) approach was proposed. It projects feature vectors onto a one-dimensional axis to maximise differences between AD, MCI, and healthy subjects. Finally, the classification task was performed by applying a decision tree based multi-class AdaBoost classifier. The authors concluded that the hippocampus and the isthmus of the cingulate cortex are closely related to the development of AD and MCI, and that

the regularisation methods and the AdaBoost classifier can significantly improve the classification performance.

3.4 Optical Coherence Tomography

OCT is a compelling imaging technique widely used in medicine, especially in ocular imaging. This technique is used to gather the data considered in this work. For this reason, the physical principle behind time- and frequency-domain OCT will be explained in this section.

An OCT system allows obtaining a series of parallel 2D cross-section representations of the eye fundus based on the reflectivity of the different retinal layers [53]. This technology is now considered a clinical standard in ophthalmology and crucial for diagnosing and monitoring several retinal diseases [54].

Current OCT technologies perform an optical *biopsy in vivo* and *in situ*, with an axial resolution ranging from 1 to 15 μm, one or two orders of magnitude higher than the standard ultrasound imaging [54]. These are evident assets when a standard excisional biopsy is delicate or impossible, as in the human retina, nervous tissues, or arteries [54]. Moreover, OCT is a non-contact method and a helpful tool for the guidance of interventional procedures since it enables the visualisation of the structural morphology beneath the surface [54].

The physical principle behind OCT is similar to that of ultrasound, except that ultrasound makes use of a mechanical wave while OCT uses light. When a beam of light or sound is directed onto a sample, it is backscattered or back-reflected differently depending on the optical or acoustic properties of the sample. Therefore, in ultrasound, the structure of the sample is estimated by measuring the time that sound takes to return from different depths [54]. However, in OCT, there is a major difference: as light travels much faster, a much higher time-resolution would be required, on the order of femtoseconds (10^{-15} s), so that the backscattering optical path could be detected with an axial resolution of 1-15 μm [53]. No current hardware is available to measure such a short time. Hence, the principle of interferometry was used in OCT to overcome this obstacle [53].

The Michelson interferometer splits the light emitted from a broadband source, with a wide range of optical wavelengths, into two identical light beams. Each beam travels a different path: the reference and the sample arm. The reference beam (light beam directed to the reference arm) travels a known distance to a

reference mirror and is backscattered. In contrast, the sample beam (light beam directed to the sample arm) travels to the sample to be partially backscattered. Both backscattered beams are recombined and the phase difference between them creates an interference pattern possible to be measured by a detector. There is a substantial interference signal when the backscattered light from the sample and reference arms have travelled the same optical distance. The reference mirror can be translated longitudinally, allowing the formation of the sample's reflectivity along its optical axis. This is the principle behind Time-Domain OCT (TD-OCT) technology, whose scheme is illustrated in Figure 3.1.

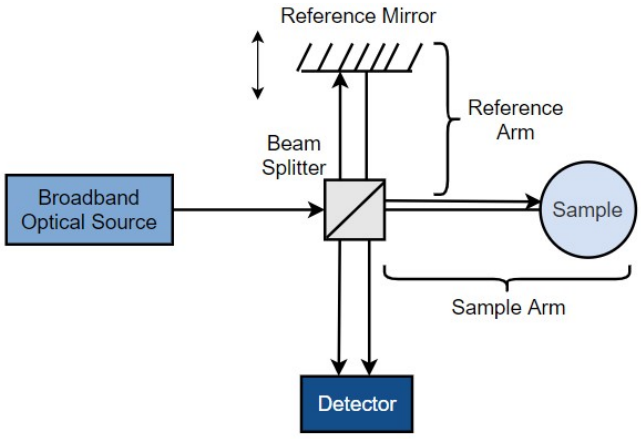


Figure 3.1: Scheme of time-domain OCT system. Image adapted from "Optical coherence tomography: A concept review" [53].

The technology of the OCT model used in the present dissertation is not of the time-domain type. Instead, it is based on the Spectral/Fourier domain technology (SD-OCT), which allows the depth information of the sample to be obtained by analysing the spectrum of the interferometer's output light by a spectrometer without the need of moving the reference mirror. Therefore, all depth information within a single A-scan is performed simultaneously, allowing for a much faster acquisition of B-scans, hence reducing the chance of saccades [55].

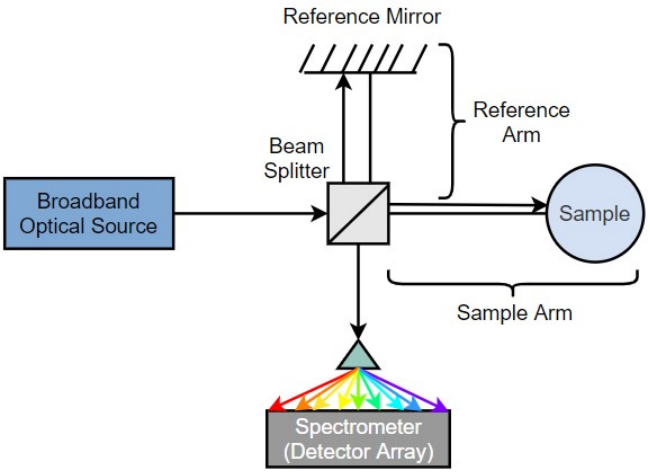


Figure 3.2: Scheme of frequency-domain OCT system. Image adapted from "Optical coherence tomography: A concept review" [53].

Several studies have focused on assessing whether data provided by OCT allows establishing biomarkers for the diagnosis and follow-up of AD. However, using ocular fundus images generated from the 3D OCT data of mice model of AD is a pioneer study rendering it impossible to compare it with other studies, as these are either based on B-scans or thickness measurements.

Güneș et al. (2015) [56] compared the RNFL thickness of patients in an early stage of AD with age-, sex- and education-matched healthy subjects. These results demonstrated a significant thinning of the RNFL on average and in all quadrants (inferior, superior, nasal, and temporal) of AD patients compared to healthy controls.

Parisi et al. (2001) [57] reported similar findings. Compared to healthy controls, AD patients showed a reduction not only in the overall RNFL thickness but also in each quadrant. Moreover, the RNFL thickness reduction in AD patients was correlated with functional retinal impairment.

Marziani et al. (2013) [58] also concluded that the RNFL and the aggregate RNFL-GCL thickness was significantly reduced in the nine examined macular fields (central, inner superior, inner temporal, inner inferior, inner nasal, outer superior, outer temporal, outer inferior, and outer nasal) of AD patients in comparison to healthy subjects.

On the other hand, Lad et al. (2018) [5] examined the aggregate Ganglion-Cell/Inner-Plexiform Layers (GCIPL) and the RNFL thickness on the macula and optic nerve of 15 MCI patients, 15 mild-moderate AD patients and 18 cognitively

normal subjects and found no statistical differences between groups.

Furthermore, Pillai et al. (2016) [59] performed a study in which they measured the RNFL thickness, the GCL thickness and macular volume among AD patients, amnestic MCI patients, non-AD dementia, Parkinson's disease patients, and healthy controls. This retinal thickness proved unable to distinguish AD from the remaining groups since no statistical differences were found. Thus, no correlation was observed between RNFL reduction and severity of cognitive impairment.

Ascaso et al. (2014) [60] performed a cross-sectional study using OCT. They measured the RNFL macular thickness and volume of patients diagnosed with AD, patients diagnosed with MCI, and healthy controls. Although an age-related thinning of RNFL in healthy subjects has been reported, AD patients showed the most significant reduction in RNFL thickness, followed by MCI patients. The analysis of the macular measurements revealed that MCI patients had the highest macular volume (considering the total retinal thickness) compared to healthy controls and AD patients, and the controls' macular volume was also higher than in AD patients.

Finally, Berisha et al. (2007) [61] evaluated the retinal hemodynamic parameters in nine patients with mild to moderate probable AD and eight age-matched controls. Patients diagnosed with AD revealed narrowed veins and a decreased venous blood flow relative to controls. They also found RNFL loss patterns in patients early diagnosed with AD and a significant RNFL thinning in the superior quadrant. Still, the inferior, temporal, and nasal quadrants showed no significant differences.

4

Deep Learning Models

This chapter describes the methods used in this study to develop the proposed deep learning approach to identify changes in retinal images in the early stages of AD.

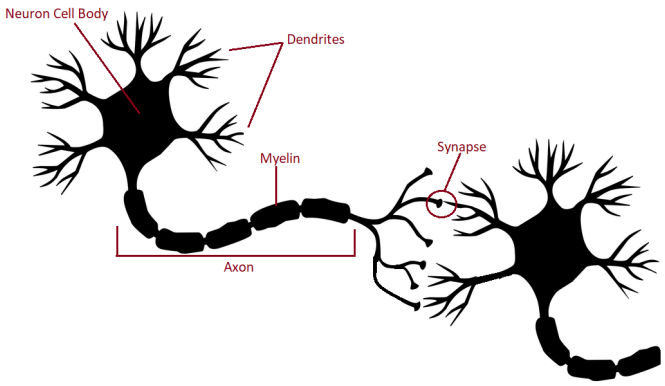
4.1 Feed-forward Neural Networks

FFNNs are inspired by the architecture of the human brain, in which electrical signals are transmitted from one neuron to the next one across a synaptic gap via action potentials (the cell either fires or it does not) and chemical neurotransmitters.

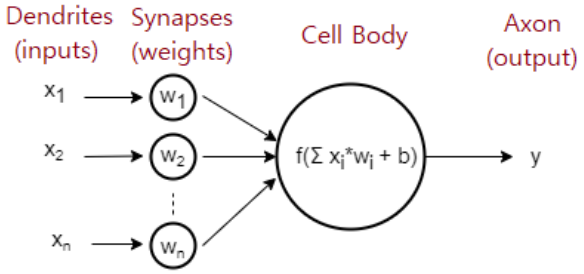
In a classification task, the main idea of a feed-forward network is to approximate a function f^* that maps an input x to a category y. The mapping $y = f(x,\theta)$ is made by learning the parameter values θ that result in the best function approximation [62].

FFNNs are composed of multiple layers. Each layer is organised in neurons that are interconnected with the neurons of the previous and following layers.

The first layer is called the input layer and is responsible for receiving the inputs that will be transmitted to the neurons of the hidden layers. These lie in between the input and output layers. The output layer receives the outputs from the previous layer and produces the final result. During training, the Neural Network (NN) decides how to make the best use of these layers to generate a function f that produces the closest value to y.



(a) Biological Neuron.



(b) Artificial Neuron.

Figure 4.1: Similarity between biological and artificial neurons.

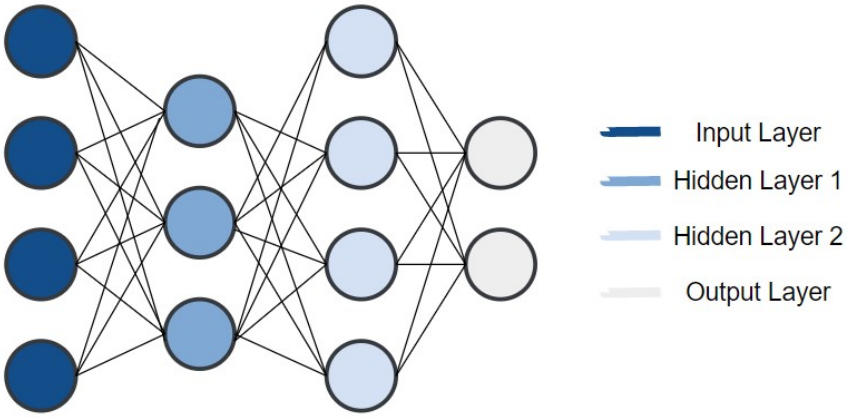


Figure 4.2: Illustration of an Artificial Neural Network architecture with multiple hidden layers.

Each connection has an associated weight that determines the influence of the neuron. Each neuron value is determined by summing the weighted input signals and, subsequently, adding a bias ($z = w^T x + b$). Then, the resulting value is fed into an activation function f and the output $a = f(z)$ determines whether the

neuron should or should not be activated. One should note that the bias shifts the activation function by adding a constant value, allowing for a better generalisation of the neural network. Both the weights and biases are learned during training.

4.1.1 Hidden Layers

Defining the hidden layers and their hidden units is an essential part of the model design process, although it does not have many guiding theoretical principles [62].

As aforementioned, each hidden neuron receives a vector of inputs x (the previous layer's outputs) and performs the transformation $z = w^T x + b$.

Rectified Linear Unit (ReLU) is the typical activation function recommended within most FFNNs [62]. ReLU is defined by the function $r(z) = max\{0,z\}$ represented in figure 4.3.

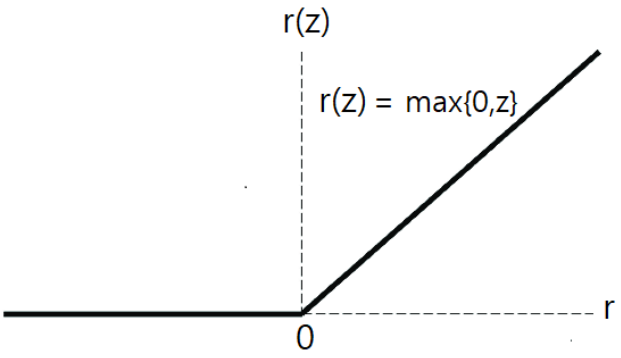


Figure 4.3: ReLU activation function.

4.1.2 Output Layer

Similarly to the hidden units, the output neuron computes the linear function $z = w^T h + b$, where h is the vector of inputs that culminated from the last hidden layer, and b is the bias. Next, an activation function is used to transform z.

For classification problems with two classes, the sigmoid function is a suitable choice. It is a non-linear activation function that bounds the output into the interval (0,1). Its mathematical expression is denoted below:

$$s(x) = \frac{1}{1 + e^{-x}} \tag{4.1}$$

Figure 4.4 shows the sigmoid activation function.

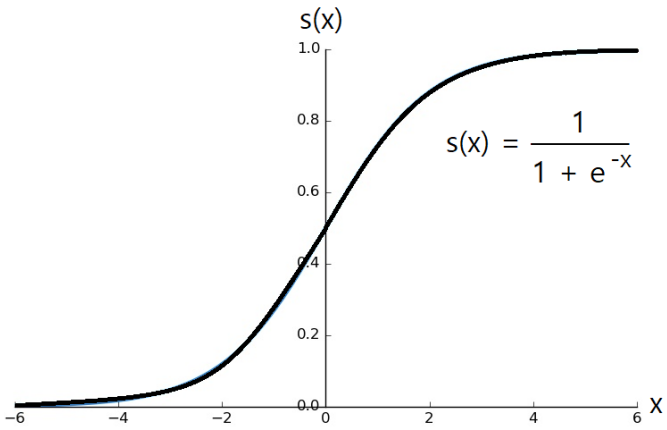


Figure 4.4: Sigmoid activation function.

4.2 Convolutional Neural Networks

CNNs are a specialised kind of neural network inspired by the organisation of the visual cortex in the animal world, where neurons are only activated by watching lines at specific orientations. CNNs are designed to adaptatively identify spatial hierarchies of features, from simpler (e.g., colours and edges) to more complex patterns (e.g., faces, objects, and shapes). These neural networks can automatically extract representations not easily identified in image data and, for this reason, they are known as feature extractors.

The convolutional network architecture is typically built with three basic building blocks: Convolution Layer, Pooling Layer and Fully-Connected Layer. The convolution layer consists of filters, or kernels, that are moved across the entire input image to create activation maps, that is, learnable feature maps that provide a helpful understanding of the internal representations of the input. These filters work as feature detectors, and this operation is known as convolution.

Convolution is a specialised linear operation that performs the sum of element-wise multiplication (Hadamard product) between the kernel and a specific location of the input matrix (image/image channel). Similarly to the traditional feed-forward network, an activation function is applied to every value of the feature map introducing the mandatory non-linearity into the model.

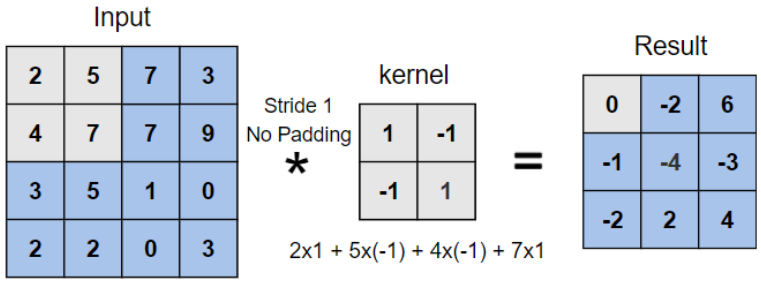


Figure 4.5: Example of a convolution operation[1].

Conversely to traditional neural layers, where each input unit interacts with each output unit, convolution layers have sparse interactions. This is accomplished by using kernels smaller than the input image. Since sparse interactions limit the number of connections for each input, computing the output requires few operations, which improves the statistical efficiency without undermining its performance [62].

In general, during the training phase, weights are initialised with random values and updated until reaching the best accuracy possible. Therefore, unlike traditional neural networks, where each element of the weighted matrix is used precisely once when computing the layer's output, the kernel weights are used at every input position. One should also note that filter weights remain fixed across the same convolution layer - parameter sharing. Each filter identifies only one kind of feature; hence, several filters should be applied in parallel to learn a wider variety of patterns. Parameter sharing means that it is necessary to learn only one set of weights per pattern for each layer, henceforth reducing the number of parameters to be learned.

A convolution layer is usually followed by a pooling layer, aiming to reduce the spatial dimension by replacing the output at specific locations by combining the nearby values into a single element. A popular pooling operation is the average pooling, which calculates the average value within a selected region of the feature map. This method helps to reduce the number of parameters to be learned and, therefore, reduces the complexity of the network leading to a more robust model to small changes (controlling overfitting) [63].

[1]Stride is the number of pixels the filter moves between positions along the input matrix and padding is adding zeros to the borders of the input to preserve and control the size of the output.

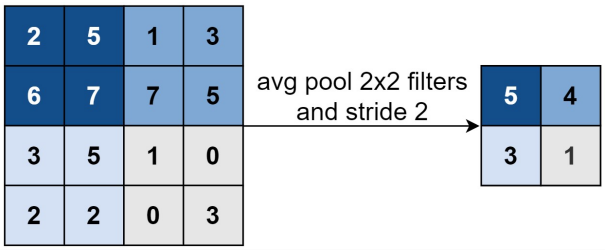


Figure 4.6: Example of an average pooling operation.

After the convolution and pooling layers, one or more fully-connected layers can be added to wrap up the convolutional neural network architecture. Here, neurons have full connectivity with all elements of the preceding and succeeding layers. Inputs are flattened into a one-dimensional vector with the same number of parameters (flattening layer). Fully-connected layers map high-level features computed by the previous convolution and pooling layers to generate the final outputs of the network, that is, the likelihood of each class in the case of a classification NN. Even though an activation function follows each fully-connected layer, the one applied to the final fully-connected layer is usually distinct and selected according to the specific task at hand.

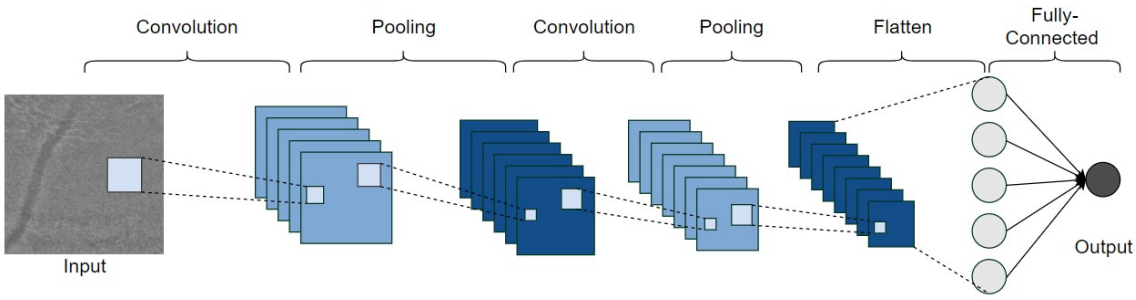


Figure 4.7: Illustration of a Convolutional Neural Network.

4.3 Gradient-based Optimisation

Deep learning algorithms solve the mathematical problems by iteratively updating solution estimates, rather than analytically derive the equations that lead to the optimal solution. In other words, deep learning algorithms imply optimisation - the task of finding the parameters θ that minimise the loss function, cost function or error function $J(\theta)$.

The optimisation is based on a few mathematical concepts. Suppose a 1D function $f = f(x)$, where x is a real value. Its derivative $f'(x)$ gives the slope of the graphic of $f(x)$ at point x. Hence, the derivative indicates how small changes in the

input x produce changes in the output through the linear approximation: $f(x + \epsilon) \approx f(x) + \epsilon f'(x)$.

For functions with multiple inputs, the partial derivatives $\frac{\partial f(x)}{\partial x_i}$ are computed. They measure how f changes with respect to each variable x_i at point x. The partial derivatives are stored in a vector called gradient, denoted as $\nabla_x f(x_t)$.

The loss function is decreased by moving the inputs in the opposite direction of the gradient. This optimisation method is known as gradient descent, whose expression is presented below:

$$\theta_{t+1} = \theta_t - \eta \nabla_\theta J(\theta_t), \tag{4.2}$$

where η is the learning rate, a positive scalar that determines the size of the step, and t is the time step. .

When $\nabla_\theta J(\theta_t) = 0$, the gradient lies in a critical or stationary point. This can either be a local minimum if it is lower than all its neighbouring points; a local maximum if it is higher than all the adjacent points or; a saddle point if its neighbours are both higher and lower.

Ideally, the optimisation algorithm should find a global minimum - the lowest value of the loss function. However, this can be challenging because there may be many saddle points surrounded by flat regions and many local minima that may not be optimal. Therefore, in deep learning, after a few epochs without improvements on the loss function, the training is ceased, and the loss function value is accepted.

Before moving to a deeper explanation of the loss function, a summary of the learning process will clarify the purpose of the methods involved.

During training, the inputs x propagate through the hidden layers until reaching the output layer and producing the output y. This is known as forward propagation. Then, the gradient $\nabla_\theta J(\theta_t)$ of the loss function is computed backwards, i.e., from the output to the input by the back-propagation method. Finally, an optimisation algorithm is used to update the parameters θ - the weights and biases - using the computed gradients. These steps are repeated at each iteration.

4.3.1 Loss Function

The loss function, also known as the objective function, compares the predicted and actual values, measuring how well the model's prediction matches the correct

value.

Binary cross-entropy is an appropriate loss function in binary classification problems where the network's output lies between 0 and 1. It is defined as:

$$L = -(y\ log(p) + (1-y)\ log(1-p)) \tag{4.3}$$

$$= \begin{cases} -log(1-p), & \text{if } y = 0 \\ -log(p) & \text{if } y = 1, \end{cases} \tag{4.4}$$

where y is the true class (0 or 1), and p is the predicted probability between 0 and 1.

4.3.2 Optimisation Algorithms

In multi-dimensional problems, the learning rate is a difficult hyperparameter to set because the loss function may increase rapidly in one direction and slowly in another one. Hence, this section focuses on two optimisation techniques with adaptive learning rates: Root Mean Square Propagation (RMSProp) and Adaptive Moment Estimation (Adam). These algorithms choose a different learning rate for each parameter, and they are updated separately throughout the learning course.

The RMSProp algorithm [64] adapts the learning rate of all model parameters. The parameters with higher partial derivative values of the loss function have a more rapid decrease in their learning rate. In contrast, parameters with lower partial derivative values have a lower reduction in their learning rate. The learning rate is updated using an exponentially decaying average of squared gradients to ignore the history so that the algorithm can converge fast after finding a convex bowl. The current average of squared gradients $E[g^2]_t$ depends on the previous average $E[g^2]_{t-1}$ and the current gradient $g_t = \nabla_\theta J(\theta_t)$.

$$E[g^2]_t = 0.9E[g^2]_{t-1} + 0.1g_t^2 \tag{4.5}$$

The parameters θ are updated according to the following equation:

$$\theta_{t+1} = \theta_t - \frac{\alpha}{\sqrt{E[g^2]_t + \epsilon}} g_t, \tag{4.6}$$

where ϵ is a smoothing term that avoids division by zero (usually in the order of

10^{-8}).

The optimisation algorithms result from improvements on the gradient descent. For example, the difference between the gradient descent (equation 4.2) and the RMSProp (equation 4.6) is that, in the latter, the learning rate is divided by the squared root of the average of squared gradients $E[g^2]_t$ plus the smoothing term ϵ.

Adam [65] is a combination of RMSProp and Stochastic Gradient Descent (SGD) with momentum. It uses estimators of the first and second moments of the gradient, the mean m_t and the uncentered variance v_t, respectively, to adapt the learning rate of each parameter, as denoted below:

$$\theta_{t+1} = \theta_t - \frac{\alpha}{\sqrt{\hat{v}_t} + \epsilon}\hat{m}_t \tag{4.7}$$

Training large training sets can be time-consuming. For that reason, at each iteration, the loss function estimates are computed using only a small group of samples.

These mini-batches are usually a small number of examples, ranging from 1 to a few hundred. Their size is often a power of 2, especially when using GPUs, for a shorter runtime. Mini-batches must be selected randomly and are independent of each other to compute an unbiased estimate of the loss function.

4.4 Regularisation Methods

A fundamental problem in deep learning is how to create a NN that will perform well not only on the training set but also on new (unseen) samples.

Thus, there are two factors crucial to determine whether an algorithm has good performance:

1. achieving a small training error, and;
2. getting a small gap between training and test error.

The ability of a model to perform well on new data is called generalisation. It is usually measured on a test set of samples that were collected separately from the training set. The error computed on the test set is called generalisation or test error, and it should be as low as possible.

There are two terms related to the generalisation ability of a model. Underfitting occurs when the model learns the underlying patterns of the data too poorly. As a result, the model fails to obtain a small training set error. On the other hand, overfitting occurs when a model fits too closely to the training data, therefore being unable to generalise well and make correct predictions on new inputs. In this case, the gap between the training error and test error is significantly large.

Some strategies to reduce the test error while maintaining the training error can be applied to the learning process. These approaches are called regularisation. Unfortunately, as there is no best form of regularisation, its choice is made according to the task at hand.

There are many regularisation strategies for deep learning models, such as L1 and L2 regularisation, dropout, and sparse representations. In this section, only data augmentation, early stopping, and batch normalisation will be described because these were the ones used in the present study.

4.4.1 Data Augmentation

An effective way of improving the generalisation ability of a model is to learn using a more extensive training set. However, the amount of data available is, sometimes, quite limited. A practical solution to increase the number of training examples is to apply transformations to the samples, creating new data.

In image classification problems, classic data augmentation techniques include cropping, rotating, and resizing the images. By adding diversity to the training set, overfitting is reduced.

However, one must be careful when applying the transformations not to change the correct class. For example, rotating a “6” by 180 degrees would turn it into a “9”.

4.4.2 Early Stopping

When training a model, a situation is frequently observed: over time,training and validation errors gradually decrease at the beginning of the learning process. However, as the learning process continues, the validation error begins to rise at some point, indicating the model starts overfitting the training set.

The early stopping motivation is that the optimal model parameters would be the ones when the validation error (and, hopefully, the test error) is the lowest.

Every time the validation error decreases, a copy of the model parameters is stored. When the training algorithm terminates, the parameters are returned rather than those obtained at the last epoch. After a pre-specified number of iterations, the learning ends if the error has not improved over the best-recorded validation error.

Early stopping is a simple and popular technique with many advantages. Besides reducing computational cost by limiting the number of training iterations, it also does not require any changes to the learning procedure.

4.4.3 Batch Normalisation

Batch normalisation is a widely used technique in the field of deep learning to improve optimisation. Although it is not considered a regularisation method, it is computed over mini-batches instead of the entire dataset, adding some noise to the model. As a result, it reduces overfitting and stabilises learning, thereby having a slight regularisation effect [62]. For that reason, batch normalisation is included in this section.

Batch normalisation standardises the mean and standard deviation of each unit over mini-batches of training data through back-propagation. This way, the gradient never suggests an operation such as increasing or decreasing the neuron's mean or standard deviation.

This method is applied to the units within a layer after the transformation $z = w^T x$ and before the activation function, in the following manner:

$$z_{normalised} = (\frac{z - \mu_z}{\sigma_z}) * \gamma + \beta \tag{4.8}$$

where μ_z is the neuron's mean, and σ_z is the standard deviation over the mini-batch.

β and γ are learned parameters that result in a layer that has a distribution with mean β and standard deviation γ. This reparametrisation is easier to learn with the gradient descent than normalising the layer to have zero mean and unit variance [62].

One should note that the regular equation for computing the neuron's output $z = w^T x + b$ was replaced by $z = w^T x$. The bias term was omitted because it becomes redundant due to adding the parameter β to the normalisation.

In CNNs, every spatial location within the feature map should have the same

mean and standard deviation so that the feature map statistics remain the same regardless of spatial location [62].

4.5 Design Methodology

When building a deep learning model, besides understanding the existing algorithms and the underlying principles, it is imperative to carefully design the model, always according to the task at hand.

The designing process includes the following steps:

- identify the objectives and consequently define the performance metrics to use - this step is essential since the error metrics will guide future actions to improve the model's performance;
- set up the dataset - the quality of data strongly influences the success of complex data analysis. It reduces model complexity, making the training process faster. Furthermore, it improves the generalisation ability of a model, thereby enhancing its performance;
- design the NN architecture, and;
- make incremental changes such as gathering more data, adjusting hyperparameters or trying new algorithms based on the findings from the previous step.

Before moving to the explanation of each step, it should be noted that the recommendations were adapted from Goodfellow et al. (2016) [62] and Yu et al. (2007) [63].

4.5.1 Performance Metrics

The performance metrics should be chosen according to the problem being tackled and the distribution of the labels on the dataset. In binary classification tasks, accuracy, sensibility, specificity, F1-score, and confusion matrix are commonly used to evaluate the models' performance and their ability as class predictors.

The formulation of the metrics are denoted below:

1. **Accuracy:** rate of predictions correctly classified;

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \tag{4.9}$$

2. **Sensitivity:** rate of positives correctly classified;

$$Sensitivity = \frac{TP}{TP + FN} \tag{4.10}$$

3. **Specificity:** rate of negatives correctly classified;

$$Specificity = \frac{TN}{TN + FP} \tag{4.11}$$

4. **F1-Score:** harmonic mean between precision and recall;

$$Precision = \frac{TP}{TP + FP} \tag{4.12}$$

$$Recall = \frac{TP}{TP + FN} \tag{4.13}$$

$$F1 - Score = 2 * \frac{Precision * Recall}{Precision + Recall} \tag{4.14}$$

5. **Confusion matrix:** summarises the model's performance by providing the correct and incorrect classifications for each class.

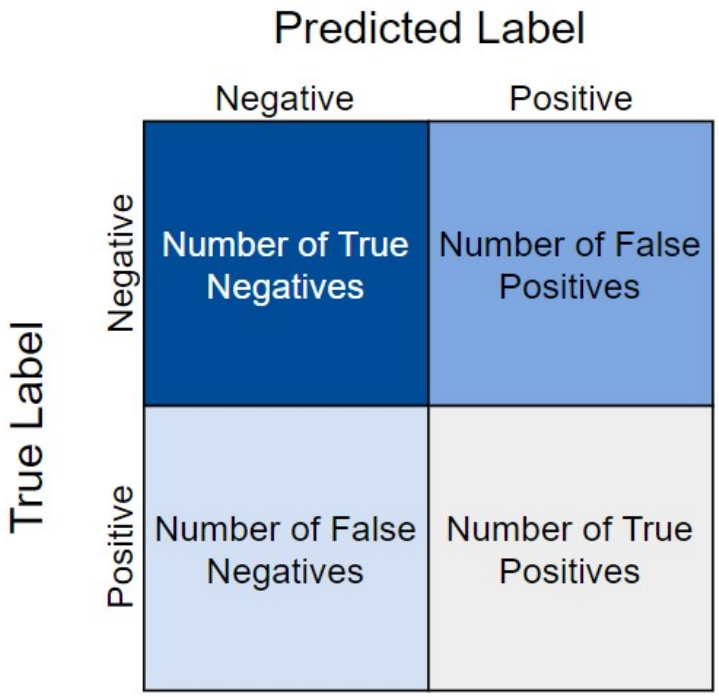


Figure 4.8: Confusion matrix.

Accuracy is one of the most intuitive performance metrics. However, it can be misleading when dealing with imbalanced datasets. For instance, a poor classifier may achieve good performance by ignoring the underrepresented class. Sensitivity and specificity are common metrics in binary problems to evaluate the ability of the model to predict the presence or absence of a specific condition. F1-score is a widely used metric, especially in imbalanced datasets, and it gives an overall idea of the model's performance by combining precision and recall.

4.5.2 Data Preparation

Data preparation consists of three phases: data pre-analysis, data preprocessing, and data post-analysis.

Data pre-analysis begins with understanding the kind of data required for a specific task. Then, based on these requirements, the data is collected. Afterwards, another two steps may be taken: feature selection to remove irrelevant or redundant features and data integration for when the data were collected from multiple sources, potentially being disordered and scattered.

The second phase, data preprocessing, identifies issues in the dataset such as data missing, noisy data with outliers, or features with different ranges. Accordingly, there are many processing techniques to fix the encountered problems - data normalisation, data denoising, image histogram equalisation, etc.

Finally, the last phase consists of splitting the dataset into training, validation, and test sets.

The training set is used to learn the parameters of the NN. Thus, the model sees and learns from this subset of data.

The validation set is constructed from the training set, i.e., the training set is divided into two disjoint subsets: the training set and the validation set. The latter provides an unbiased evaluation during training, allowing the hyperparameters to be selected accordingly. Thus, the model sees the data but never learns from it.

The test set is used to estimate the generalisation error of the model after the learning process has terminated. It is used only once and provides an unbiased evaluation of the final model, simulating how it would behave in the real-world, i.e., in unseen data.

4.5.3 Model Design

After choosing the performance metrics and preparing the dataset, it is essential to establish an end-to-end model [62].

Firstly, one should choose the NN based on the structure of the dataset. For example, if the inputs are fixed-sized vectors, a reasonable choice would be FFNNs. On the other hand, if the inputs are images, then a CNN would be a wise option.

Besides, it is also necessary to choose an optimisation algorithm. SGD, RMSProp, or Adam are popular choices. Furthermore, the loss function should be selected according to the type of task at hand. For example, in the case of a binary classification problem, binary cross-entropy is a reasonable option.

If the collected dataset does not contain millions of samples, then some regularisation methods should be included. For example, one can use early stopping in almost any kind of task [62]. Data augmentation is also an attractive technique to reduce the generalisation error in a wide range of computer vision models.

Lastly, if the problem is similar to another task already studied with success, a good strategy is to use the model trained in that task and use it to solve the problem at hand. This is called transfer learning.

4.5.4 Hyperparameters Optimisation

Deep learning models have parameters and hyperparameters. Parameters are learned during training, while hyperparameters cannot be directly estimated from the data and must be specified by the user. Thus, hyperparameters optimisation is the process of finding the best combination of hyperparameters that maximises the model performance on the validation set.

A common method is grid search. It methodically tries combinations of hyperparameter values and evaluates the model for each combination. This can be implemented using k-fold cross-validation. This type of cross-validation works by randomly splitting the training set into k-folds of roughly equal size, where each observation is assigned to one group and remains in the group until the end of the process. The model is then trained using k-1 folds and evaluated on the remaining one. The process repeats k times, and the model performance estimate is the average of the performance values obtained in each iteration. Figure 4.9 shows an illustration of how this method works. Although it demonstrates a five-fold cross-validation, the same principle applies to any k-fold cross-validation. The k-fold cross-validation is

repeated to every combination of the hyperparameters to decide the best one.

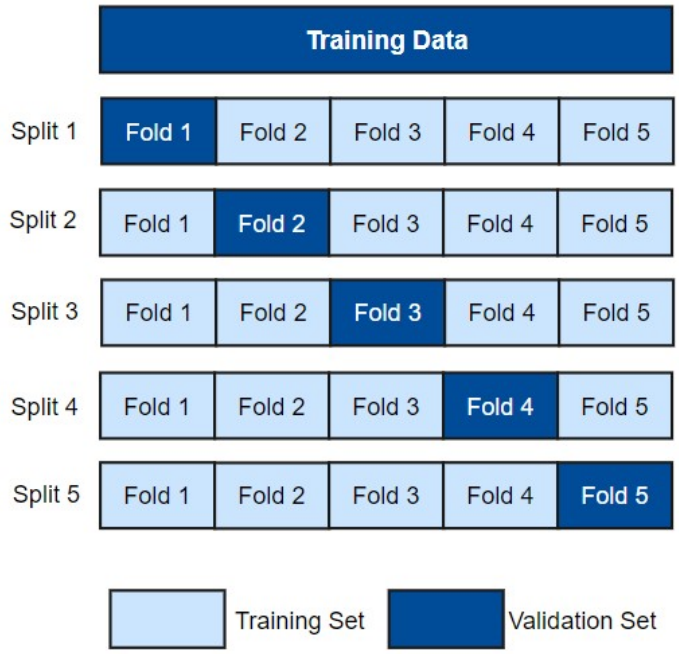


Figure 4.9: Demonstration of a five-fold cross-validation.

4.6 Transfer Learning

Transfer learning can be defined as transferring the knowledge learned by a network while solving a task and applying that knowledge to a new, but related, problem.

As aforementioned, CNNs automatically generate powerful discriminative features using a hierarchical learning approach [66]. Feature maps in the earlier layers identify more simple patterns, such as lines, edges, and colours (for multichannel images). The last layers progressively detect more specific details of the classes contained in the dataset, integrating local and global information.

Pre-trained CNN models have been successfully applied to several computer vision and medical imaging tasks as a feature generator or baseline for transfer learning [67]. Training a deep CNN from scratch requires a substantial dataset to ensure proper convergence and computational and memory resources. A promising alternative is to fine-tune a CNN that has been trained using a different dataset with a large number of labelled data.

Fine-tuning a model is the process of retraining some of the deeper layers of the CNN. This way, the high-level features learned by the last layers are adjusted to

make them more relevant to the specific domain at hand. The weights of the shallow layers are kept frozen because, as already mentioned, they identify low-level features independent of the target task, suitable for addressing a large set of domains.

Depending on the differences between the source and target datasets a deeper fine-tuning may be required, i.e., retraining a higher number of the deep layers so that the CNN learns relevant features needed to classify the new dataset.

CNNs have been applied to computer vision tasks since the late 80s. However, very deep CNNs only gained tremendous popularity when Krizhevsky et al. (2012) [68] won the ImageNet Large Scale Visual Recognition Challenge (ILSVRC) 2012 competition with "AlexNet". This was in view of the developments in computing power and substantial amounts of labelled data [69]. AlexNet was trained over one million images from the ImageNet dataset containing one thousand object categories.

Since then, researchers have been focusing on finding higher-performing CNNs, resulting in deeper and wider networks with significantly improved performance.

In the ILSVRC 2014 contest, the winners Szegedy et al. (2015) [70] introduced a new architecture called Inception (also referred to as GoogleNet). It introduced dimensionality reduction, achieving outstanding performance. The runners-up of the same contest, Simonyan and Zisserman (2014) [71], presented another classic deep network - VGGNet - that keeps the hyperparameters constant along with the network's depth. The ILSVRC 2015 contest winners, He et al. (2015) [72], used an even deeper CNN with a residual learning framework - the ResNet.

The Inception architecture has been modified over the years, culminating in new versions: Inception-v2 [73] in which batch normalisation was combined with the Inception architecture; Inception-v3 [73] that introduced factorised convolutions and a more aggressive dimensionality reduction; Inception-v4 [74], a simplified architecture with more Inception modules; and Inception-ResNet [74], an Inception architecture integrated with residual functions.

In the following section, Inception-v3 architecture will be described. It was developed for assisting in image analysis and object detection. As a result, this network has learned valuable feature representations from a wide range of images and objects.

4.6.1 Inception-v3

This network was proposed in the paper "*Rethinking the Inception Architecture for Computer Vision*" [73], and it has been extensively used in medical applications through transfer learning [75].

When designing the first version of the Inception network, some principles guided the researchers:

1. highly performant deep neural networks need to be large in both depth - number of layers - and width - number of units at each layer;

2. similarly to the biological human visual cortex, which identifies patterns at different scales and aggregates them to enlarge further the perception of objects, CNNs benefit from extracting features at varying scales as well, and;

3. consideration of the Hebbian Principle — *neurons that fire together, wire together.*

However, these propositions led to two considerable shortcomings: large networks are prone to overfitting due to the significant number of parameters and having multiple kernels of varying sizes increases the requirements in computing power.

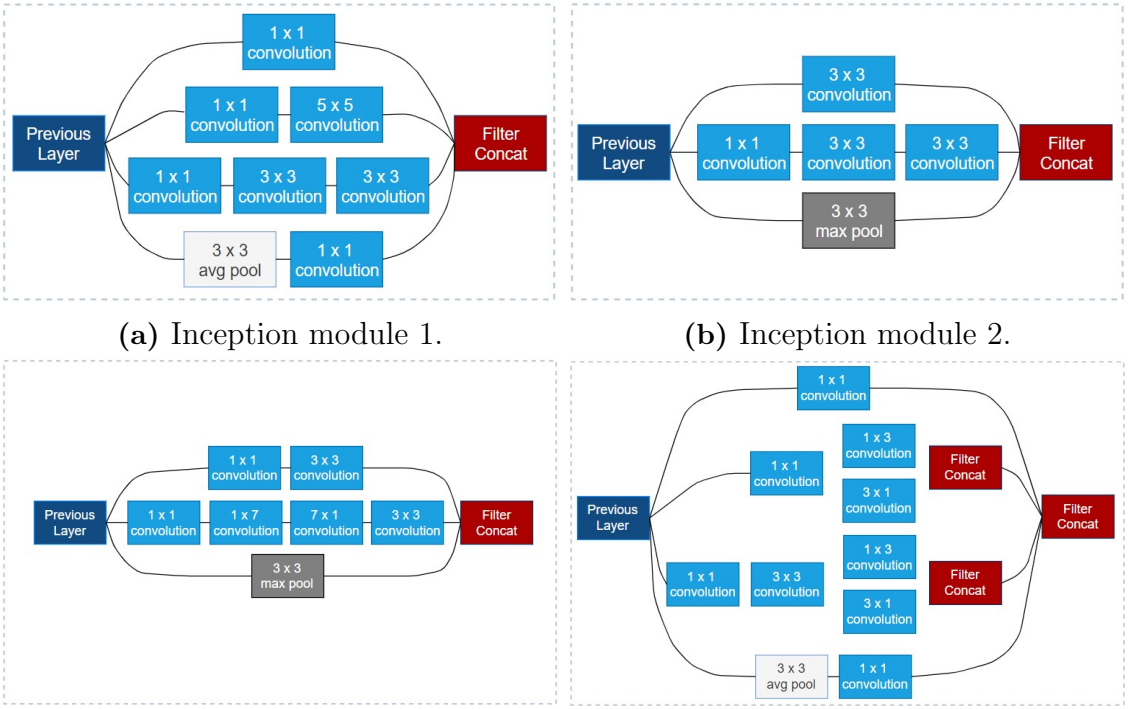
All versions of Inception architectures are organised in Inception modules stacked on top of each other to overcome the above obstacles. Each module consists of pooling layers and convolutional filters of different sizes with ReLU as the activation function. The output of each operation is concatenated and sent to the next Inception module.

The Inception modules of an Inception-v3 network are built with the following individual components:

i. **1 x 1 convolutions**

A 1 x 1 convolution maps an input pixel with all its respective channels to one output pixel. Each filter has the dimension of 1 x 1 x n, where n is the number of filters.

By introducing a 1 x 1 convolution, computational cost can be drastically reduced and hence the network's depth and width can be increased without any performance penalisation. Although a 1 x 1 kernel does not learn spatial patterns within the image, it learns patterns across the channels (in-depth),



(a) Inception module 1. **(b)** Inception module 2.

(c) Inception module 3. **(d)** Inception module 4.

Figure 4.10: Inception modules of Inception-v3.

enabling the network to learn more [76].

ii. **3 x 3 and 5 x 5 convolutions**

The 3 x 3 and 5 x 5 convolutions learn spatial patterns at different scales within each channel.

To utilise the computation even more efficiently, the Inception-v3 was added with the factorised convolutions explained below:

(a) **Factorisation of 5 x 5 convolutions into two 3 x 3 convolutions**

Convolutions with larger spatial filters, such as 5 x 5 or 7 x 7, tend to be computationally expensive. For this reason, some of the 5 x 5 convolutions in GoogleNet were replaced with two layers of 3 x 3 convolutions in Inception-v3, reducing the number of parameters by 28%.

(b) **Factorisation in asymmetric convolutions**

The 3 x 3 convolutions of the last Inception modules were divided into asymmetric convolutions of 1 x 3 followed by 3 x 1 convolutions. These

two operations have the same receptive field as a 3 x 3 convolution but are computationally more effective.

The researchers found that this factorisation does not work well on early layers. Consequently, the asymmetric convolutions 1 x 7 followed by 7 x 1 and 1 x 3 followed by 3 x 1 were applied only on the last three Inception modules of the network.

iii. **Maximum and average pooling layers**

As pooling operations have been essential for the success of convolutional networks, the researchers considered that a parallel pooling path in each Inception module would have an additional beneficial impact.

iv. **Concatenation layer**

In the concatenation layer, all the feature maps from the convolution layers are combined into one object to create a single output of the Inception module increasing the network representational power.

Finally, batch normalisation and the ReLU activation function are applied after every convolutional layer. Moreover, the Inception-v3 also includes a *label-smoothing regularisation* mechanism to increase the ability of the model to adapt by preventing it from becoming too confident about its predictions [73].

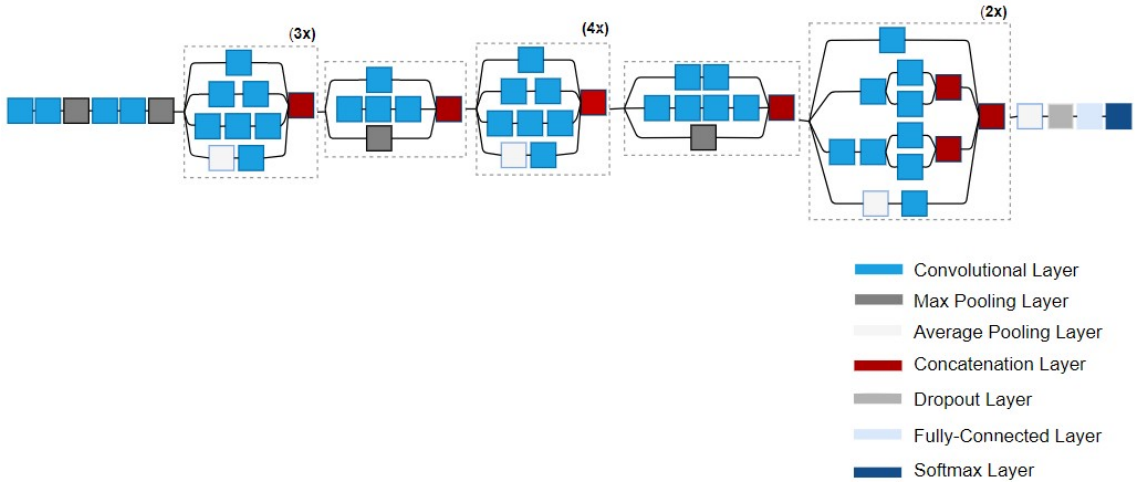


Figure 4.11: Schematic diagram of Inception-v3.

4.7 Explainability

Deep learning methods have been very effective for a myriad of medical diagnostic tasks, even outperforming human experts on some of those [77]. However, despite attaining impressive results, the underlying black-box nature of these algorithms has restricted their deployment in clinics. It arises from the fact that even understanding the underlying statistical principles and knowing the weights of the neurons of such models, they lack the ability to represent the knowledge for a given task explicitly .

To gain the trust of physicians, regulators and patients, a medical diagnosis model needs to be transparent, understandable by itself, and explainable. It should clarify the whole logic behind each decision while maintaining a high level of performance.

Explainability is very important to build safe, ethical, and trust-able deep learning models, and it is a key enabler for its deployment in medicine. To begin with, medical professionals can be provided with the reasoning for the model decision and thereby build trust among end-users and help recognise questionable choices to improve the model further [77]. Secondly, deep learning practitioners can analyse the model features, such as the pixels of an image that contribute to the output decision and potentially make corrections to the model's deficiencies.

4.7.1 Visual explanations in Convolutional Neural Networks

To explain how the classification decisions are made, a backpropagation-based attribution method can be applied to the CNNs. These algorithms are powerful tools to determine which area(s) of the input image contribute to the output prediction.

The Gradient-weighted Class Activation Mapping (Grad-CAM) is a backpropagation technique developed by Selvaraju et al. (2017) [78]. This approach is a generalisation of Class Activation Mapping (CAM) introduced by Zhou et al. (2016) [79]. While CAM is only applicable to a particular kind of CNNs, the ones that do not contain fully-connected layers, Grad-CAM is relevant to a broader range of CNN families since it does not require any modification to the network architecture.

Grad-CAM works as follows: firstly, the gradient of the output class prediction y^c (before the sigmoid) concerning the feature map activations A^k of the last convolutional layer is computed. Using the score y^c before the sigmoid layer is important to ensure the gradients are correctly computed. An activation function would mod-

ify the algorithm's outcome. Commonly, it is the last convolutional layer because the later layers in a CNN capture high-level patterns and CNNs naturally retain spatial information lost in fully-connected layers. Consequently, the last convolutional layer is expected to have the best compromise between high-level semantics and detailed class-specific spatial information.

$$\text{gradient} = \frac{\partial y^c}{\partial A^k} \qquad \text{(computed via backpropagation)} \tag{4.15}$$

Then, this gradient is spatially pooled using global average pooling over the width and height dimensions (indexed by i and j respectively) to obtain the weights α_k^c. These weights capture the importance of each feature map for the class prediction.

$$\alpha_k^c = \underbrace{\frac{1}{Z}\sum_i \sum_j}_{\text{global average pooling}} \overbrace{\frac{\partial y^c}{\partial A_{i_j}^k}}^{\text{partial derivatives via backpropagation}} \tag{4.16}$$

The Grad-CAM heatmap is a weighted sum of feature map activations, A^k, with weights α_k^c. Positive pixel values in the heatmap mean they positively influence the class of interest, while negative pixels are likely to contribute to other categories in the image.

Finally, the heatmap is passed through a rectified linear activation unit (ReLU) to filter out the negative weights, allowing the visualisation of the image regions that led to the class decision.

$$L_{Grad-CAM}^c = ReLU \overbrace{(\sum_k \alpha_k^c A^k)}^{\text{linear combination}} \tag{4.17}$$

4.7.2 Feature Importance in Neural Networks

There are several methods whose aim is to assess the contribution of each input feature for the output prediction. In this section, two methods will be described: the connection weights algorithm and the perturbation method.

The connection weights algorithm, proposed by Olden et al. (2004) [80], aims to quantify the importance of each input variable in the prediction process of a shallow

NN with one hidden layer. This method considers the magnitude and direction of each weight. The greater the weight's magnitude, the higher the contribution of the input feature to the prediction process.

For each input feature, the algorithm calculates the sum of the product of the connection weights between the input neurons and the hidden neurons with the weights between the hidden neuron and the output neuron.

Features with positive input-hidden and positive hidden-output weights, or negative input-hidden and negative hidden-output weights have a positive contribution. In contrast, positive input-hidden and negative hidden-output and negative input-hidden and positive hidden-output weights have a negative effect.

Thus, the multiplication of the two connection weights directions indicates the contribution of the input feature on the output prediction. The higher the sum, the more significant the contribution of the corresponding input feature.

The mathematical expression is denoted below.

$$R_I = \sum_{H=1}^{h} (w_{I-H} \cdot w_{H-O}), \tag{4.18}$$

where R_I is the relative importance of the input feature I, h is the total number of neurons in the hidden layer, w_{I-H} is the weight of the connection between the input neuron I and the hidden neuron H, and w_{H-O} is the weight of the connection between the hidden neuron H and the output neuron O.

Differently, the perturbation method analyses the impact on the output of changing each input variable [77]. This is implemented by modifying the values of one input feature at a time while retaining all other features' original values. The difference between the original and the current performance on the test set is noted down.

The input features affecting the performance the most are considered the most important ones. In the end, input variables are ordered by importance according to the achieved performance.

While in the connection weights method the estimates are obtained by training the NN only once, in the perturbation method the model needs to compute a forward pass for every modified feature, which is computationally more expensive. On the other hand, the perturbation method can be applied in a wide range of deep learning

models, such as FFNNs with many hidden layers and CNNs, whereas the connection weights algorithm can only be used in shallow NNs.

5

Dataset

5.1 Animals

The dataset comprises OCT volume data of 57 triple transgenic mice model of Alzheimer's disease (3×Tg-AD) and 57 wild-type mice (healthy controls), of both eyes, at ages of one, two, three, and four-months-old. The triple transgenic mice express three human genes associated with familial AD: Swedish Amyloid Precursor Protein (APPswe), Presenilin 1 (PSEN1), and Microtubule-Associated Protein Tau (MAPT).

The dataset is intentionally balanced to mitigate the effects caused by any imbalanced and avoid the need for additional measures to tackle it.

All mice were kept at the vivarium of the Coimbra Institute for Clinical and Biomedical Research (iCBR), Faculty of Medicine, University of Coimbra, and were on a 12-h light/dark cycle with free access to food and water. Each animal was followed closely throughout the study.

Mice models of disease have been widely used in biomedical research, as they are small and easily maintained, conserving almost 99% of the human genome and physiologically resembling humans [81]. Because of their similarity to humans, mice are good models to address senescence, despite their short lifespan - about 24 months.

As illustrated in figure 5.1, one to four months of mice's age is equivalent to 20 years in humans, the period when cerebral changes precede AD symptoms, according to the literature [9]. Thus, the time points one, two, three, and four months of age were chosen, as they cover the onset period of the disease to its detection by current state-of-the-art diagnostic tools.

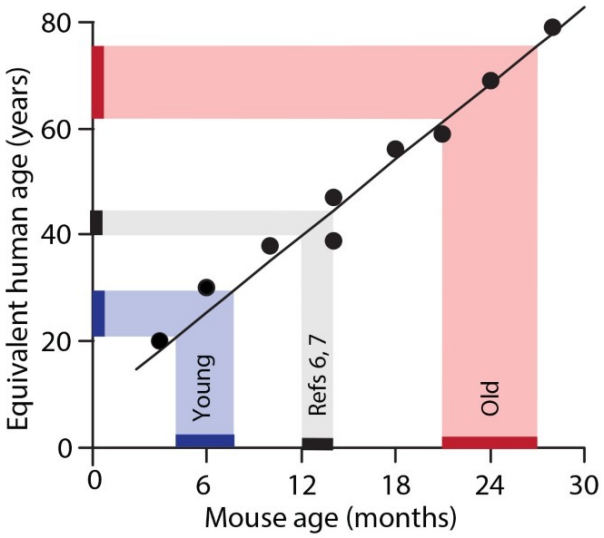


Figure 5.1: Correspondence between the ages of humans and mice - for mice ages between 3 and 30 months. Image from "Diminished KCC2 confounds synapse specificity of LTP during senescence" [82].

5.2 Data Preprocessing

The retinas were imaged with a Micron IV OCT System (Phoenix Technology Group, Pleasanton, CA, USA). From each eye, at each time point, OCT data was segmented to compute six different images corresponding to the following layers/layer-aggregates: two coupled layers – RNFL-GCL complex -, four anatomical layers – IPL, INL, OPL, and ONL -, and the TR, which encompasses all anatomical layers. RNFL and GCL are coupled as it is still not possible to segment them apart.

Some eye scans were excluded from the dataset due to poor image quality, yielding 4 698 fundus images (six images per eye scan). The dataset is composed of:

Table 5.1: Number of eye scans (OCT volumes) per group, eye and time-point. Legend: OD is right eye, and OS is left eye.

	One month		**Two months**		**Three months**		**Four months**	
	OD	OS	OD	OS	OD	OS	OD	OS
WT	55	52	50	47	49	43	39	37
3×Tg-AD	49	50	52	54	53	53	49	51
Total	104	102	102	101	102	96	88	88

The ocular fundus images were generated from the 3D OCT data by computing the average of the A-scan values comprised within the boundaries of each layer/layer-aggregate of interest. This method is known as Mean Value Fundus (MVF) [83].

The preprocessing step in neural networks is utterly important to achieve faster convergence and the most accurate results. In this thesis, the preprocessing phase consisted of the following steps:

1. Contrast normalisation - the Contrast-Limited Adaptive Histogram Equalisation (CLAHE) algorithm was used to attenuate any potential differences between mice groups. CLAHE computes several histograms corresponding to different image regions, and, based on their analysis, intensity values are adjusted to create a uniform grayscale distribution [84];

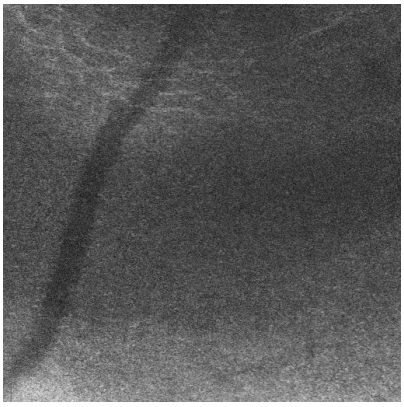
2. Intensity normalisation - a Gaussian low-pass filter was applied to the images resulting from the previous step. A low-pass filter is used to attenuate the high-frequency components while preserving the low-frequency ones, therefore allowing to spot image intensity variations at the scale of interest. This filter needs two values for its implementation: a sigma and a kernel size. The sigma controls the variance around the mean. In other words, a filter with a small sigma has a steeper peak, and filter weights decrease sharply from the central value. With a large sigma value, a broader effect is spotted. Considering the image size (512 x 512 pixels) and the modulation effects present in some images, the sigma value was set to 25 and the filter size set to 121x121 (to allow the Gaussian to attain values close to zero at the limits of the filter). Finally, intensity normalisation is achieved by the division between each CLAHE image and its low-pass filtered version. This process renders normalised images across eyes, timepoints and mice groups, ensuring the model receives ocular fundus images less prone to contain biased information, which could mislead the model.

Each step of the image enhancement process is illustrated in figure 5.2.

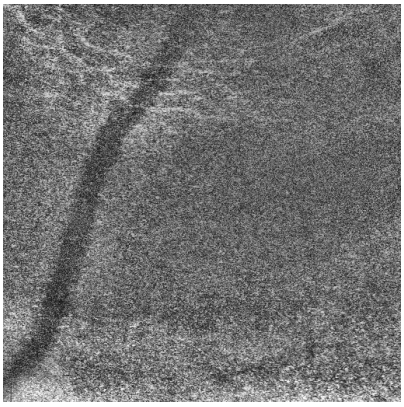
After the above correction process, images may still contain different ranges of pixel intensity values between images. However, it is highly desired that images have the same statistical distribution as the shear average and standard deviation may be due to acquisition conditions and not on actual differences. Therefore, a technique to standardise these images was implemented. This method involves rescaling the image data to zero mean and unitary variance (Z-scoring) followed by

a multiplication and an addition to achieve a mean intensity value of 128 and a variance of eight. While the mean intensity was simply defined as the mean value of the range of intensities (8-bit images), the variance value was chosen by probing the dataset to keep the pixels' intensity within the range 0-255. These images were saved in a non-compressed image file format.

All preprocessing was implemented in MATLAB Release 2021a (The MathWorks, Inc., Natick, Massachusetts, United States).

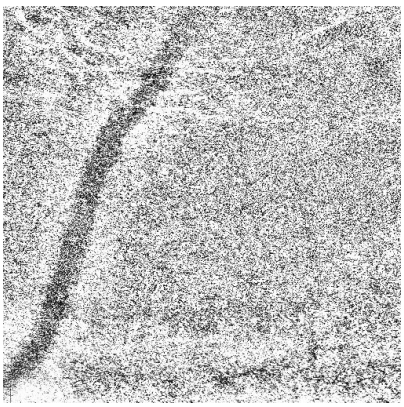


(a) Original Image.



(b) CLAHE image.

(c) Low-pass filtered CLAHE image.



(d) Fully-normalised image.

Figure 5.2: Illustration of image enhancement process. Mouse ID: D_D953_T01_OS_INL

6

Results and Discussion

The ocular fundus images dataset was partitioned into training, validation, and test sets. Each subset has the same ratio of transgenic and wild-type mice, and the split was done randomly. Moreover, the division was made by mice, thereby ensuring that data from one mouse is not present in different subsets.

Eleven mice were picked from each class to set the test set, corresponding to 20% (157 eye scans) of the total data. These were kept across all CNNs and FFNN models. The remaining 80% (626 eye scans) of the data were used for hyperparameters optimisation and model training, with 20% (125 eye scans) of these constituting the validation set.

The dataset includes the time points one, two, three, and four months of age. These four time points were joined together so that the deep learning models do not depend excessively on a particular time point. Instead, they should attend to the global characteristics of AD progression. Getting transverse discriminators to time points is important because, in contrast to the mice data used in this study, in which we know the duration of AD, in humans, we cannot assuredly know when changes related to AD have started.

All deep learning models used in this study were built in Python 3.7.9 using the Keras [85] framework with TensorFlow [86] as the back-end. The experiments were run on a 3.30GHz i7-5820K computer with an NVIDIA GeForce RTX 2080 SUPER graphics card.

6.1 Deep Learning Models to Detect Changes in the Retina

This study aims to identify changes in retinal images of animal models of AD using a deep learning approach. Therefore, the first goal was to assess the ability

of CNN models to distinguish between wild-type and transgenic mice using ocular fundus images from a particular layer or layer-aggregate (RNFL-GCL, IPL, INL, OPL, ONL, and TR).

Each CNN learns from images belonging to only one layer/layer-aggregate, resulting in six distinct CNN models. Their evaluation provides two important observations: whether retinal changes exist between both groups and whether the CNNs manage to classify correctly based on those changes.

The six CNNs were trained using transfer learning. Considering its performance and computational resources, Inception-v3 [73] was the chosen pre-trained CNN for the task at hand. Nevertheless, few adjustments to both the Inception-v3 architecture and training methodology were made to leverage the knowledge gained by the original model on the ImageNet dataset and integrate it with the knowledge learned from the ocular fundus dataset:

1. The classification layer was replaced by a global average pooling layer followed by a fully-connected layer with a sigmoid activation function;

2. The Inception-v3 models were trained in two separate phases. Firstly, only the weights of the classification layer were updated, while the weights of the remaining layers were kept unchanged. Afterwards, the models were fine-tuned by allowing the retraining of the 15 last layers. This decision was made by looking at the Inception-v3 architecture. As it consists of Inception modules of varying filter sizes, the reasonable option was retraining entire modules. Furthermore, although the ocular fundus dataset size is considerably large, it does not contain millions of examples. Thus, to avoid overfitting, the CNN was fine-tuned from the last Inception module onwards.

Fine-tuning the models was applied after the classification layer had been trained. Otherwise, the unfrozen layer weights would randomly initialise. Consequently, the error signal propagating through the network would be too large, disrupting the patterns previously learned by the layers being fine-tuned [87].

Although CNNs have many hyperparameters possible to optimise, only three were tuned: the learning rate, batch size, and number of epochs. The optimisation approach consisted of grid-search with five-fold cross-validation. As the dataset is nearly balanced, accuracy was the metric used for evaluation. The range of possible values was chosen based on papers whose aim was to classify retinal images using deep learning [88–90], a very close area of application to the one herein presented, although based on substantial different images.

Table 6.1 compiles the list of values tested for each hyperparameter.

Table 6.1: List of values used in the CNNs optimisation.

Hyperparameters	**Grid-search Values**
Learning Rate	{0.01, 0.001, 0.0001}
Batch Size	{8, 16, 32, 64}
Number of Epochs	{20, 30, 40, 50}

Regarding the optimisation algorithm, RMSProp was chosen because it was the one used by the Inception-v3 authors to train their network on the ImageNet dataset [73].

Also, binary cross-entropy was the selected loss function since one deals with a binary classification problem, and the CNNs' outputs are prediction values between 0 and 1.

Table 6.2 summarises the selected set of hyperparameters for each CNN, along with the mean accuracy of the models on the validation set.

Table 6.2: Selected hyperparameters and mean validation performance of each CNN model.

Layer	**Learning Rate**	**Epochs**	**Batch Size**	**Optimiser**	**Loss Function**	**Mean Accuracy on the Validation Set** (%)
RNFL-GCL	0.0001	30	8	RMSProp	Binary Cross-Entropy	92.0 ±1.9
IPL	0.01	40	32			90.7 ±2.7
INL	0.001	30	32			88.2 ±1.5
OPL	0.001	20	32			87.5 ±2.8
ONL	0.001	40	16			85.1 ±1.7
Total Retina	0.01	50	32			87.7 ±4.5

The CNNs were then trained using the chosen hyperparameters. Four additional methods were combined during the learning process: reducing learning rate, early stopping, model checkpoint, and data augmentation. When the evaluation metric had stopped improving after six iterations, the learning rate was reduced by half. Additionally, early stopping was applied to cease training if there was no

improvement over ten epochs. Furthermore, model checkpoints saved the model giving the best performance on the validation set, independently of the finishing epoch. Finally, the number of training examples was increased by adding the following random transformations: rotation by a multiple of 90 degrees (90^{o}, 180^{o}, or 270^{o}) and a random vertical or horizontal flip. These transformations made the models' learning process independent of the images' orientation and invariant to the symmetry between right and left eyes.

Additionally, these were tested on the hold-out test set to evaluate the models' performance on unseen data. The obtained results are shown in Table 6.3.

Table 6.3: Performance metrics obtained on the test set for the final models of each layer.

	Test Set			
Layer	**Accuracy (%)**	**Sensitivity (%)**	**Specificity (%)**	**F1-Score (%)**
RNFL-GCL	89.2	97.6	80.0	90.4
IPL	81.5	96.3	65.3	84.5
INL	86.0	81.7	90.7	85.9
OPL	79.0	74.4	84.0	78.7
ONL	84.7	84.1	85.3	85.2
Total Retina	86.6	86.6	86.7	87.1

All six CNNs performed well on the test set, with accuracies between 79.0% and 89.2%. These findings thus suggest that each particular layer/layer-aggregate has indeed helpful information to distinguish wild-type from transgenic mice.

The CNNs that achieved higher accuracy were those trained using the RNFL-GCL and TR layer-aggregates, whose accuracies were, respectively, 89.2% and 86.6%. These results came without any surprise as both aggregate two or more layers.

Additionally, the classifications per eye and time points were analysed. This allowed determining whether there was a pattern for incorrect classifications on the test set. For example, CNNs being able to classify correctly some time points but performing poorly in other ones.

The results showed that, overall, the distribution of incorrect classifications

between right- and left-eyes were quite balanced for the RNFL-GCL model (41.2% and 58.8%, respectively), IPL model (51.7% and 48.3%), INL model (50% and 50%), OPL model (51.5% and 48.5%), and total retina model (42.9% and 57.1%).

Only the ONL model suggested a substantial difference in accuracy between eyes, with 70.8% of the wrong predictions belonging to the left eye and only 29.2% to the right eye.

Regarding the classification between time points, the distribution of incorrect predictions for each CNN was reasonably balanced between time points. For the RNFL-GCL model, the distribution of incorrect classifications were 11.8%, 29.4%, 17.6%, and 41.2% for one, two, three, and four-month-old, respectively. For the IPL model, these were 17.3%, 20.7%, 31.0%, and 31.0%. Regarding the INL model, the distribution was 31.8%, 13.6%, 36.4%, and 18.2%. Concerning the OPL model, they were 27.3%, 30.3%, 21.2%, and 21.2%. For the ONL model, the percentages were 29.2%, 29.2%, 12.5%, and 29.2%. Finally, the TR model had the following distribution: 28.6%, 38.1%, 14.3%, and 19.0%. Thus, any of the models strove to correctly classify at a specific time point, which is explained by the fact that the six CNNs were trained using images from all time points simultaneously, thereby learning global characteristics independent of a particular time point.

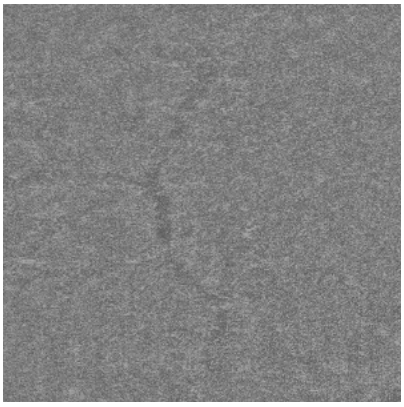
6.1.1 Visual Explanations from Deep Networks

After ascertaining that it was possible to differentiate both mice's groups and demonstrating that all CNNs could attain good performance in the classification task, this study intended to assess which regions in the ocular fundus images conveyed relevant information; if the decisive characteristics are located in a specific structure, such as the vascular network, or diffusely spread throughout the image.

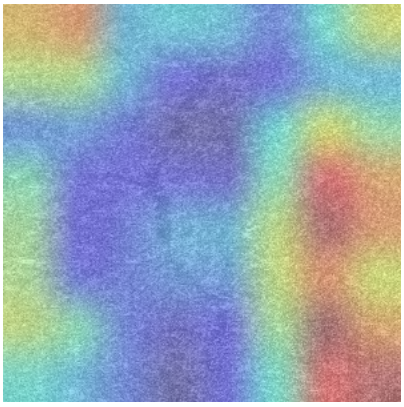
This was achieved by generating heatmaps based on the trained models with heatmaps computed by the Grad-CAM algorithm.

The heatmaps had the size of the convolutional layer picked (the layer *mixed10* of size 8x8) and were resized to the input image size for ease of matching with the input image.

Figure 6.1 shows an example of an ocular fundus image and the respective heatmap. Blueish regions correspond to lower importance areas, while the orange-red areas correspond to those with higher significance to the CNN.



(a) Original Image.

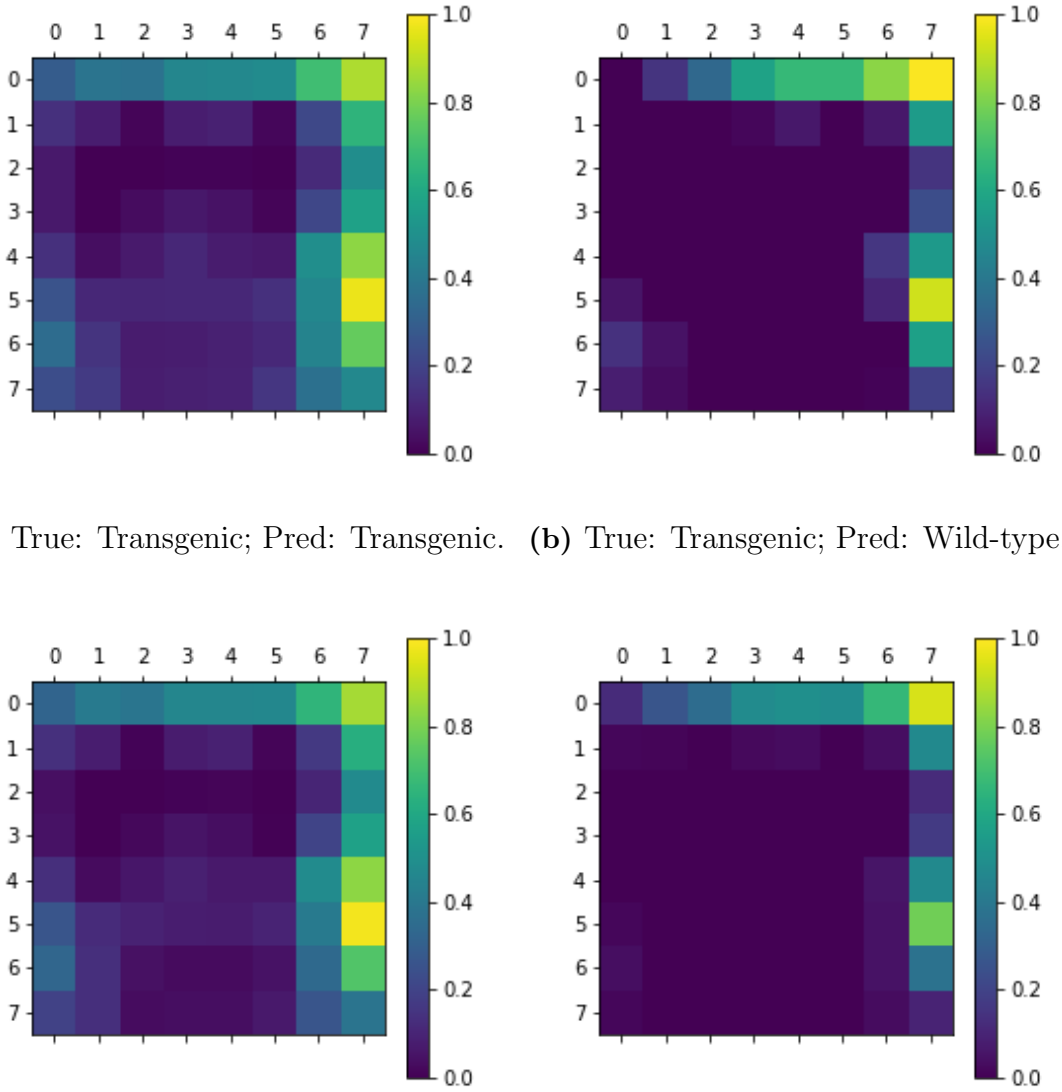


(b) Heatmap superimposed to the input image.

Figure 6.1: Illustration of fundus image and overlaid heatmap. Mouse ID: D_D957_T02_OD_INL

For each layer/layer-aggregate and each class, the Grad-CAM method was applied to locate the areas in the images that have the largest impact on prediction and compute an average heatmap for each pair {true class, predicted class}. Therefore, four average heatmaps were produced for each of the six layers: {true transgenic, predicted transgenic}, {true transgenic, predicted wild type}, {true wild-type, predicted transgenic}, {true wild-type, predicted wild type}.

Figures 6.2 to 6.7 display the average heatmaps obtained for RNFL-GCL, IPL, INL, OPL, ONL, and TR, respectively.



(a) True: Transgenic; Pred: Transgenic. **(b)** True: Transgenic; Pred: Wild-type.

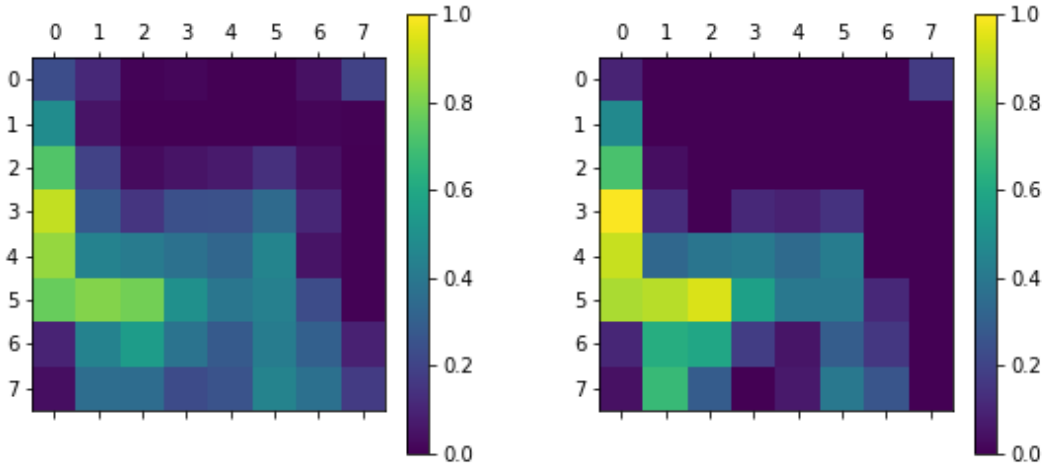
(c) True: Wild-type; Pred: Transgenic. **(d)** True: Wild-type; Pred: Wild-type.

Figure 6.2: Average heatmaps for the different outcomes for the RNFL-GCL layer-aggregate.

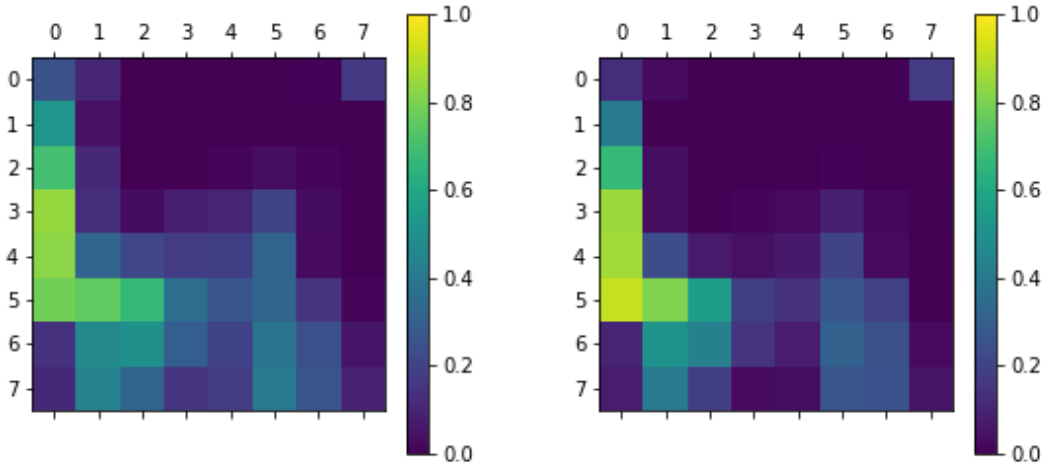
The analysis of the average heatmaps shows that, for the RNFL-GCL model, the areas with the highest impact in the class decision are located on the right side and top of both 'Transgenic' and 'Wild-type' heatmaps, (i.e., average heatmaps corresponding to images classified as transgenic and wild-type, respectively), being the maximum found on the rightmost area.

The main difference found is that the 'Transgenic' heatmaps have a more extensive relevant area than the 'Wild-type' heatmaps, comprising a vertical band located

on the left side and a few in the centre.



(a) True: Transgenic; Pred: Transgenic. **(b)** True: Transgenic; Pred: Wild-type.

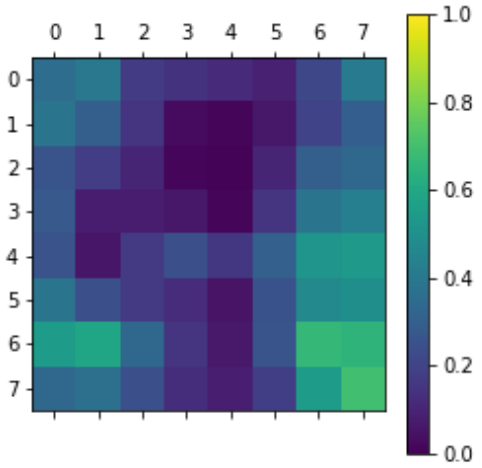


(c) True: Wild-type; Pred: Transgenic. **(d)** True: Wild-type; Pred: Wild-type.

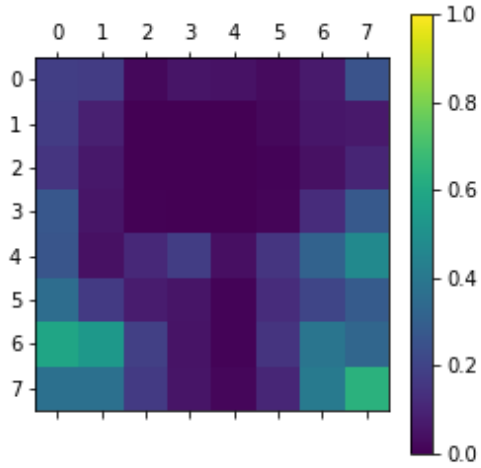
Figure 6.3: Average heatmaps for the different outcomes for the IPL.

For the IPL model, the 'Transgenic' and 'Wild-type' have similar heatmaps. Both cover a vertical band on the left (this having the maximum importance), a relatively small area located on the upper right corner, and a big relevant area going from the bottom to the upper centre of the images.

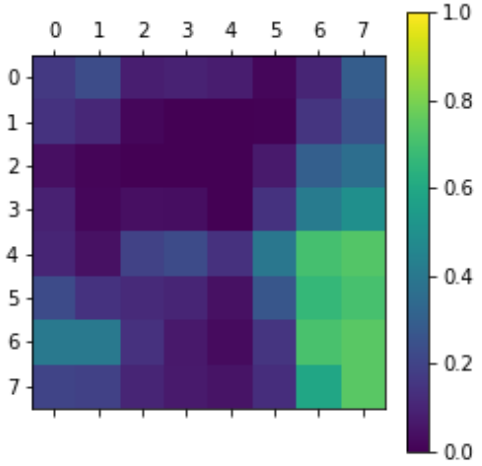
Although the similarity between heatmaps, 'Wild-type' images have a slightly smaller area that impacts the network's output than 'Transgenic' heatmaps.



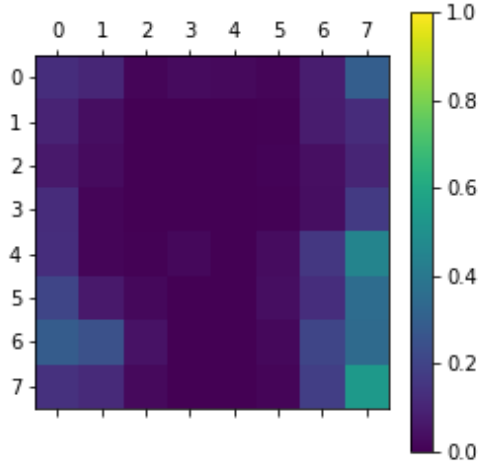
(a) True: Transgenic; Pred: Transgenic.



(b) True: Transgenic; Pred: Wild-type.



(c) True: Wild-type; Pred: Transgenic.

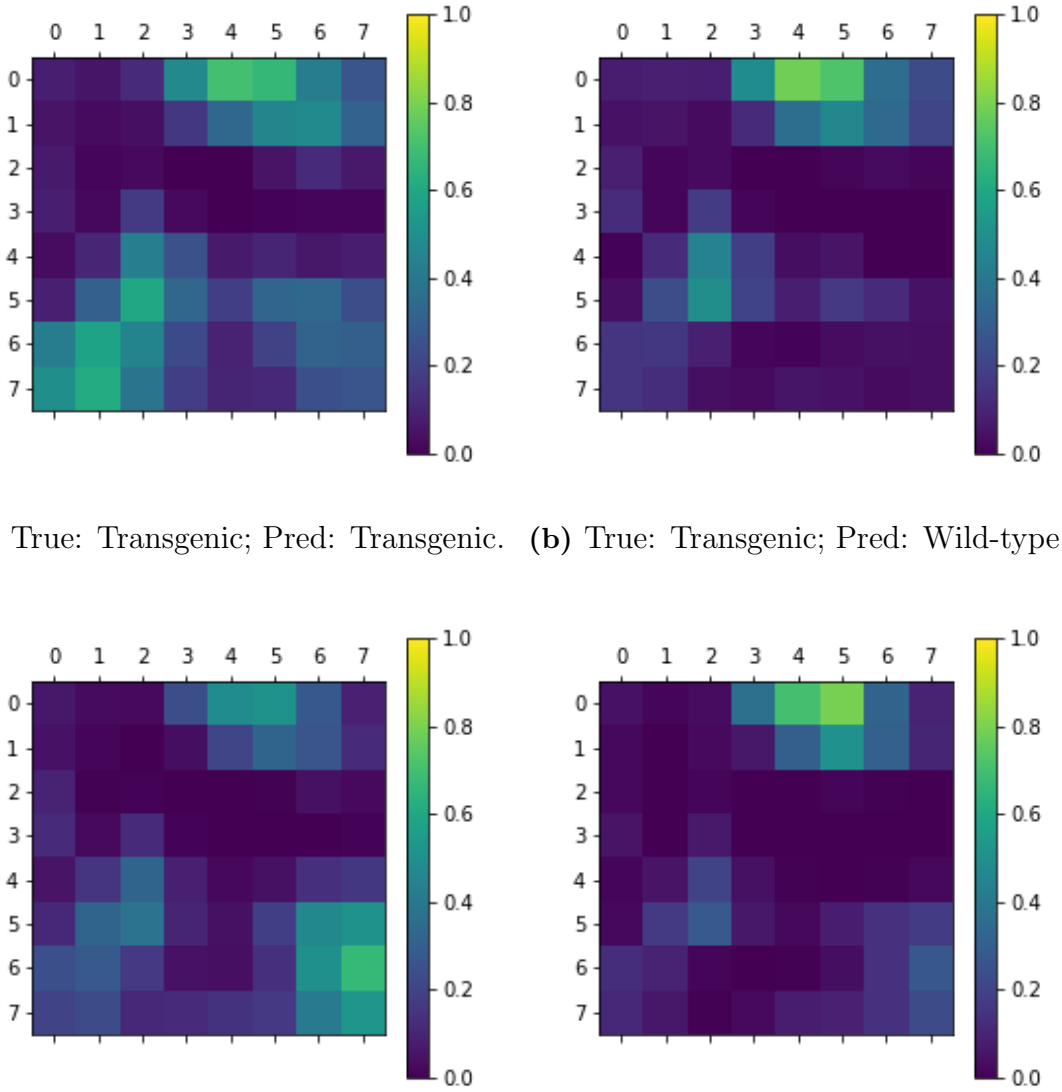


(d) True: Wild-type; Pred: Wild-type.

Figure 6.4: Average heatmaps for the different outcomes for the INL.

The INL 'Wild-type' heatmaps are shown as addressing two vertical approximately symmetrical bands at both sides of the heatmaps. The areas with the highest impact are located at the bottom corners.

The 'Transgenic' heatmaps focus on almost the entire image, except the upper centre. The maximum is located on the right-side band from the bottom to the centre.



(a) True: Transgenic; Pred: Transgenic. **(b)** True: Transgenic; Pred: Wild-type.

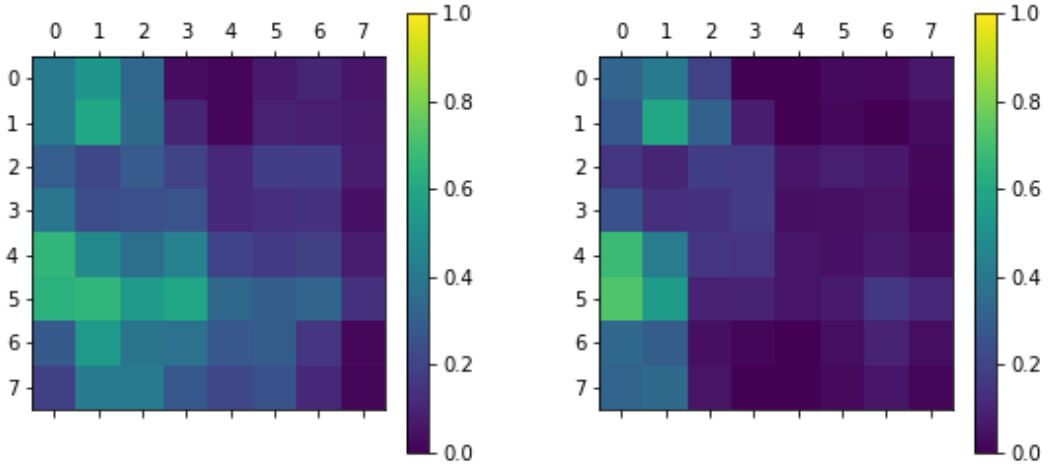
(c) True: Wild-type; Pred: Transgenic. **(d)** True: Wild-type; Pred: Wild-type.

Figure 6.5: Average heatmaps for the different outcomes for the OPL.

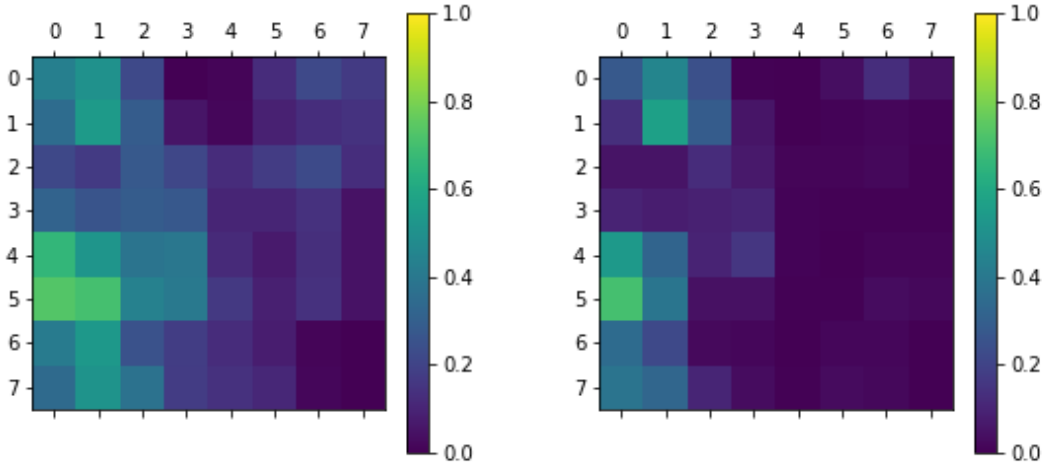
For the OPL model, the 'Transgenic' heatmaps cover mainly two horizontal bands on the top and bottom of the images, leaving only an area on the upper centre that has no relevance. However, if the actual class is transgenic, the areas with higher impact are located at the top and left bottom corner. On the other hand, if the actual class is wild-type, the most relevant area is placed at the right bottom corner.

Regarding the 'Wild-type' heatmaps, the area with the highest impact is a horizontal band at the top, and, to some extent, the left-centre is also relevant for

class prediction.



(a) True: Transgenic; Pred: Transgenic. **(b)** True: Transgenic; Pred: Wild-type.

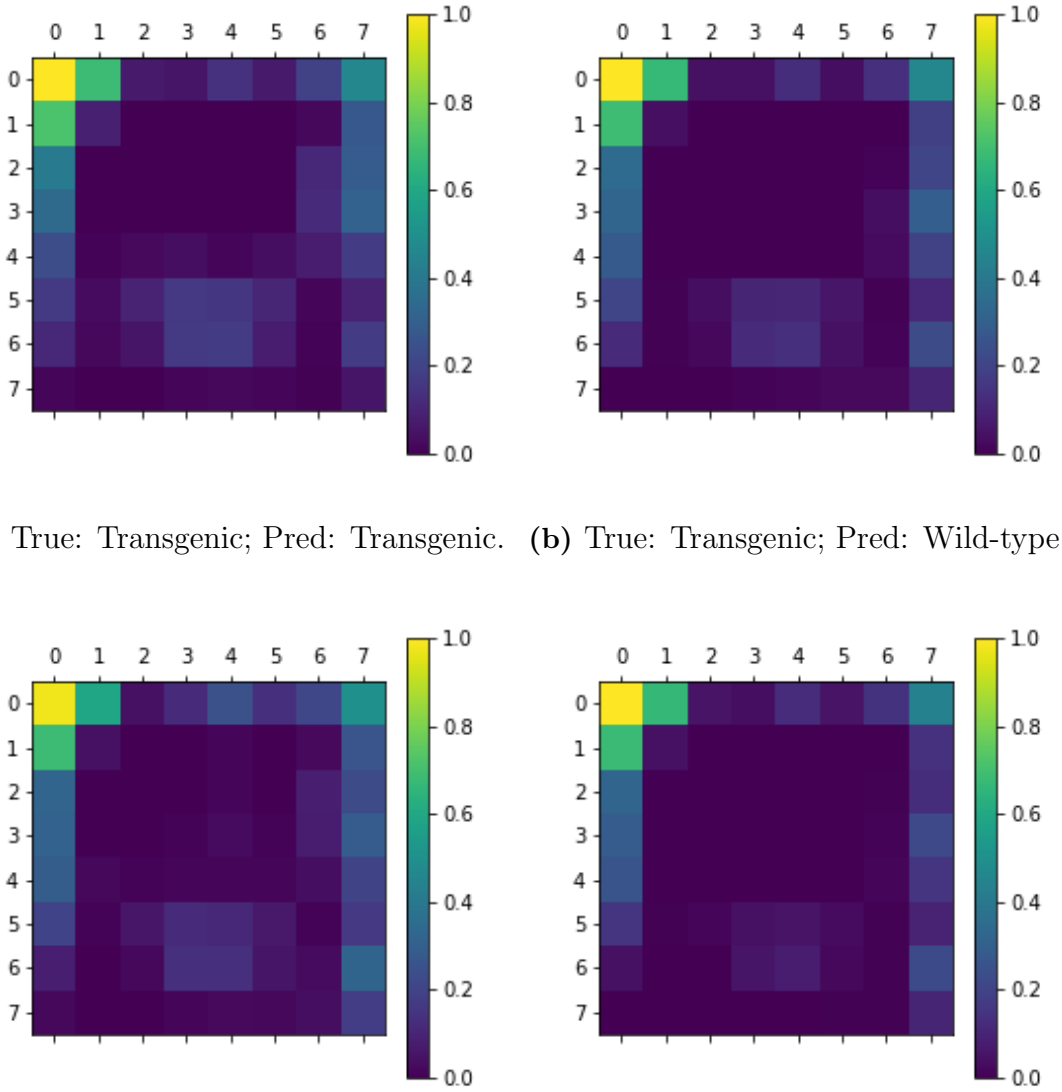


(c) True: Wild-type; Pred: Transgenic. **(d)** True: Wild-type; Pred: Wild-type.

Figure 6.6: Average heatmaps for the different outcomes for the ONL.

The ONL 'Transgenic' heatmaps cover nearly the whole image, having the left side higher impact.

The 'Wild-type' heatmaps also consider the left-side as being the relevant area.



(a) True: Transgenic; Pred: Transgenic. **(b)** True: Transgenic; Pred: Wild-type.

(c) True: Wild-type; Pred: Transgenic. **(d)** True: Wild-type; Pred: Wild-type.

Figure 6.7: Average heatmaps for the different outcomes for the TR layer-aggregate.

For the TR model, the four average heatmaps are pretty similar to each other. They assign importance to a horizontal band placed on the top, vertical bands on the rightmost and leftmost sides, and a band on the bottom centre of the images. The maximum is located around the upper left corner.

Overall, the analysis of all layers/layer-aggregates heatmaps indicates that the images classified as transgenic have a more extensive area that is essential to assign

the classification than those classified as wild-type. Moreover, it was also found that the helpful characteristics are not located in the same places for all retinal layers/layer-aggregates.

6.1.2 Dependence between Eyes

The third goal of this thesis was to check whether the fundus images characteristics that allow distinguishing wild-type from transgenic mice are the same for the right and left eyes. In other words, it is intended to assess whether both eyes convey the same information or not.

The objective was met by building six CNNs, one for each layer/layer-aggregate. Their architecture was the one used in Section 6.1, including the hyperparameters. During training, the networks learned from right-eye images and the performance was evaluated on images from the left eye.

Table 6.4 discloses the performance results for each layer on the test set.

Table 6.4: Performance metrics obtained on the test set for the model training on data from the right eye and evaluated on the left eye data.

	Test Set			
Layer	**Accuracy (%)**	**Sensitivity (%)**	**Specificity (%)**	**F1-Score (%)**
RNFL-GCL	81.9	92.3	69.8	84.6
IPL	53.7	100.0	0.0	69.9
INL	80.9	97.1	62.0	84.5
OPL	77.5	96.2	55.9	82.1
ONL	71.1	96.2	41.9	78.1
Total Retina	53.7	100.0	0.0	69.9

From these results, it is possible to observe that the CNNs for the IPL and TR images performed very poorly, both getting an accuracy of 53.7%. These values indicate that the right eye is quite different from the left eye. As a result, the right-eye characteristics learned by the CNNs were not helpful to classify left-eye fundus images. Indeed, these two CNNs classified all test set images as belonging to the class transgenic, resulting in a sensitivity of 100% and specificity of 0%.

The accuracies of the RNFL-GCL, INL, OPL, and ONL models are moderately better in comparison to the IPL and TR models, ranging from 71.1% to 81.9%. However, the sensitivity values are much higher than the specificity values, revealing that most test images were also classified as transgenic mice.

Therefore, these findings indicate that the right and left eyes characteristics differ, and these differences are not due to the symmetry between eyes since random flips were added to the training set; instead, these may be intrinsic to the development of the eye.

6.2 Combining Convolutional Neural Networks' Predictions

As previously explained, for each mouse, at each time point, six fundus images corresponding to six layers/layer-aggregates (RNFL-GCL, IPL, INL, OPL, ONL, and TR) are computed. The CNNs produce an output between 0 and 1 to each of these images. Then, if the output value is higher than 0.5, the image is assigned to the class transgenic; otherwise, it is classified as wild-type. As a result, each mouse's retina, at each time point, has six prediction values assigned to, corresponding to each layer/layer-aggregate as in table 6.5.

Table 6.5: CNNs' output values for each mouse of the dataset.

Mouse ID	Class	RNFL-GCL	IPL	INL	OPL	ONL	TR
E038_T01_OD	0	0.164	0.030	0.001	0.028	0.025	0.011
E038_T02_OD	0	0.039	0.021	0.001	0.001	0.072	0.060
E038_T02_OS	0	0.064	0.002	0.002	0.021	0.003	0.007
...	...	...	...	...	...	...	...
E289_T01_OD	1	0.726	0.995	0.995	0.954	0.967	0.821
E289_T01_OS	1	0.989	0.999	0.873	0.987	0.835	0.803
E289_T02_OD	1	0.965	0.993	0.954	0.982	0.027	0.791

Thus, an additional objective of this study is to assess whether combining the prediction values from the six layers/layer-aggregates would improve the classification performance and find which layers contributed the most.

This classification task is simple, not requiring a complex deep learning model

with many hidden layers to solve it. Therefore, the classification problem was solved using an FFNN with a single input, hidden, and output layers.

A second hidden layer was added during preliminary tests to analyse whether it would improve further the model performance. However, as the performance on the validation set did not improve, a NN with a single hidden layer was set.

The network architecture consisted of six input neurons - one for each layer -, a single hidden layer followed by a ReLU activation function; and lastly, an output layer with one neuron, followed by a sigmoid activation function. The final classification of each mouse's retina is of the wild-type if the output is below 0.5 and transgenic otherwise.

The learning rate, batch size, number of epochs, and number of neurons in the hidden layer were optimised using a grid-search with five-fold cross-validation. Table 6.6 summarises the set of values tested for each hyperparameter.

Table 6.6: List of values used in the FFNN optimisation.

Hyperparameters	**Grid-search Values**
Learning Rate	{0.1, 0.01, 0.001, 0.0001}
Batch Size	{8, 16, 32, 64}
Number of Epochs	{10, 20, 30, 40, 50}
Hidden Neurons	{1,2,3,..., 20}

The chosen loss function was binary cross-entropy since this is a binary classification problem. Furthermore, Adam was the selected optimiser due to usually achieving good performances on this kind of tasks.

Table 6.7 compiles the selected hyperparameters and the mean accuracy achieved on the validation set.

Table 6.7: Selected hyperparameters and mean validation performance of the FFNN.

Learning Rate	**Epochs**	**Batch Size**	**Number of Hidden Neurons**	**Optimiser**	**Loss Function**	**Mean Accuracy on the Validation Set (%)**
0.001	50	64	5	Adam	Binary Cross-Entropy	97.0 ±1.2

During the learning process of the FFNN, two regularisation methods were combined: early stopping and model checkpoint. After 10 epochs with no improvements on the validation accuracy, the training was ceased. The saved model was the one achieving the best performance on the validation set.

Finally, the performance of the NN was evaluated on unseen data - the hold-out test set. Results can be found in table 6.8.

Table 6.8: Performance metrics obtained on the test set for the final models of each layer.

Test Set			
Accuracy (%)	**Sensitivity (%)**	**Specificity (%)**	**F1-Score (%)**
85.4	97.6	72.0	87.4

Upon comparing the performances achieved on the six CNNs that learned from images of a particular layer and the FFNN, the results show that combining the six prediction values does not improve further the classification performance and that the RNFL-GCL alone provides better performance in all metrics.

It was expected that performance values would be higher for the FFNN classifier since it comprises information from the six CNNs. However, whereas the highest performing CNN (the RNFL-GCL model) achieved a test accuracy of 89.2%, the FFNN only attained an accuracy of 85.4%. The accuracy obtained was similar to the mean accuracy of the CNNs (85.4% and 84.5%, respectively). Also, there is a notable difference between the mean accuracy obtained on the validation set (97.0 $\pm$ 1.2 %) and the test accuracy (85.4%).

A different approach to classify the mice's retinas was also tried. It consisted of a majority-vote classifier that aggregated the predictions of all CNNs and predicted the class with the most votes. Interestingly, this voting classifier achieved an accuracy of 89.2%, higher than the test accuracy obtained with the FFNN (85.4%). Moreover, its accuracy equals the one attained in the highest performing CNN (the RNFL-GCL model).

For this voting classifier, all retinal layers/layer-aggregates have equal importance. However, the following part of the approach assessed which layers contributed the most to the classification task. This was accomplished with the FFNN model that determined the optimal combination of the CNNs' predictions and identified the most important layers/layer-aggregates.

6.2.1 Retinal Layers/Layer-Aggregates Importance

The last objective of the present work was to discover which layers/layer-aggregates contributed the most to the classification. For this purpose, two different methods were used: the connection weights algorithm and the perturbation method (for details, see section 4.7.2).

Here, the perturbation method consisted of setting one of the six layer/layer-aggregate values to 0.5 while retaining all other layers' original values. The amount 0.5 is equidistant between groups, thereby being an ambivalent input. The modified layers/layer-aggregates producing worse accuracy (in comparison to the performance obtained in the original test set) were considered the most contributing to the classification task.

Table 6.9: Comparison between connection weights and perturbation methods results.

(a) Connection Weights Algorithm.

Layer	Importance
INL	1.874
OPL	1.086
IPL	0.661
TR	0.316
ONL	0.297
RNFL_GCL	-0.042

(b) Perturbation Algorithm.

Modified Layer	Accuracy (%)	∇ Accuracy (%)
INL	52.2	-33.1
OPL	77.1	-8.3
TR	83.4	-1.9
ONL	84.7	-0.6
IPL	85.4	0.0
RNFL_GCL	85.4	0.0

Both algorithms indicate that the INL is the most contributing layer/layer-aggregate in the classification task. Indeed, in the connection weights method, the INL has an importance value of 1.874, 0.788 higher than the most important layer after the INL. Regarding the perturbation algorithm, when the INL feature values were changed to 0.5, the obtained accuracy was 52.2%, corresponding to a difference of -33.1% compared to the accuracy attained in the original test set (85.4%). This result suggests that the INL is so important that the FFNN model cannot make predictions accurately, assigning the class nearly randomly.

The OPL is the second most contributing layer according to both algorithms. For the connection weights method, it has an importance of 1.086. Moreover, it achieved an accuracy of 77.1% in the perturbation method, affecting the test accu-

racy by -8.3%. However, the model still obtained a reasonable accuracy.

The other layers/layer-aggregates got different orders of importance between algorithms. The connection weights assigned the following ranking: IPL (0.661), TR (0.316), ONL (0.297), RNFL-GCL (-0.042). Hence, the IPL, TR, and ONL had small positive importance, whereas the RNFL-GCL slightly negatively impacted the classification task.

The perturbation method considered that the TR and the ONL barely affected the final accuracy, with -1.9% and -0.6% differences, respectively. Thus, these two layers had minor importance. Finally, the IPL and the RNFL-GCL obtained the same accuracy as the original test set, implying that these had a residual contribution to the prediction task.

7

Conclusion

The main goal of this study was to develop a deep learning approach to identify the characteristic changes in ocular fundus images of the triple-transgenic mouse model of AD. This way, valuable insights about ongoing changes in the early stages of disease were obtained for each retinal layer/layer-aggregate.

This study began with implementing a preprocessing phase. This step was essential to normalise images across eyes, time points, and mice groups. Hence, potential differences between images caused by the acquisition conditions were attenuated.

In the following stage, six CNNs were created to discriminate wild-type from transgenic fundus images using only a particular layer/layer-aggregate (RNFL-GCL, IPL, INL, OPL, ONL, TR). All six CNNs performed well on unseen data, attaining accuracies between 79.0% and 89.2%. This allowed confirming previous findings claiming that each layer/layer-aggregate presented sufficient information to distinguish between wild-type and transgenic mice.

Furthermore, heatmaps were computed to indicate which image areas contained relevant information for the classification task. These findings suggested that the valuable characteristics in transgenic images were present in a broader area within the image compared to the wild-type images. Moreover, the relevant areas were different between layers, i.e., they were not placed in the same regions for all retinal layers/layer-aggregates.

Six additional CNNs were developed. They learned from right-eye images and were evaluated on images from the left eye. In this experiment, the models performed poorly, revealing that the right eye was different from the left eye; the characteristics learned by the models to classify the right-eye images were not helpful to classify the left-eye ones.

A further step in this study was to build a FFNN to assess whether joining the predictions of all six CNNs would improve the classification performance. Results showed that this was not verified - combining the six predictions led to an accuracy of 85.4% which is similar to the mean accuracy of the CNNs (84.5%).

Finally, although the FFNN did not outperform the CNNs, two algorithms were applied to discover which layers/layer-aggregates contributed the most to this classification task - the connection weights algorithm and the perturbation method. The results demonstrated that the INL and OPL were the most contributing layers, obtaining an importance of 1.874 and 1.086, respectively, in the connection weights algorithm, and an accuracy of 52.2% and 77.1% in the perturbation method.

Regarding the other layers, their order of importance differed between algorithms. The connection weights assigned the following ranking: IPL (0.661), TR (0.316), ONL (0.297), RNFL-GCL (-0.042). Hence, the IPL, TR, and ONL had small positive importance, whereas the RNFL-GCL slightly negatively impacted the classification task.

The perturbation method considered that the TR and the ONL had minor importance, scarcely affecting the final accuracy (with -1.9% and -0.6% differences to the original test set accuracy). Moreover, the IPL and RNFL-GCL had a residual contribution to the prediction task. When the layer's original values were set to 0.5, the accuracy was the same as that obtained on the original test set.

In what concerns future work, to further advance this field of research, it is essential to keep focusing on data collection to increase the amount and diversity of samples gathered. This is crucial to improve the deep learning models generalisation ability on unseen data.

Furthermore, additional efforts should be made to include a broader range of hyperparameters in the grid-search and try different pre-trained CNNs to ascertain whether the performance would improve.

Finally, the CNNs trained on the mice dataset could be used to classify human ocular fundus images. This would be valuable to discover if the characteristics identified by the CNNs would prove helpful to diagnose patients in the early stages of AD.

Bibliography

[1] C. Reitz, C. Brayne, and R. Mayeux, "Epidemiology of Alzheimer disease," *Nature Reviews Neurology*, vol. 7, no. 3, p. 137–152, 2011.

[2] V. Polo, M. Rodrigo, E. Garcia-Martin, S. Otin, J. Larrosa, M. Fuertes, M. Bambo, L. Pablo, and M. Satue, "Visual dysfunction and its correlation with retinal changes in patients with Alzheimer's disease," *Eye*, vol. 31, no. 7, pp. 1034–1041, 2017.

[3] R. Mayeux and Y. Stern, "Epidemiology of Alzheimer Disease," *Cold Spring Harbor Perspectives in Medicine*, vol. 2, no. 8, 2012.

[4] P. J. Snyder, J. Alber, C. Alt, L. J. Bain, B. E. Bouma, F. H. Bouwman, D. C. Debuc, M. C. Campbell, M. C. Carrillo, E. Y. Chew, and et al., "Retinal imaging in Alzheimer's and neurodegenerative diseases," *Alzheimer's Dementia*, vol. 17, no. 1, p. 103–111, 2020.

[5] E. M. Lad, D. Mukherjee, S. S. Stinnett, S. W. Cousins, G. G. Potter, J. R. Burke, S. Farsiu, and H. E. Whitson, "Evaluation of inner retinal layers as biomarkers in mild cognitive impairment to moderate Alzheimer's disease," *Plos One*, vol. 13, no. 2, 2018.

[6] L. P. Cunha, A. L. Almeida, L. V. Costa-Cunha, C. F. Costa, and M. L. Monteiro, "The role of optical coherence tomography in Alzheimer's disease," *International journal of retina and vitreous*, vol. 2, no. 24, 2016.

[7] M. Koronyo-Hamaoui, Y. Koronyo, A. V. Ljubimov, C. A. Miller, M. K. Ko, K. L. Black, M. Schwartz, and D. L. Farkas, "Identification of amyloid plaques in retinas from Alzheimer's patients and noninvasive in vivo optical imaging of retinal plaques in a mouse model," *NeuroImage*, vol. 54, 2011.

[8] R. Bernardes, G. Silva, S. Chiquita, P. Serranho, and A. F. Ambrósio, "Retinal biomarkers of alzheimer's disease: insights from transgenic mouse mod-

els," in *International Conference Image Analysis and Recognition*, pp. 541–550, Springer, 2017.

[9] H. Ferreira, J. Martins, A. Nunes, P. I. Moreira, M. Castelo-Branco, A. F. Ambrósio, P. Serranho, and R. Bernardes, "Characterization of the retinal changes of the 3×Tg-AD mouse model of Alzheimer's disease," *Health and Technology*, vol. 10, no. 4, p. 875–883, 2020.

[10] M. B. Graeber and P. Mehraein, "Reanalysis of the first case of Alzheimer's disease," *European Archives of Psychiatry and Clinical Neurosciences*, vol. 249, no. S3, pp. S10–S13.

[11] R. Dahm, "Alois Alzheimer and the beginnings of research into Alzheimer's disease," *Alzheimer: 100 Years and Beyond Research and Perspectives in Alzheimer's Disease*, p. 37–49.

[12] K. Maurer, S. Volk, and H. Gerbaldo, "Auguste D and Alzheimer's disease," *The lancet*, vol. 349, no. 9064, pp. 1546–1549, 1997.

[13] R. Dahm, "Alzheimer's discovery," *Current Biology*, vol. 16, no. 21, pp. R906–R910, 2006.

[14] M. Prince, A. Wimo, M. Guerchet, G. Ali, Y. Wu, and M. Prina, "World Alzheimer Report 2015 The Global Impact of Dementia An analysis of prevalence, incidence, cost and trends," tech. rep., Alzheimer's Disease International, 57A Great Suffolk Street, London, SE1 0BB, UK, Aug. 2015.

[15] M. Prince, J. Jackson, C. P. Ferri, R. Sousa, E. Albanese, W. S. Ribeiro, and M. Honyashiki, "World Alzheimer Report 2009 Executive Summary," tech. rep., Alzheimer's Disease International, 57A Great Suffolk Street, London, SE1 0BB, UK, 2009.

[16] C. Reitz and R. Mayeux, "Alzheimer disease: Epidemiology, diagnostic criteria, risk factors and biomarkers," *Biochemical Pharmacology*, vol. 88, no. 4, p. 640–651, 2014.

[17] W. Xu, C. Ferrari, and H.-X. Wang, *Epidemiology of Alzheimer's Disease*, pp. 229–358. Inga Zerr, IntechOpen, DOI: 10.5772/54398, Feb 2013.

[18] I. Santana, F. Farinha, S. Freitas, V. Rodrigues, and A. de Carvalho, "Epidemiologia da demência e da doença de Alzheimer em Portugal: Estimativas da prevalência e dos encargos financeiros com a medicação," *Acta Médica Portuguesa*, vol. 28, pp. 182–188, 2015.

[19] C. R. Jack and D. M. Holtzman, "Biomarker Modeling of Alzheimer's Disease," *Neuron*, vol. 80, no. 6, p. 1347–1358, 2013.

[20] C. Bauer, H. Cabral, and R. Killiany, "Multimodal Discrimination between Normal Aging, Mild Cognitive Impairment and Alzheimer's Disease and Prediction of Cognitive Decline," *Diagnostics*, vol. 8, no. 1, p. 14, 2018.

[21] N. Tolboom, M. Yaqub, W. M. V. D. Flier, R. Boellaard, G. Luurtsema, A. D. Windhorst, F. Barkhof, P. Scheltens, A. A. Lammertsma, B. N. V. Berckel, and et al., "Detection of Alzheimer Pathology In Vivo Using Both 11C-PIB and 18F-FDDNP PET," *Journal of Nuclear Medicine*, vol. 50, no. 2, p. 191–197, 2009.

[22] J. Islam and Y. Zhang, "Brain MRI analysis for Alzheimer's disease diagnosis using an ensemble system of deep convolutional neural networks," *Brain Informatics*, vol. 5, no. 2, 2018.

[23] B. Schlosshauer and H. Steuer, "Comparative Anatomy, Physiology and In Vitro Models of the Blood-Brain and Blood-retina Barrier," *Current Medicinal Chemistry-Central Nervous System Agents*, vol. 2, no. 3, p. 175–186, 2002.

[24] J. A. Fernández-Albarral, E. Salobrar-García, R. Martínez-Páramo, A. I. Ramírez, R. D. Hoz, J. M. Ramírez, and J. J. Salazar, "Retinal glial changes in Alzheimer's disease – A review," *Journal of Optometry*, vol. 12, no. 3, p. 198–207, 2019.

[25] E. Salobrar-García, R. de Hoz, A. I. Ramírez, I. López-Cuenca, P. Rojas, R. Vazirani, C. Amarante, R. Yubero, P. Gil, M. D. Pinazo-Durán, *et al.*, "Changes in visual function and retinal structure in the progression of Alzheimer's disease," *PloS one*, vol. 14, no. 8, p. e0220535, 2019.

[26] D. A. Valenti, "Alzheimer's Disease and Glaucoma: Imaging the Biomarkers of Neurodegenerative Disease," *International Journal of Alzheimer's Disease*, vol. 2010, p. 1–9, 2010.

[27] G. Mckhann, D. Drachman, M. Folstein, R. Katzman, D. Price, and E. M. Stadlan, "Clinical diagnosis of Alzheimer's disease: Report of the NINCDS-ADRDA work group* under the auspices of department of health and human services task force on Alzheimer's Disease," *Neurology*, vol. 34, no. 7, p. 939–939, 1984.

[28] K. Blennow, M. J. D. Leon, and H. Zetterberg, "Alzheimer's disease," *The*

Lancet, vol. 368, no. 9533, p. 387–403, 2006.

[29] B. Dubois, H. H. Feldman, C. Jacova, S. T. Dekosky, P. Barberger-Gateau, J. Cummings, A. Delacourte, D. Galasko, S. Gauthier, G. Jicha, and et al., "Research criteria for the diagnosis of Alzheimer's disease: revising the NINCDS–ADRDA criteria," *The Lancet Neurology*, vol. 6, no. 8, p. 734–746, 2007.

[30] "Biomarkers and surrogate endpoints: Preferred definitions and conceptual framework," *Clinical Pharmacology & Therapeutics*, vol. 69, no. 3, pp. 89–95, 2001.

[31] R. Craig-Schapiro, A. M. Fagan, and D. M. Holtzman, "Biomarkers of Alzheimer's disease," *Neurobiology of Disease*, vol. 35, no. 2, p. 128–140, 2009.

[32] M. Tible, Sandelius, K. Höglund, A. Brinkmalm, E. Cognat, J. Dumurgier, H. Zetterberg, J. Hugon, C. Paquet, and K. Blennow, "Dissection of synaptic pathways through the CSF biomarkers for predicting Alzheimer disease," *Neurology*, vol. 95, p. 10.1212/WNL.0000000000010131, 06 2020.

[33] S. C. Rice, "Cerebral Spinal Fluid (CSF) Analysis." Available at: `https://www.healthline.com/health/csf-analysis#risks`, May 2018. Accessed on 15 September 2021.

[34] R. Khoury and E. Ghossoub, "Diagnostic biomarkers of Alzheimer's disease: A state-of-the-art review," *Biomarkers in Neuropsychiatry*, vol. 1, p. 100005, 2019.

[35] H. Hampel, G. Wilcock, S. Andrieu, P. Aisen, K. Blennow, K. Broich, M. Carrillo, N. C. Fox, G. B. Frisoni, M. Isaac, and et al., "Biomarkers for Alzheimer's disease therapeutic trials," *Progress in Neurobiology*, vol. 95, no. 4, p. 579–593, 2011.

[36] M. M. Miller-Thomas, A. L. Sipe, T. L. Benzinger, J. McConathy, S. Connolly, and K. E. Schwetye, "Multimodality review of amyloid-related diseases of the central nervous system," *Radiographics*, vol. 36, no. 4, pp. 1147–1163, 2016.

[37] L. Saint-Aubert, L. Lemoine, K. Chiotis, A. Leuzy, E. Rodriguez-Vieitez, and A. Nordberg, "Tau PET imaging: present and future directions," *Molecular Neurodegeneration*, vol. 12, no. 1, 2017.

[38] S. Minoshima, N. L. Foster, A. A. F. Sima, K. A. Frey, R. L. Albin, and D. E. Kuhl, "Alzheimer's disease versus dementia with Lewy bodies: Cerebral

metabolic distinction with autopsy confirmation," *Annals of Neurology*, vol. 50, no. 3, p. 358–365, 2001.

[39] R. L. Albin, S. Minoshima, C. J. Damato, K. A. Frey, D. A. Kuhl, and A. Sima, "Fluoro-deoxyglucose positron emission tomography in diffuse Lewy body disease," *Neurology*, vol. 47, no. 2, p. 462–466, 1996.

[40] P. Santens, J. D. Bleecker, P. Goethals, K. Strijckmans, I. Lemahieu, G. Slegers, R. Dierckx, and J. D. Reuck, "Differential Regional Cerebral Uptake of 18F-Fluoro-2-Deoxy-D-Glucose in Alzheimer's Disease and Frontotemporal Dementia at Initial Diagnosis," *European Neurology*, vol. 45, no. 1, p. 19–27, 2001.

[41] W. Jagust, "Positron emission tomography and magnetic resonance imaging in the diagnosis and prediction of dementia," *Alzheimer's Dementia*, vol. 2, no. 1, p. 36–42, 2006.

[42] N. Regoli, "9 Advantages and Disadvantages of Pet scans." Available at: `https://connectusfund.org/9-advantages-and-disadvantages-of-pet-scans`, Oct 2017. Accessed on 15 September 2021.

[43] A. Khvostikov, K. Aderghal, A. S. Krylov, G. Catheline, and J. Benois-Pineau, "3D Inception-based CNN with sMRI and MD-DTI data fusion for Alzheimer's Disease diagnostics," *CoRR*, vol. abs/1809.03972, 2018.

[44] F. Zhang, S. Tian, S. Chen, Y. Ma, X. Li, and X. Guo, "Voxel-Based Morphometry: Improving the Diagnosis of Alzheimer's disease Based on an Extreme Learning Machine Method from the ADNI cohort," *Neuroscience*, vol. 414, p. 273–279, 2019.

[45] H. Li, M. Habes, D. A. Wolk, and Y. Fan, "A deep learning model for early prediction of Alzheimer's disease dementia based on hippocampal magnetic resonance imaging data," *Alzheimer's Dementia*, vol. 15, no. 8, p. 1059–1070, 2019.

[46] S. Rebbah, D. Delahaye, S. Puechmorel, F. Nicol, P. Maréchal, and I. Berry, "Support Vector Machine-based classification of Alzheimer's Disease population using a combination of structural MRI biomarkers," in *ISMRM Workshop on Machine Learning Part II -2018*, (Washington, United States), Oct. 2018.

[47] J. Feng, S. Zhang, and L. Chen, "Extracting roi-Based Contourlet Subband Energy Feature from the sMRI Image for Alzheimer's Disease Classifica-

tion," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, p. 1–1, 2021.

[48] T. U. of Edinburgh, "Functional MR (fMRI)." Available at: `https://www.ed.ac.uk/clinical-sciences/edinburgh-imaging/research/themes-and-topics/medical-physics/imaging-techniques/functional-mr-fmri`, Mar 2021. Accessed on 15 September 2021.

[49] B. C. Dickerson, D. H. Salat, D. N. Greve, E. F. Chua, E. Rand-Giovannetti, D. M. Rentz, L. Bertram, K. Mullin, R. E. Tanzi, D. Blacker, and et al., "Increased hippocampal activation in mild cognitive impairment compared to normal aging and AD," *Neurology*, vol. 65, no. 3, p. 404–411, 2005.

[50] X.-a. Bi, Q. Shu, Q. Sun, and Q. Xu, "Random support vector machine cluster analysis of resting-state fMRI in Alzheimer's disease," *PloS one*, vol. 13, no. 3, p. e0194479, 2018.

[51] W. Li, X. Lin, and X. Chen, "Detecting Alzheimer's disease Based on 4D fMRI: An exploration under deep learning framework," *Neurocomputing*, vol. 388, pp. 280–287, 2020.

[52] Z. Wang, Y. Zheng, D. Zhu, A. Bozoki, and T. Li, "Classification of Alzheimer's Disease, Mild Cognitive Impairment and Normal Control Subjects Using Resting-State fMRI Based Network Connectivity Analysis," *IEEE Journal of Translational Engineering in Health and Medicine*, vol. PP, pp. 1–1, 10 2018.

[53] P. Serranho, A. M. Morgado, and R. Bernardes, "Optical Coherence Tomography: A Concept Review," in *Optical Coherence Tomography*, pp. 139–156, 2012.

[54] W. Drexler and J. G. Fujimoto, *Optical coherence tomography: technology and applications* (1st ed.). Berlin, Heidelberg, Germany: Springer, Cham, 2008.

[55] J. S. Duker, N. K. Waheed, and D. Goldman, *Handbook of Retinal OCT: Optical Coherence Tomography E-Book* (1st ed.). Elsevier Saunders, 2013.

[56] A. Güneş, S. Demirci, L. Tök, Ö. Tök, and S. Demirci, "Evaluation of retinal nerve fiber layer thickness in Alzheimer disease usingspectral-domain optical coherence tomography," *Turkish journal of medical sciences*, vol. 45, no. 5, pp. 1094–1097, 2015.

[57] V. Parisi, R. Restuccia, F. Fattapposta, C. Mina, M. Bucci, and F. Pierelli, "Morphological and functional retinal impairment in Alzheimer's disease pa-

tients," *Clinical neurophysiology : official journal of the International Federation of Clinical Neurophysiology*, vol. 112, pp. 1860–7, 10 2001.

[58] E. Marziani, S. Pomati, P. Ramolfo, M. Cigada, A. Giani, C. Mariani, and G. Staurenghi, "Evaluation of Retinal Nerve Fiber Layer and Ganglion Cell Layer Thickness in Alzheimer's disease using spectral-domain optical coherence tomography," *Investigative Opthalmology Visual Science*, vol. 54, no. 9, p. 5953, 2013.

[59] J. A. Pillai, R. Bermel, A. Bonner-Jackson, A. Rae-Grant, H. Fernandez, J. Bena, S. E. Jones, J. P. Ehlers, and J. B. Leverenz, "Retinal nerve fiber layer thinning in Alzheimer's disease: a case–control study in comparison to normal aging, parkinson's disease, and non-alzheimer's dementia," *American Journal of Alzheimer's Disease & Other Dementias®*, vol. 31, no. 5, pp. 430–436, 2016.

[60] F. J. Ascaso, N. Cruz, P. J. Modrego, R. Lopez-Anton, J. Santabárbara, L. F. Pascual, A. Lobo, and J. A. Cristóbal, "Retinal alterations in mild cognitive impairment and Alzheimer's disease: an optical coherence tomography study," *Journal of neurology*, vol. 261, no. 8, pp. 1522–1530, 2014.

[61] F. Berisha, G. Feke, C. Trempe, J. McMeel, and C. Schepens, "Retinal Abnormalities in early Alzheimer's disease," *Investigative ophthalmology visual science*, vol. 48, pp. 2285–9, 06 2007.

[62] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. MIT Press, 2016. `http://www.deeplearningbook.org`.

[63] L. Yu, S. Wang, and K. K. Lai, "Data Preparation in Neural Network Data Analysis," *Foreign-Exchange-Rate Forecasting With Artificial Neural Networks*, pp. 39–62, 2007.

[64] G. Hinton, N. Srivastava, and K. Swersky, "Lecture 6d-a separate, adaptive learning rate for each connection," *Slides of lecture neural networks for machine learning*, vol. 1, pp. 1–31, 2012.

[65] D. P. Kingma and J. Ba, "Adam: A Method for Stochastic Optimization," *arXiv preprint arXiv:1412.6980*, 2017.

[66] Z. N. K. Swati, Q. Zhao, M. Kabir, F. Ali, Z. Ali, S. Ahmed, and J. Lu, "Brain tumor classification for MR images using transfer learning and fine-tuning," *Computerized Medical Imaging and Graphics*, vol. 75, p. 34–46, 2019.

[67] N. Tajbakhsh, J. Y. Shin, S. R. Gurudu, R. T. Hurst, C. B. Kendall, M. B. Gotway, and J. Liang, "Convolutional Neural Networks for Medical Image Analysis: Full Training or Fine Tuning?," *IEEE Transactions on Medical Imaging*, vol. 35, no. 5, p. 1299–1312, 2016.

[68] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Advances in neural information processing systems*, vol. 25, pp. 1097–1105, 2012.

[69] W. Rawat and Z. Wang, "Deep convolutional neural networks for image classification: A comprehensive review," *Neural computation*, vol. 29, no. 9, pp. 2352–2449, 2017.

[70] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.

[71] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[72] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," *CoRR*, vol. abs/1512.03385, 2015.

[73] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the Inception Architecture for Computer Vision," *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[74] C. Szegedy, S. Ioffe, and V. Vanhoucke, "Inception-v4, inception-resnet and the impact of residual connections on learning," *CoRR*, vol. abs/1602.07261, 2016.

[75] M. Raghu, C. Zhang, J. M. Kleinberg, and S. Bengio, "Transfusion: Understanding Transfer Learning with Applications to Medical Imaging," *CoRR*, vol. abs/1902.07208, 2019.

[76] R. Alake, "Deep learning: Understand The Inception Module." Available at: `https://towardsdatascience.com/deep-learning-understand-the-inception-module-56146866e652`, Dec 2020. Accessed on 15 September 2021.

[77] A. Singh, S. Sengupta, and V. Lakshminarayanan, "Explainable Deep Learning Models in Medical Image Analysis," *Journal of Imaging*, vol. 6, no. 6, p. 52, 2020.

[78] 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," *2017 IEEE International Conference on Computer Vision (ICCV)*, 2017.

[79] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, "Learning Deep Features for Discriminative Localization," *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[80] J. D. Olden, M. K. Joy, and R. G. Death, "An accurate comparison of methods for quantifying variable importance in artificial neural networks using simulated data," *Ecological modelling*, vol. 178, no. 3-4, pp. 389–397, 2004.

[81] S. Dutta and P. Sengupta, "Men and mice: relating their ages," *Life sciences*, vol. 152, pp. 244–248, 2016.

[82] I. Ferando, G. C. Faas, and I. Mody, "Diminished KCC2 confounds synapse specificity of LTP during senescence," *Nature Neuroscience*, vol. 19, no. 9, p. 1197–1200, 2016.

[83] P. Guimarães, P. Rodrigues, C. Lobo, S. Leal, J. Figueira, P. Serranho, and R. Bernardes, "Ocular fundus reference images from optical coherence tomography," *Computerized medical imaging and graphics*, vol. 38, no. 5, pp. 381–389, 2014.

[84] J. A. Stark, "Adaptive image contrast enhancement using generalizations of histogram equalization," *IEEE Transactions on image processing*, vol. 9, no. 5, pp. 889–896, 2000.

[85] F. Chollet, "Keras." "Available at: `https://github.com/fchollet/keras`", 2015. Accessed on 15 September 2021.

[86] M. Abadi, P. Barham, J. Chen, Z. Chen, A. Davis, J. Dean, M. Devin, S. Ghemawat, G. Irving, M. Isard, *et al.*, "Tensorflow: A system for large-scale machine learning," in *12th {USENIX} Symposium on Operating Systems Design and Implementation ({OSDI} 16)*, pp. 265–283, 2016.

[87] M. Zanotti, "Transfer Learning in Image Classification: how much training data do we really need?." Available at: https://towardsdatascience.com/transfer-learning-in-image-classification-how-much-training-data, Jun 2020. Accessed on 15 September 2021.

[88] A. G. Roy, S. Conjeti, S. P. K. Karri, D. Sheet, A. Katouzian, C. Wachinger,

and N. Navab, "ReLayNet: retinal layer and fluid segmentation of macular optical coherence tomography using fully convolutional networks," *Biomedical optics express*, vol. 8, no. 8, pp. 3627–3642, 2017.

[89] J. Y. Choi, T. K. Yoo, J. G. Seo, J. Kwak, T. T. Um, and T. H. Rim, "Multi-categorical deep learning neural network to classify retinal images: A pilot study employing small database," *PloS one*, vol. 12, no. 11, p. e0187336, 2017.

[90] F. J. Martinez-Murcia, A. Ortiz, J. Ramírez, J. M. Górriz, and R. Cruz, "Deep residual transfer learning for automatic diagnosis and grading of diabetic retinopathy," *Neurocomputing*, vol. 452, pp. 424–434, 2021.