Università degli Studi di Padova

DIPARTIMENTO DI INGEGNERIA DELL'INFORMAZIONE

Dipartimento di Ingegneria dell'Informazione

Corso di Laurea Magistrale

in Bioingegneria

Cardiovascular risk assessment from retinal images in a diabetic population

Laureanda:

Sara Poletto

Relatore:

Fabio Scarpa

Correlatore:

Emanuele Trucco

Anno accademico 2021-2022

17 Ottobre 2022

University of Dundee

DUNDEE

To my beloved family
who stayed with me during these five years,
supporting and bearing me in every moment.
You made possible this goal.
I love you and I thank you.

Contents

Abstract

This research work has been carried out at the Computer Vision and Image processing department of the University of Dundee in Scotland. We used the retinal images of the GoDARTS dataset, which was born from a project in 1996 that aimed to identify all diabetic patients in the Tayside region. Since diabetic patients may suffer from diabetic retinopathy, they were offered retinal screenings to monitor the retina's health and prevent the disease.

Differently from the fasting glucose test, measuring the glycated haemoglobin offers data about the trend of glycaemia over the last three months. Our idea was to measure the exposure to glycated haemoglobin, which has been calculated as the integral of the curve interpolating all the values of HbA1c for each patient.

This thesis aimed to predict the cumulative glycated haemoglobin (HbA1c) from retinal images with a deep learning technique. Then, we aimed to build a model to classify the risk of cardiovascular disease in the population of interest.

The main experiment concerned the prediction of the cumulative HbA1c for each available image with the Efficient Net B2 neural network. However, the performance obtained on the test set is not the expected one. The range of the predicted values is significantly narrower than the actual range. Moreover, it seems that the error is linearly dependent on the actual measurement of HbA1c.

Nevertheless, we repeated the same experiment on the left-eye and the right-eye datasets and obtained the two average predicted values very similar (303075.2 versus 301050.1 mmol/mol). We concluded that we could use indistinguishably left-eye and right-eye images.

Then we considered the trend of the predicted values for each patient and compared it with the trend of the actual values. We aimed to study if the neural network, on average, was able to capture the average trend of the cumulative HbA1c. Despite the values being quite different, the average predicted trend is increasing as the actual trend.

As the last step, we considered the position of the predicted line with respect to the actual line. We aimed to investigate a potential association between an above-predicted line and a higher risk of mortality, death due to cardiovascular disease or experiencing a MACE. In all our qualitative analyses, we did not find this association but the numerousness of the datasets was very small to draw any conclusion.

For each task of risk assessment, we investigate if males and females were differently distributed and if there were differences in their slope distributions. Thus, we investigated if males have a higher risk of mortality, death due to cardiovascular disease or experiencing a MACE than females and vice versa. We concluded that they have the same probability of experiencing the three evaluated risks.

Acknowledgement

This thesis was made possible thanks to the collaboration between the University of Dundee (Scotland, UK) and the Università degli Studi di Padova (Italy).
The project Erasmus+, published by the Europen Union, made possible my stay in Dundee for the six months needed to develop this work.

The team CVIP (Computer Vision and Image Processing) of the University of Dundee and its members, who are specialised in biomedical image analysis, computer vision, and applied machine learning, have been fundamental for the creation and development of this research work.

A special thanks to Syed Ghouse, who patiently introduced us to Python coding and neural network implementation, and to Dr Huan Wang, who provided the dataset of cumulative glycated haemoglobin. I would like to thank Oluwafemi Samuel for the interesting seminars and for having integrated us into the research group.

An appreciated thanks to Doctor Alex Doney, who has always been available to discuss and share the obtained result and gave us his precious medical point of view and interpretation.

A grateful thanks to our supervisor, professor Emanuele Trucco, who guided and supported us during these months, let us learn as many things as possible and taught us a critical and objective way to carry out biomedical research.

This work was developed in parallel with the research carried out by my colleague, Andrea Quinto, a student at the University of Padua. Our theses are to be considered a single research work and, for this reason, they share parts and complement each other.

Chapter 1

Introduction

1.1 Work Motivation

Nowadays the rate of people in the United Kingdom suffering from type 2 diabetes is significantly increasing. Their total number is estimated to grow to 5.5 million in less than 10 years. Therefore, it is very important to supply doctors and health care systems with powerful instruments that are able to detect diabetes, prevent it and provide useful advice to slow down its development.

The measurement of glycated haemoglobin is widely used to monitor the average glycaemia over the last three months and to classify the diabetic status of a patient. Since the average lifespan of red blood cells is about three months, measuring the glycated haemoglobin gives important information about the level of blood glucose concentration during the last months.

Retinal images and neural networks are a beneficial combo useful to predict some medical outcomes of interest. For example, the age of a patient can be accurately predicted from retinal images using different types of neural networks. Eye images are a non-invasive and non-expensive way to monitor the health of a patient, in particular, his/her cardiovascular system.

The idea at the basis of this research work is to merge together all the methodology just mentioned in order to use retinal images as input of a neural network to predict the exposure of glycated haemoglobin with the aim of assessing the risk of the patient to develop cardiovascular disease.

1.2 Diabetes

One in ten people over 40 lives in the UK with a diagnosis of diabetes, with a total of 3.8 million people and 90% of those with type 2 diabetes. Moreover, almost 1 million more people with type 2 diabetes do not know they have it because they have not been diagnosed, bringing the total number up to 4.7 million. By 2030 the number of diabetic people is predicted to rise to 5.5 million.

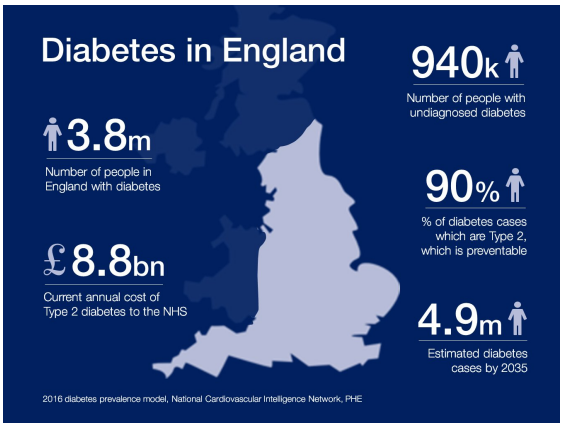


Figure 1.1: Diabetes situation in England (UK).

Diabetes is a pathology caused by a modification of the insulin-glucose control system, and there are two different types:

- **Type 1 diabetes** is an autoimmune pathology in which the beta cells of the pancreas, which are responsible for insulin production, are destroyed by antibodies and cytokines produced by the immune system. A patient who suffers from this type of diabetes must be cured with pharmacological therapy because his organism does not produce insulin anymore. So it can not use glucose to produce the energy necessary for its functioning.

- **Type 2 diabetes** is caused by a combination of a deficit in insulin production and a reduced response to insulin action. The phenomenon is called insulin-resistance and it happens when the organs responsible for controlling glucose concentration become less sensitive to insulin.

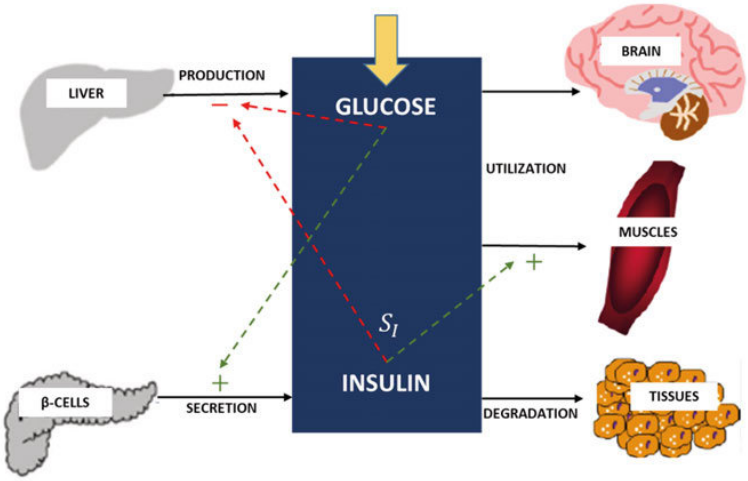


Figure 1.2: Scheme of the glucose metabolism

Glucose metabolism is a fundamental process in a living organism, both from the physiological point of view because it provides the energy necessary for the different vital functions of cells and organs, and from the pathological point of view because its malfunctioning may cause glucose intolerance and diabetes. The most important system in its regulation is the endocrine system, which produces two hormones, insulin and glucagon. In particular, in the pancreas, there are two types of cells: the beta cells that produce insulin and are destroyed in T1D subjects, and alpha cells that produce glucagon. These two hormones are continuously secreted and act with opposing actions to maintain the glucose concentration in a specific range.

The liver has a crucial role in glucose metabolism as a glucose-sensor organ: it can detect its concentration and react with an appropriate secretory response. It can store glucose when its concentration is high and produce and release it in the bloodstream when the organism is deficient.

It is essential for the subject's health that the glucose blood concentration is maintained at a precise interval: the average values of fasting glucose concentration are between 60mg/dL and 110mg/dL, and they may rise to 140mg/dL two hours after an Oral Glucose Tolerance Test (OGTT). When the glycaemia goes under 60mg/dl, the subject experiences a hypoglycaemia episode that is perceived with weakness, headache, sweat and/or trepidation due to the suffering of the central nervous system. In the most severe cases, it may bring hypoglycaemic coma [1].

Hyperglycaemia, on the contrary, happens when the glycaemia is too high. If this condition is maintained for an extended period, it may cause diabetic ketoacidosis and coma caused by dehydration due to a blood accumulation of ketone bodies.

Patients with type 2 diabetes suffer from insulin-resistant which consists of the cell's incapability to use the insulin. They have fewer insulin receptors that make glucose not enter the cells and accumulate in the bloodstream. As the first reaction, the organism produces more insulin to maintain low glucose concentration, and this phenomenon is known as hyperinsulinism. However, in the next stage, the increased insulin production does not maintain glycaemia to normal values and hyperglycaemia happens.

1.3 Glycated haemoglobin

HbA1c refers to glycated haemoglobin, which develops when haemoglobin binds with glucose in the blood, becoming glycated. By measuring it, clinicians can get an overall picture of the average blood sugar levels in the last couple of months. For diabetic patients, this is very important since a higher value of HbA1c means a greater risk of developing diabetes-related complications, like eye and kidney damage, dementia and cardiovascular problems [2].

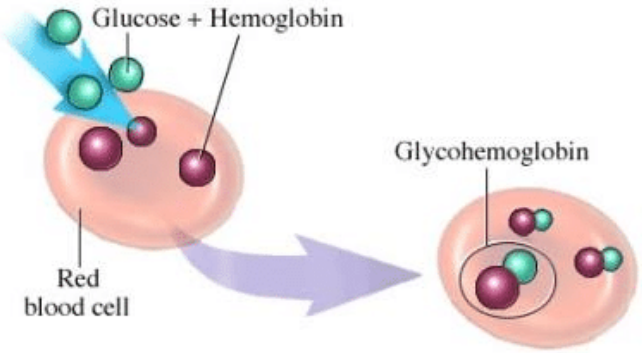


Figure 1.3: Glycated haemoglobin

Haemoglobin is a protein within red blood cells that carries oxygen throughout the body. When the body processes sugar, glucose in the bloodstream naturally

attaches to the haemoglobin. The attached amount is directly proportional to the total amount of sugar in the organism.

Since the lifespan of red blood cells is around three months before renewal, HbA1c measurement reflects the average glucose concentration over that duration, providing a proper longer-term gauge of blood glucose control. On the contrary, fasting glucose and oral glucose tolerance tests only indicate the current concentration and may be biased by the day-to-day variability. Moreover, they need the person to fast and have preceding dietary preparations.

Measuring HbA1c has many advantages. It can be measured at any time of the day and taken from just a finger and it does not require special preparation such as fasting. Furthermore, it is an important instrument for the early identification and treatment of diabetes. For these reasons, it has become an interesting diagnostic test for people with diabetes and a screening test for people at high risk of diabetes. However, it is crucial to consider that HbA1c levels may be affected by some genetic, hematologic and illness-related factors. Some of them are haemoglobinopathies, certain anaemias and disorders associated with accelerated red cell turnover, such as malaria [3].

The average values of HbA1c in healthy individuals are below or equal to 5.6%, a pre-diabetic condition is identified within the range from 5.7% to 6.4%, and a diabetic condition when levels are higher than 6.5%. The target value for diabetic people is 6.5%, which corresponds to 48 mmol/mol, i.e. mmol of glycated haemoglobin per mol of haemoglobin.

1.4 Eye structure

The human eye is responsible for one of our five senses, the vision. When light enters the eye through the pupil, the lens reflects it onto the retina, where messages are encoded thanks to millions of specialised cells, the rods and the cones. These cells transform the image into an electrical message that is in turn sent to the brain through the optic nerve.

The eye is located in orbit, a bone cavity formed by several bones that contains the eyeball, muscles, nerves and blood vessels. The eyeball comprises three chambers:

the anterior, the posterior and the vitreous chamber, divided by three layers: the outer, the middle and the inner layer. The first two chambers are filled with aqueous humour, a watery fluid nourishing interior eye structures. The vitreous chamber is filled with vitreous humour, a thicker transparent gel composed of 99% water that comprises about two-thirds of the eye's volume and helps maintain a round shape. The liquid component is essential because it generates a pressure that keeps the eyeball inflated.

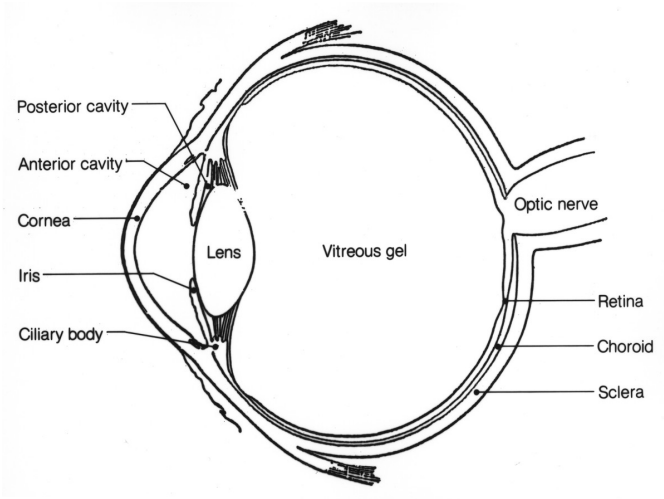


Figure 1.4: Eye structure

The most important elements of the human eye are the followings [4]:

- **cornea**, the transparent curved layer that covers the front of the eyeball, without blood vessels and extremely sensitive to pain. It protects the eye from extraneous and harmful bodies and refracts light onto the lens.

- **crystalline lens**, a transparent structure located directly behind the iris, whose function is to focus light rays onto the retina by changing shape. The lens becomes thicker to focus on nearby objects and thinner to focus on distant ones.

- **iris**, the coloured part of the eye regulates the amount of entering light by controlling the pupil's dilation and constriction, the iris's black centre.

- **choroid**, the middle layer between the retina and the sclera, spongy and filled with blood vessels. Its functions are to absorb the light in excess to prevent blurring of vision and supply oxygen and nutrients to the outer layers of the retina.

- **sclera**, the white and robust part of the eye that forms, together with the cornea, the outer protective coat. It is covered by the conjunctiva, a transparent membrane that protects and lubricates the eye. The function of the sclera is to provide protection and to serve as the attachment for ocular muscles responsible for eye movements.

- **retina**, a light-sensitive layer covering the interior of the eye, whose function is to sense light and create impulses sent through the optic nerve to the brain.

- **macula**, a yellow area at the retina's centre with a diameter of 5.0 mm, which is highly sensitive and responsible for detailed central vision. The centre of the macula is the fovea, a circular area with a diameter of approximately 1.0 mm that contains no rods and the highest concentration of cones and where the focused image is the most accurately registered by the brain.

- **optic disc**, the visible portion of the optic nerve with a diameter of 1.5 mm that identifies the start of the optic nerve. Since it does not have any photoreceptor, it creates a blind spot on the retina. The optic nerve is a structure composed of millions of nerve fibres, each linked to a photoreceptor responsible for connecting the retina to the brain. It transmits the electrical messages from cones and rods to the brain's visual processing centre.

- **cone and rod cells**, the light-sensitive cells or photoreceptors present in the retina. There are between 6 and 7 million cones, thick and short cells divided into three types, each sensitive to the wavelength of a different primary colour. Cones provide acute and detailed central vision and work at best in bright light. On the other hand, the 120 million rods, long cylindrical structures, are responsible for peripheral and night vision and can transmit only shades of grey.

1.5 The retina

Retina architecture

The retina is a part of the central nervous system and is the only visible one. At its centre, as we can see in Fig.1.5, there is the optic nerve, a circular to an oval white area measuring about 2×1.5 mm, across from which the major retina blood vessels branch. It contains the ganglion cell axons running to the brain and the blood vessels that vascularise the retinal layers and neurons.
The fovea is a slightly oval-shaped and blood vessel-free reddish spot which is located around $4.5 - 5$mm to the left of the disc. The central area of the fovea is called the macula.

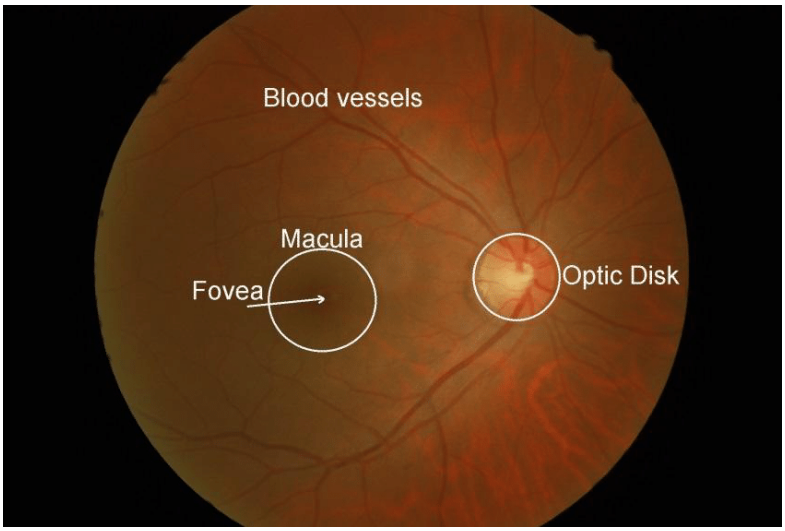


Figure 1.5: Images of the retina architecture.

The retina is divided into two zones: the central retinal, a circular field of ~ 6 mm around the fovea which is the thickest region because of the high concentration of cones, and the peripheral retina, the area outside the macula. The total retina is a circular disc with a diameter between 30 and 40 mm and a thickness of 0.5 mm. The two photosensors, the rods and the cones lie innermost in the retina closest to the lens against the pigment epithelium and choroid.

Human retinas, as well as all vertebrate retinas, are composed of three layers of nerve cell bodies:

- the outer nuclear layer, with cell bodies of the rods and cones,
- the inner nuclear layer that contains cell bodies of the amacrine cells, interneurons that operate in the second synaptic retinal layer-,
- the ganglion cell layer, which contains cell bodies of ganglion cells and displaced amacrine cells.

Moreover, two layers of neuropils, where synapses occur, divide the previous three layers. The neuropil is divided into the outer plexiform layer (OPL) and the inner plexiform layer (IPL). The OPL is a connection layer where rods connect with cones. Instead, the IPL has a carrying-information function between the ganglion cells and the bipolar and amacrine ones. Then the information reaches the brain along the optic nerve.

The two most important sources of blood supply to the retina are:

- the choroidal blood vessels, which receive the greatest blood flow, around $65 - 85\%$, that is vital for the maintenance of the outer retina, including photoreceptors,
- the central retinal artery, that has four main branches -divided in turn into three layers of capillary networks- and nourishes the inner layers with the remaining $20 - 30\%$ of the flow.

When light enters the eye, it must travel through the retina before striking and activating the photosensors that absorb the photons carried by light rays. Their activation is translated firstly into a biochemical message and then into an electrical message that stimulates the retinal neurons and is transmitted to the brain.

Fovea

The fovea is the essential part of the retina for human vision. Protective mechanisms are present to protect the delicate cones and to avoid bright light and ultraviolet irradiation damage. If the cones of the fovea are destroyed, we become incapable of seeing.

The centre of the fovea is called the foveal pit. It is a highly specialised region of the retina where cones are concentrated at maximum density and organised into a hexagonal mosaic, and rods are assent. Below the foveal pit, the other layers are

displaced concentrically and form the foveal slope, a thicker region in both human and monkey retinas. The rim of this area is the parafovea, the thickest portion of the entire retina with ganglion cells organised into six layers.

The foveal area, including the foveal pit, the foveal slope and the parafovea, is considered the macula. The macula lutea has a yellow pigmentation due to the carotenoids present in the axons of the cones. It acts like an additional short-wavelength filter after the lens. It helps improve the achromatic resolution of the foveal cones and blocks harmful UV light irradiation [5].

Degenerative diseases

In some diseases, the retina becomes damaged or compromised, and its imaging can be used as an important diagnostic tool to check the illness's progress. Degenerative changes should be monitored because they may lead to severe damage to the eye, the sight capability and thus the messages sent to the brain.

One of the world's most widespread causes of blindness is age-related macular degeneration. The macula area becomes compromised because the pigment epithelium degenerates and fluid leakages behind the fovea. The cones in this region die, causing an irrecoverable central visual loss.

Another common disease is glaucoma, which happens when the eye pressure is too high and consequently, the eye can not exchange fluids properly because of the elevated pressure. Moreover, the blood vessels and the ganglion cells are damaged.

Diabetic patients may suffer from diabetic retinopathy. It is a side effect of diabetes, which causes the distortion and uncontrollable multiplication of the eye's blood vessels and may cause blindness [5]. In its early stages, there are no symptoms so people may not realise they are developing the disease. Thus screening is important because if the condition is caught early, treatment is effective at reducing or preventing visual impairment and sight loss.

1.6 Retinal imaging

Retina is an important part of the human body that contains essential information about the health of a subject. Retinal images help clinicians in understanding the pathophysiology of a disease and identifying those features that can be used as a diagnostic tool in evaluating the progress of a disease, as well as stratifying patients according to their risk of developing a particular disease in the future and probing the role of the microvasculature in the development of clinical eye and systemic disease.

Retinal image analysis (RIA) has been widely used to study different pathologies, such as diabetic retinopathy, cardiovascular disease and hypertension. The retina must receive a constant supply of blood through a network of tiny blood vessels but, for example, a persistently high blood sugar level -as happens in diabetic patients- can damage its blood vessels and lead to blindness. For these reasons, the study of the retina is a handy tool to identify pathologies early, prevent the development of new diseases, and assess the efficacy of new treatments.

Retinal imaging has developed greatly in the last century and is now a mainstay in treating and screening patients with retinal and systemic diseases. The main imaging techniques are the followings:

- **Fluorescein Angiography**, according to which a fluorescent dye is injected into a vein and fluoresces in the blood, including the retina's blood vessels. When it passes through the retina, photographs record the blood flow and reveal abnormal blood vessels exhibiting hypo or hyper fluorescence and damage. FA is an invasive and time-consuming technique that only records the superficial vascular plexus.

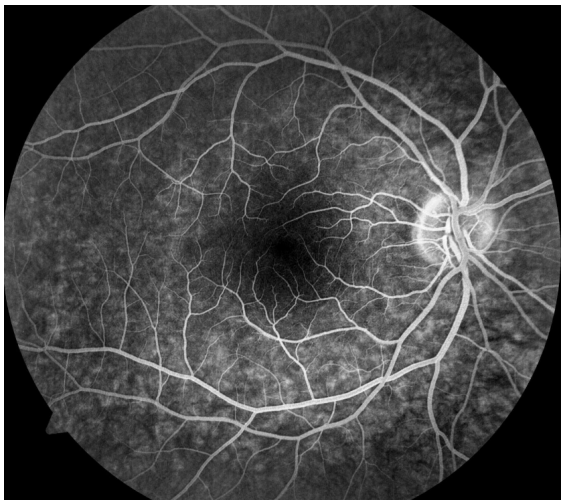
- **Autofluorescence imaging**, based on natural fluorescence, some retinal cells glow without any injected dye when the retina is illuminated with blue light. This fluorescence creates a black-and-white image which can be interpreted by recognising characteristic patterns and used to check the health of the retina layers.

- **Optical Coherence Tomography Angiography**, a non-invasive imagining technology that uses rays of light, that are non-harmful nor dangerous. OCTA records fovea-centred images of the retina blood flow, including the radial peripapillary capillary plexus, the intermediate one and the deep capillary network, features that were not visible with the previous techniques.

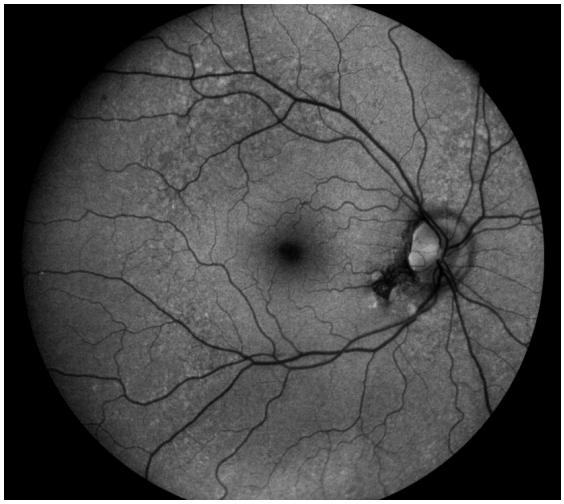
- **Color Fundus Photography**, a non-invasive and fast technique that uses a fundus camera -a specialised low-power microscope with an attached camera- to detect the presence of some diseases and to monitor the retina's health and changes over time. It records colour images of the interior surface of the fundus of the eye; in particular, the central and peripheral retina is visible, together with the optic disc and the macula. It can be used with fluorescein angiography to obtain a better interpretation and more complete results. It requires dilation of the patient's pupil, increasing the angle of observation and the photographed area.

1.7 Document structure

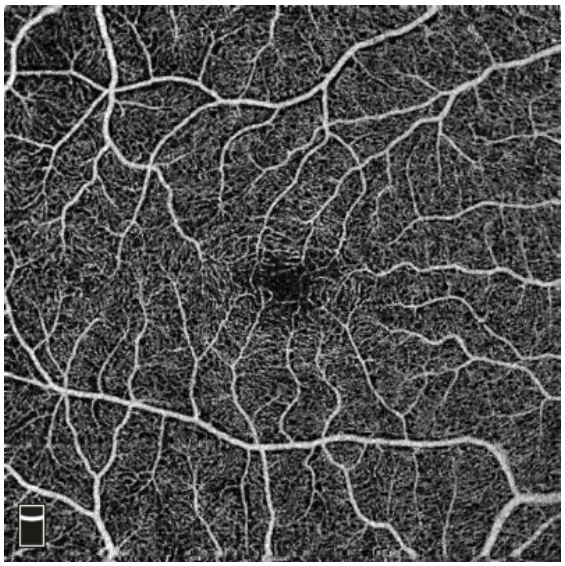
This work is structured into six chapters. After the introduction (Chapter 1), Chapter 2 is about related works, i.e. past research and achievements linked to our main research topic. In Chapter 3 the materials used in this thesis are explained, such as the dataset, retinal images and the Safe Haven environment. In Chapter 4 neural networks are explored in detail, particularly their architecture and their training process. The experiments and results are presented in Chapter 5, while the final chapter contains the conclusions: a brief summary of our work, the main achievements, limitations and future works.



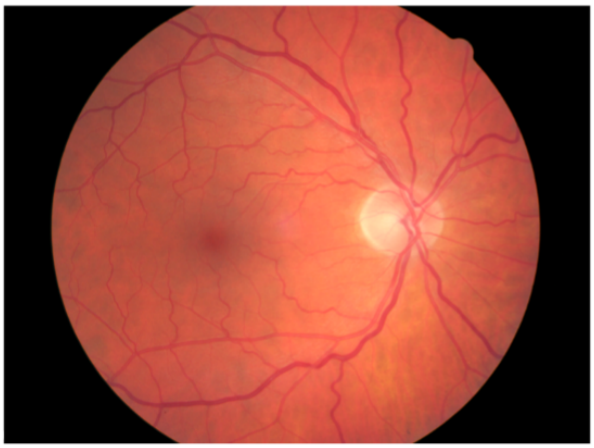
(a) Fluorescein Angiography



(b) Autofluorescence Imaging



(c) Optical Coherence Tomography Angiography



(d) Color Fundus Photography

Figure 1.6: Examples of retinal image techniques.

Chapter 2

Related work

2.1 About this chapter

The purpose of this chapter is to present the main achievements of the last years in applying neural networks to retinal images to predict some biological outcomes. Many recent studies have highlighted the critical role of these images, which provide high-resolution in-vivo images of the internal structure of the eye, particularly of its vasculature, without any invasive or expensive procedure. The human retina is considered a fundamental predictive biomarker for cerebral and systemic vascular diseases because its vasculature is affected by the physiologic and pathologic conditions related to the subject's global vasculature health. Deep neural networks are widely used to extract clinical information from retinal images, such as cardiovascular risk factors and individual characteristics. They are shown to achieve excellent performance in these tasks.

2.2 Age prediction from retinal images

Poplin et al. [6] used a deep learning model to predict for the first time cardiovascular risk factors not previously thought to be present or quantifiable in retinal images, such as age, sex, smoking status and major adverse cardiac event. They used two different datasets to train and test the model, the UK Biobank and the EyePACS, which contain mostly diabetic patients presenting for diabetic

retinopathy screening and with a mean of the glycated haemoglobin HbA1c higher than average values.

They used the Inception-v3 neural network, which predicted age and sex highly accurately, with the MAE for age in the two datasets equal to 3.26 and 3.42 years. It also predicted the HbA1c, blood pressure and BMI but with a low coefficient of determination R^2.

Then they stratified the model's performance by diabetic retinopathy severity, previously assessed by a retinal specialist, and found no significant difference between the groups. Moreover, they trained a model to predict the onset of major adverse cardiovascular events (MACE) within five years in the UK Biobank dataset. The model achieved an area under the receiving operating characteristic curve (AUC) of 0.70. To make a comparison, AUCs were generated using individual factors alone, and the combination of these factors was seen to perform better.

They demonstrated that applying deep learning techniques to fundus images can predict multiple cardiovascular risk factors, such as age and sex because the human retina contains important information about a subject's health.

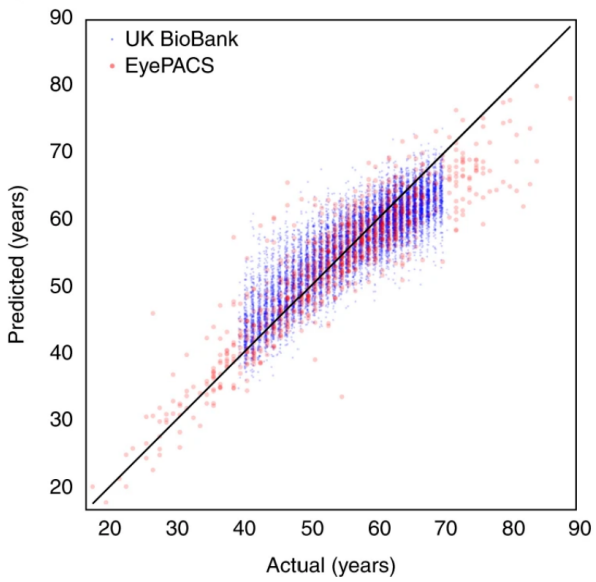


Figure 2.1: Actual vs predicted age in the two validation sets.
The black line represents $y = x$ values.

Two years later *Kim et al.* [7] used a convolutional neural network (CNN) applied to retinal images to predict age and sex in two different groups of participants: normal participants and participants with a systemic vascular-altered status. They used about 412 thousand fundus images from the Seoul National University Bundang Hospital Retinal Image Archive (SBRIA).

The images were pre-processed and normalised to a z-score to ensure the classification results were invariant to intensities and colour contrasts. They used a ResNet-152, the deepest residual network that achieves better performance, with the following structure: the set consists of a convolutional layer, batch normalisation and ReLU activation with 151 layers and a fully connected last layer.
The CNN was made a direct regressor for age prediction by defining the unprocessed numerical output of the network as the predicted age. On the contrary, it was made a classifier for sex prediction by defining the output of the sigmoid function as the probability of predicted sex.

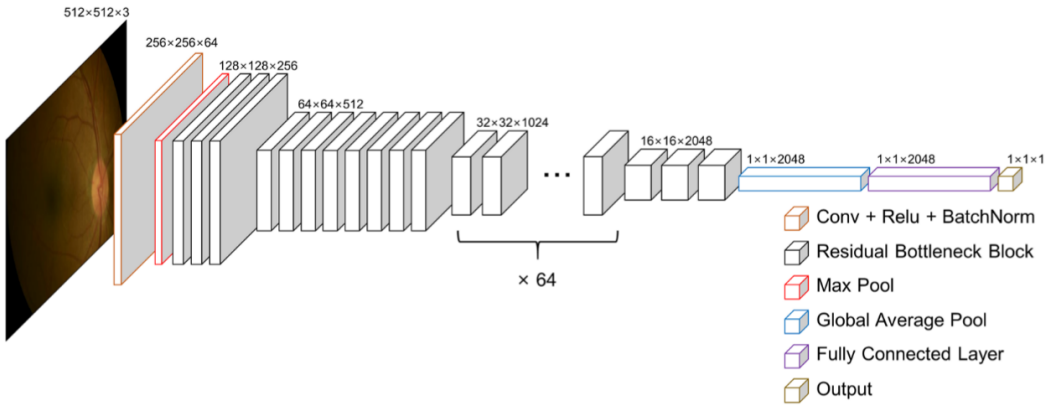


Figure 2.2: Scheme of the convolutional neural network used by *Kim et al.*

The age in the normal population was predicted accurately with an MAE of 3.06 years and a coefficient of correlation between chronologic and predicted age equal to R^2=0.92. The R^2 was lower in the vascular-altered population, suggesting that the ageing process and the pathologic vascular changes affect the retina.
The sex was predicted with an AUC of 0.97 in the normal test set, and similar performance was achieved in the test set of the population with underlying vascular conditions.

Moreover, the effect of retinal blood vessels was studied by predicting age and sex from vessel-erased images. The blood vessel region was extracted using the scale-space approximated CNN. Then an inpainting technique was applied to fill holes by extrapolation from the surrounding background. Age was predicted with an MAE of 3.19 years.

The class activation map (CAM) technique was applied to highlight the core regions where the network focuses on when it predicts age and sex. In the vessel-erased images, the neural network's attention is still focused on where blood vessels are present.

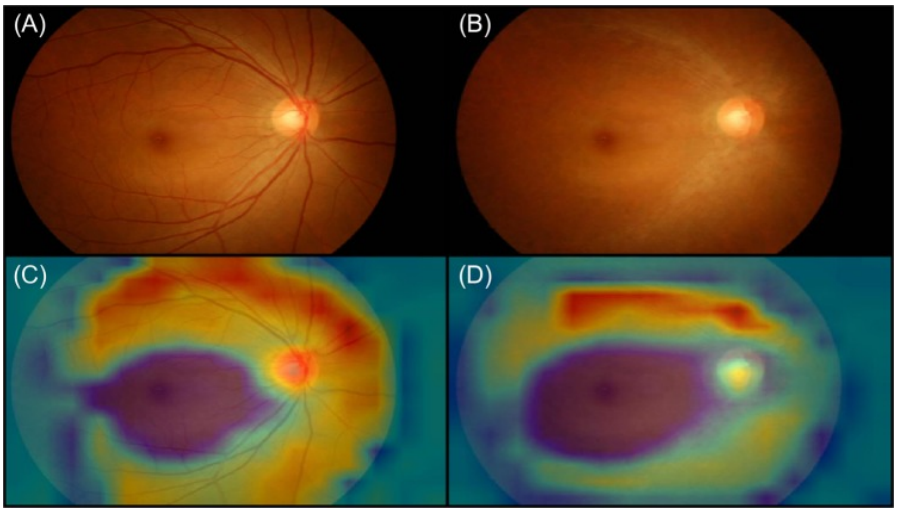


Figure 2.3: CAM heat-maps (C, D) of an original image (A) and a vessel-erased image (B) evaluated in the age prediction model.

The human retina and optic disc continuously change with ageing, and they share pathologic and physiologic characteristics with brain and systemic vascular status. Retina changes are caused by metabolism because cells and tissues are damaged while by-products accumulate in this process. However, the ageing process observed in retinal fundus images may saturate at age 60; age prediction's accuracy is the best in participants aged 20-40 years old, then decreases gradually and deteriorates in subjects older than 60.

2.3 The effects of age and sex

Mortality from cardiovascular disease (CVD) remains the leading cause of death worldwide today, with an increasing burden on the Middle East population. This burden may be reduced by the early identification of individuals developing the disease and by providing the proper lifestyle and medication to alter its course. Therefore, CVD risk stratification instruments are beneficial and retinal images can be one of them to identify subjects at risk. Deep learning can rapidly extract detailed information from fundus images to aid CVD risk determination.

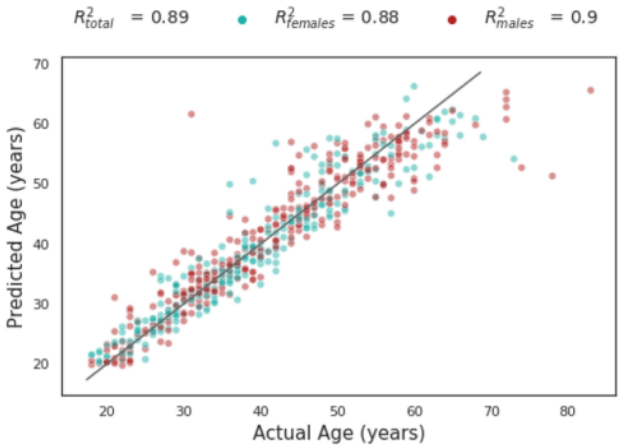
Gerrits et al. [8] investigated the prediction of cardiometabolic risk factors from fundus images and how age and sex, as mediating factors, can affect these predictions. They used retinal images from 3000 Qatari citizens participating in the Qatar Biobank study. The images, initially with a resolution of 1600x1059, were rescaled to 400x400, then filtered through a Gaussian filter and processed with different data augmentation techniques.

To maintain a lightweight and efficient deep learning method, they used a MobileNet-V2 neural network based on depth-wise separable convolutions. A global average pooling layer and two fully connected layers were added to the baseline. The first fully connected layer had 512 neurons and a ReLU activation, while the second was different according to the type of predicted variable. For continuous variables, the second fully connected layer had one output and a linear activation, while for categorical variables, it had one output and a sigmoid activation.

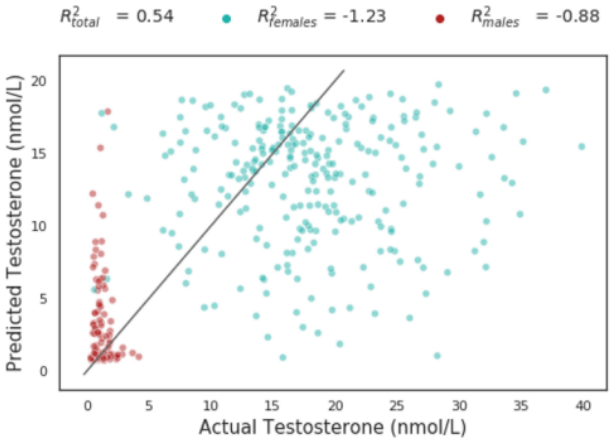
The predictions of age and sex were very accurate using the four images available per person. Age was predicted with a mean absolute error (MAE) of 2.78 years and an R^2 of 0.89, while sex with an area under the curve (AUC) equal to 0.97.

They also evaluated the predictive value of these two variables in predicting a clinical variable by constructing a simple linear regression model with age and sex as independent covariates. This step was done to understand if the trained neural network was reporting inherent correlations among age, sex and the variable being examined. Relatively good performance was achieved in predicting the blood pressure, the HbA1c, relative fat mass and testosterone.

The test set was split according to sex and to age into four subgroups with an almost equal number of individuals in each subgroup. An example of this experiment can be seen in Fig.2.4. They demonstrated that when the neural network predicts testosterone, it indirectly predicts sex. On the contrary, there are no differences between males and females for age and HbA1c predictions.
Furthermore, there are no differences in the model's performances in predicting sex in the groups divided by age, but there are in predicting the HbA1c and systolic blood pressure. They supposed that age influences the model's performance for testosterone and relative fat mass predictions.



(a) Age: actual vs predicted.



(b) Testosterone: actual vs predicted.

Figure 2.4: Age and testosterone predictions in males and females.

These results suggest that when predicting a cardiovascular risk factor other than age and sex, the model pays attention to the same characteristics already explaining age and sex. They demonstrated that the retina stores unique information about an individual's health status, comprehending information related to blood pressure and HbA1c.

2.4 An application to GoDARTS database

Accelerate ageing can be detected from the vascular systems using molecular and cellular biomarkers and functional and structural ones. An important biomarker in the human body is the brain, a highly vascular organ whose images contain recognised shreds of evidence of age-related tissue health, such as manifestations of white matter disease and other age-related structural changes. The retina is embryologically derived from the brain and is a highly vascularised neurological tissue, but, unlike the brain, it can be imaged quickly and inexpensively with digital photography.

Deep learning applied to retinal images has recently been shown to accurately predict a subject's age. Moreover, the difference between the chronological age and the predicted one can be exploited and used as an important biomarker of the subject's health. An individual with a predicted age greater than their chronological age is supposed to have a more significant risk of all-cause death.

Ghouse et al. [9] applied to retinal images a neural network to predict biological vascular age in order to investigate how the difference between chronological and retinal vascular predicted age (predicted age difference, PAD) was associated with major adverse cardiovascular events (MACE) and all-cause death in a large population of individuals with Type 2 Diabetes.

GoDARTS dataset was used, selecting patients with a MACE at the date of the earliest available image but no history of hospitalisation. Images were pre-processed to reduce the effect of image variations, such as brightness, colour and focus. They were resized to the standard size of 260x260 pixels, which is the one recommended to improve accuracy in the used neural network [10].

The EfficientNet-B2 network was used since it achieves excellent performance in imaging tasks. The fully connected layer was replaced with a global average pooling layer followed by a single output node with linear activation. Grad-CAM heatmaps, applied to the last convolutional layer, showed that the network identified the macula and the optic disc as the most important features to predict age.

Results are encouraging because the MAE in predicting age was 3.96 years for the whole cohort with an R2 equal to 0.798. Despite the limitations of the work, it can be considered an important achievement in understanding how retinal images can be used in deep learning to predict a biological outcome.

2.5 Contribution of the University of Dundee

The VAMPIRE team of the University of Dundee investigated how the combination of clinical, retinal and genomic features can successfully stratify the cardiovascular risk in patients with type-2 diabetes selected from GoDARTS database [11]. The performance suggested that multimodal features could capture essential knowledge for MACE risk assessment, and bootstrap analysis was performed to assess the robustness of all three sources of features.

There were 157 retinal features extracted semi-automatically from retinal images with the VAMPIRE 3.1 software of the University of Dundee and Edinburgh and 343 single nucleotide polymorphisms (SNPs). A total of 519 features were used as independent variables in a predictive model developed using L1-regularised logistic regression to predict the binary outcome of MACE onset before censoring. The output probability was used to stratify patients into two groups, high-risk and low-risk. They chose Lasso regression because it performs simultaneous feature selection and model estimation.

The λ parameter, which controls the strength of regularisation, was tuned using 10-fold cross-validation and the model corresponding to the lowest deviance λ was selected (Fig 2.5). It contained 51 features from all three categories. On the contrary, the model corresponding to λ_{1SE}, defined by only seven predictors, contained only clinical features and one gene score and it achieved comparable performance. The features selected in the two models are represented in Tab 2.1.

Category	**Using $\lambda_{\mathbf{min}}$**	**Using $\lambda_{\mathbf{1SE}}$**
Retinal	7 features selected	None selected
SNPs	34 features selected	None selected
Gene scores	CVD gene scores	CVD gene scores
Clinical	• Age at imaging • Blood pressure lowering drugs taken • History of smoking • Evidence of CVD before imaging • Diastolic blood pressure • High density lipoprotein • Glycated Haemoglobin • Triglycerides • Duration of diabetes	• Age at imaging • Blood pressure lowering drugs taken • History of smoking • Evidence of CVD before imaging • Diastolic blood pressure • High density lipoprotein

Table 2.1: Features selected in the models corresponding to λ_{min} and to λ_{1SE}.

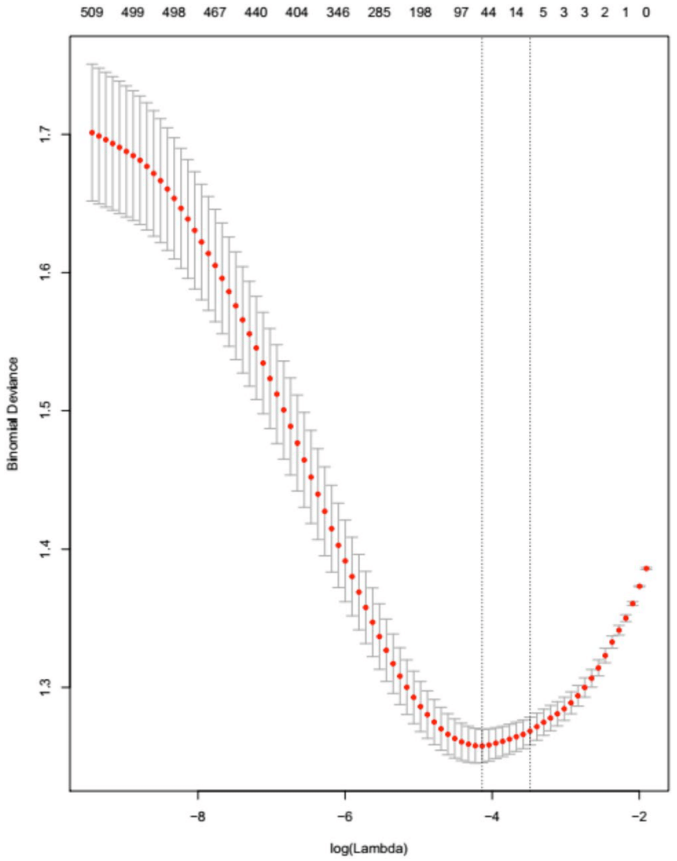


Figure 2.5: Results of the cross-validation, showing how changing λ affects binomial deviance. The vertical line on the left represents λ_{min}, the one on the right λ_{1SE}.

Bootstrap analysis was performed to assess the robustness of the feature selection with Lasso regression over variations of the training set. A total of 500 iterations were completed on the model corresponding to λ_{min}. A significant result is that the variables *age at imaging* was selected in every bootstrap iteration. Other features selected with a frequency higher than 75% were clinical features, such as the history of CVD, history of smoking and glycated haemoglobin and some genetic features. No individual retinal measurements were selected at high frequency because they can be highly correlated.

A former PDRA of the University of Dundee, Boyle Liam, started with the basis of this work and studied whether retinal and genomic measurements can stratify cardiovascular risk in the GoDARTS dataset's population. The focus was on major adverse cardiovascular events (MACE), defined as nonfatal stroke, nonfatal myocardial infarction and cardiovascular death. His dataset contained clinical markers, such as age, sex and haemoglobin levels, retinal markers extracted from retinal images through the software VAMPIRE of the University of Dundee and genetic markers.

Features	Number of Occurences (All)	Percentage of Occurences	Key
e_age	500	100%	Clinical
pre	500	100%	Genomic
dbp	500	100%	Retinal
gh	495	99%	Gene Score
cvd_time	491	98%	
rs2493292	485	97%	
rs409558	469	94%	
hdl	464	93%	
rs79089478	459	92%	
rs4308	459	92%	
rs10953541	457	91%	
ther	455	91%	
rs12941318	440	88%	
rs4511593	430	86%	
eversmoker	427	85%	
rs8258	422	84%	
rs6712094	409	82%	
rs2291435	408	82%	
rs10943605	408	82%	

Figure 2.6: Features selected in the bootstrap analysis made on the dataset containing all the features.

He performed five experiments with different sets of features and the main results he achieved are the followings:

- *age at imaging* was selected 100%, together with the glycated haemoglobin (gh) and the history of a previous CVD (pre), among all the clinical features,
- in the model with clinical and retinal features, clinical ones are still predominantly, and the first retinal one was selected 83%,
- in the model with clinical, retinal and genomic features and the model with all the features, the most often selected are clinical and genomic.

2.6 Conclusion

In this chapter, some important achievements in applying deep learning techniques to retinal images have been presented. They highlight the importance of retinal images, which represent a non-invasive and non-expensive diagnostic tool useful to understand better a subject's health, particularly his cardiovascular health.

Retina blood vessels are affected by systemic diseases, such as diabetes and glycated haemoglobin, as well as hypertension and smoking conditions, and its vasculature can be studied to prevent some diseases and stratify patients.
Retinal images can be used efficiently in neural networks to predict some important individuals characteristic. For example, age can be predicted with an MAE of only three years.

Moreover, it has been shown that retinal features are less important than clinical features in assessing cardiovascular risk. The features selected most often in bootstrap analysis with Lasso regression are the age at imaging, i.e. the age of the patient when the image was taken, the level of glycated haemoglobin and the history of a previous cardiovascular event.

Chapter 3

Materials

3.1 About this chapter

The purpose of this chapter is to present and describe the dataset, the images and the environment used in our experiments. We used the GoDARTS dataset, which contains more than 10 thousand patients that suffer from type 2 diabetes and live in the Tayside region, a region in east Scotland that includes the cities of Dundee and Perth. Since the disease of these subjects, they were offered retinal screening to prevent, diagnose and monitor diabetic retinopathy. More than 102 thousand fundus photography images of the retina were collected and are available for research purposes. Using images and clinical data is made possible through Safe Haven's remote-access environment. It has been implemented by the Health Informatics Centre (HIC) of the University of Dundee to protect data confidentiality, satisfy public concerns about data loss and reassure data controllers about HIC's secure management and processing of their data.

3.2 GoDARTS

The DARTS - Diabetes Audit and Research in Tayside Scotland - study started in 1996 as a collaboration between the University of Dundee, three Tayside Health Care Trusts: the Ninewells Hospital and Medical School, the Perth Royal Infirmatory, the Stracathro Hospital and a group of Tayside general practitioners to identify all

diabetic patients within the Tayside region and to improve health care [12]. The collected data, including hospital diabetes clinics, diabetes prescription database and all diabetes-related records, is continually updated by a dedicated team of clinicians and forms an important longitudinal dataset of clinical data.

In 1998 genetic data started to be collected through a blood sample for DNA extraction, together with phenotypic data through lifestyle questionnaires and clinical examination. This more comprehensive study, called GoDarts – Genetic of DARTS – aims to study and identify if there are correlations between specific genetic and environmental factors and the disease onset, its progression and response to treatment.

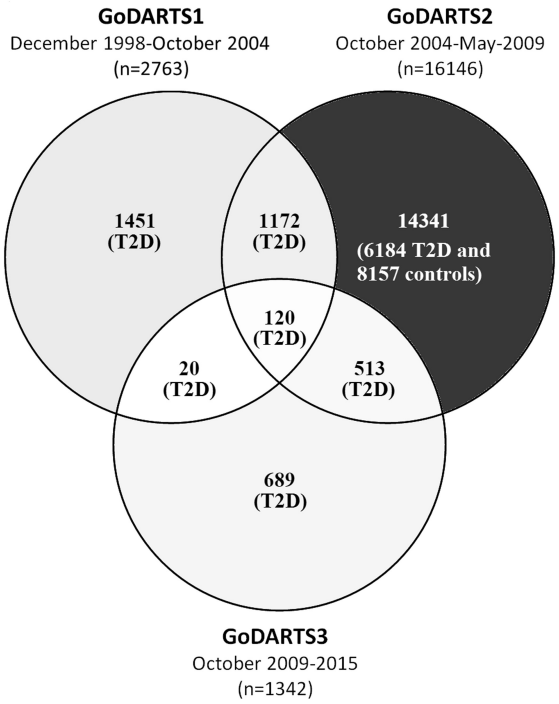


Figure 3.1: Venn diagram showing the overlap in patient recruitment.

The GoDARTS study has been created in three different phases:

(i) The first one, GoDARTS1, was the pilot phase to test the recruitment processes and ability to anonymously link patient clinical data from electronic records to the study. In this phase, only blood samples were taken when the patient was recruited and no baseline data were recorded.

(ii) In the second phase, GoDARTS2, two groups of patients were recruited, one with type 2 diabetes patients and one control group, with on average one control patient per case of diabetes. Baseline clinical and lifestyle measurements were recorded for all patients, for example, the smoking history, level of physical activity and menopause history for women, as well as height, weight, blood pressure and heart rate.

(iii) During the third phase, GoDARTS3, other patients were recruited, and urine and blood samples for RNA extraction were collected. Some of them also participated in phase one or two, where baseline data was missed or the quality of the extracted DNA was poor, and only 1451 patients involved only in the GoDARTS1 do not have these data.

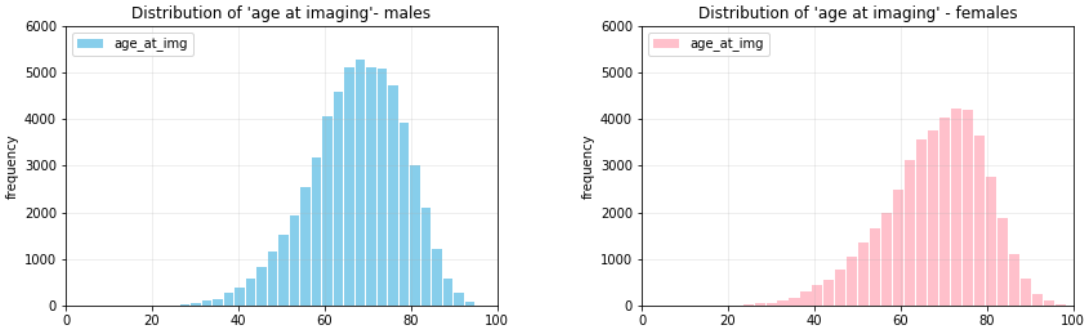
GoDARTS dataset contains a total of 18306 participants, 10149 with type 2 diabetes and 8157 healthy controls at baseline. All participants are asked to provide informed consent for their data to be used for research purposes and explicit consent of use in collaboration with the industry. This way, it is possible to access longitudinal data relating to routine diabetes management.

The linkage between different databases is made possible thanks to the community health index (CHI), a 10-digit unique numerical identifier issued to each patient on the first registration with a GP or on the first admission to a Scotland hospital. This index is then converted into a study pro-CHI, patients' identity is protected, but the linkage across multiple datasets is still possible.

The main strengths of GoDARTS are its large size, the availability of genetic and phenotypic data, the ability to link patients' data to routine electronic medical records and the consent to use these data for research purposes and to contact for possible future research participation. However, there are some weaknesses, like the missing baseline data for some GoDARTS1 patients and the lifestyle questionnaires; since they are self-completed, they may have some bias.

In our experiments we used a subset of GoDARTS, which contains 8750 T2D subjects, of which 3751 are females and 4819 males. A total of 102 082 retinal images are available. The age of the patient when an image is taken is recorded in the variable *'age_at_img'*, whose average of the entire dataset is 67.64 years.

In Fig.3.2 we see its distribution in males and females. They are very similar: females distribution has a mode equal to 69.86 years and a mean to 67.96 years with a skewness coefficient of -0.42, males distribution has a mode equal to 71.31 years and a mean to 67.39 years with a skewness coefficient of -0.57.



(a) Distribution in males

(b) Distribution in females

Figure 3.2: Age at imaging distribution.

The aim of our experiments is predicting the cumulative glycated haemoglobin from retinal images, thus we are interested in understanding how this variable is distributed in our population. As we can see, it has a skewed distribution with a skewness coefficient equal to 2.43.

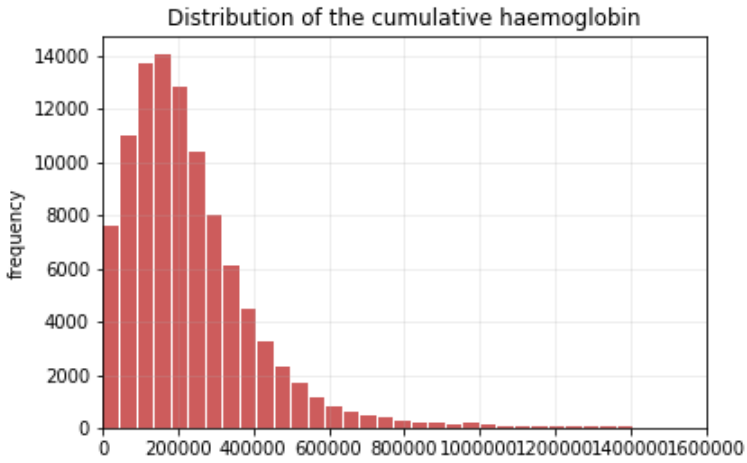
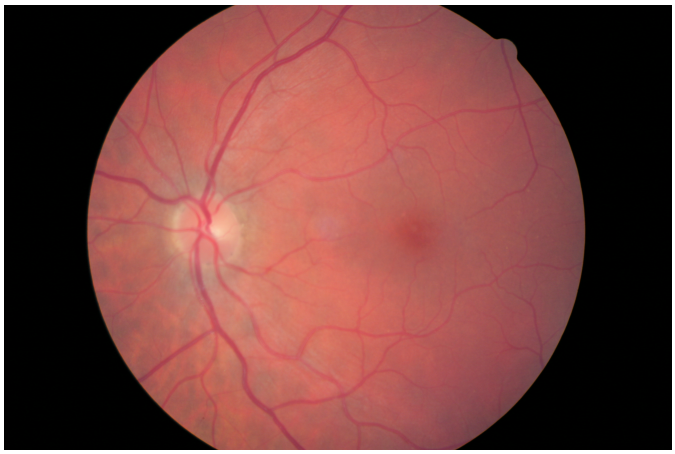


Figure 3.3: Distribution of the cumulative glycated haemoglobin.

3.3 Retinal images

All patients with diabetes in Scotland are offered annual retinal screening with digital retinal photography. Since 2006 these images have been stored centrally by the Scottish National Diabetes Eye Screening (DES) service. The capture of retinal image follows the standard Scottish diabetes retinal screening protocol, which includes a 45° field of view and macula-centred [9].
The dataset contains 102,082 retinal images taken at multiple time points for approximately ten years starting from 2006. We used these images in our experiments to train and test our neural network.



(a)

(b)

Figure 3.4: Examples of GoDARTS fundus images.

Images were pre-processed to reduce the effect of image variations, such as brightness, colour, focus and overall quality. Firstly, the excess black regions of the images were discarded. Images were resized to the standard size of 260×260 pixels, which is recommended to obtain optimal performance when using neural networks on images. Then, they were equalised by contrast-limited adaptive histogram equalisation on the R, G, and B colour channels separately. Finally, pixel intensities were normalised to [0, 1]. An example of this process can be seen in Fig. 3.5.

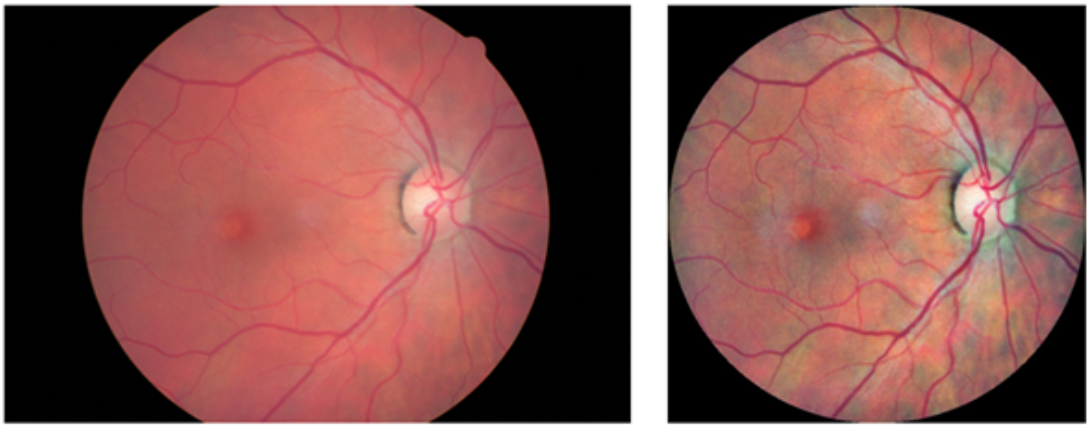


Figure 3.5: Pre-processing: original image (left) and processed image (right).

3.4 Safe Haven

The Safe Haven environment is a web-accessible Virtual Desktop Environment that allows secure remote access to research data provided by Health Informatics Centre (HIC) service and it is based on the VMWare View Horizon VDI technology.

To ensure the safe use of research data, some restrictions are imposed. Within the Safe Haven environment, no internet access is available, applications' installation is disabled and only applications installed by HIC service are available. Data are not released externally to data users for analysis on their computers but placed on a server at HIC, within a restricted, secure IT environment, where the data users are given secure remote access to carry out their analysis. Copying research data supplied by HIC out of the environment is not permitted; however, analysis results, such as reports, summaries and graphs, that do not contain patient-level data are

allowed to be removed and exported after a HIC Data Analyst has reviewed them.

To have access Safe Haven environment, participation in the "Good research practise: principles and guidelines" course held by the Medical Research Council (MRC) is required [13]. The MRC is dedicated to improving human health through excellent medical research and expects that all MRC-funded research is conducted according to the highest possible standards of research practice to ensure the integrity, clarity and good management of the research and outputs. Achieving these ethical and quality standards depends on the integrity, honesty and professionalism of all individuals involved in the research process. Thus, promoting and delivering good research practice is fundamental. Fostering a culture that supports good research practice and aims to prevent research misconduct is a duty for research organisations.

Good research practice provides solid foundations for a research career, supporting high-quality education and training, it delivers assurance to those work builds on the findings of others, and it also helps to increase public confidence and trust in the research process and its outputs.

The principles relating to the conduct of research are the following:

- research excellence and integrity: the MRC is dedicated to excellence and high ethical standards in the design, conduct, reporting and exploitation of publicly-funded research,
- respect, ethics and professional standards: all research must respect and maintain the dignity, rights, safety and wellbeing of all involved. Moreover, all researchers should be familiar with the relevant legal and ethical requirements and take appropriate steps to manage data appropriately, maintain confidentiality and minimise any adverse impact their work may have on people, animals and the natural environment.
- honesty and transparency: all those involved should be honest in respect of their actions and their responses to the actions of others, and this applies to the whole range of research activity,

- openness and accountability: MRC-funded researchers are expected to foster the exchange of ideas and to be as open as possible in discussing their work with other scientists and the public, furthermore the findings must be made available to the research community and the public and a complete and accurate account of scientific evidence must be presented to support the appropriate and effective use of this knowledge,
- supporting training and skills: all those involved in MRC-funded research have a responsibility to develop and maintain the skills necessary for their research and to assist and mentor others with their personal development.

3.5 Conclusion

In this chapter, the materials used in our experiments were presented. The dataset is the GoDARTS, which contains around 8 thousand patients with type 2 diabetes and around 102 thousand retinal images. These images were pre-processed to reduce the effect of image variations and ensure optimal performance of the neural network. The processing of sensible data was done inside the Safe Haven environment, which protects patients' confidentiality and ensures clinical data's correct and safe use.

The next chapter will illustrate the methods used to compute our analysis and results. In particular, we will see what a neural network is, how it works, and why we chose the EfficientNet-B2 to perform our experiments.

Chapter 4

Methods

4.1 About this chapter

In this chapter, neural networks and their structures will be investigated, particularly what is their architecture and how they learn from the data. Multiple algorithms will be presented in order to understand the steps that had been made to develop the Nadam algorithm. We will focus on convolutional neural networks that are commonly used and have excellent performance in classification tasks and image recognition. Then, we will focus on the neural networks mentioned in Chapter 2, which will be examined in more detail. Finally, the neural network that we used in our experiments, the Efficient Net B2 network, will be illustrated.

Neural networks are used for statistical analysis, data modelling and classification tasks. Some examples are image and speech recognition, textual character recognition, medical diagnosis or financial market indicator prediction. Neuroscientists and psychologists are interested in neural networks as computational models of the animal brain. Physicists and mathematicians are willing to understand their fundamental properties as complex systems. Commercial and industrial people use neural networks to model and analyse large and misunderstood datasets [14].

We can state that neural networks are a tool widely used in different sectors with different purposes. They are also finding natural applications in the healthcare sector and are the basis of artificial intelligence (AI).

4.2 Neural network

The human brain consists of an estimated 8.6×10^{10} nerve cells, called *neurons*, and each has, on average, 7000 synaptic connections. Neurons are the main component of the nervous tissue and communicate via electrical signals that are short-lived impulses or spikes in the cell membrane's voltage.
They consist of three components: the cell body, dendrites and a single axon. The terminal part of the axons is responsible for signal transmission and interneuron connections. Here synapses happen, and the electrical signal is converted into an electrochemical one which leaves the axon of a neuron and is received by the dendrites of another neuron. Synapses are electrochemical junctions located on cell branches, i.e. on the dendrites.

Typically, each neuron receives many thousands of connections from other neurons and these signals are integrated. Only if the resulting signal exceeds a threshold, the neuron will generate a voltage impulse in response which is transmitted to other neurons. This is the so-called *all-or-none* response: if a neuron responds at all, it must respond completely.

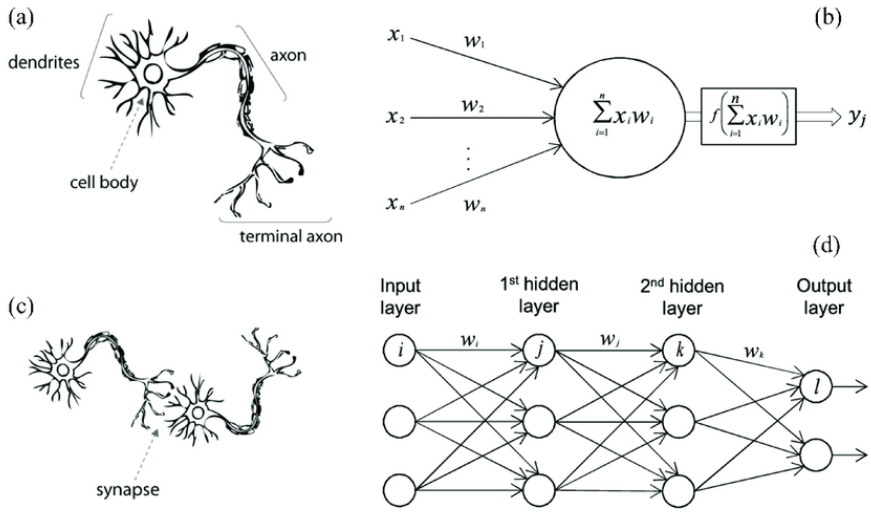


Figure 4.1: a) human neuron architecture, b) translation of a human neuron into a neuron in neural networks, c) connections between human neurons, d) connections between neurons in neural networks.

This architecture and communication system is simulated in neural networks composed of different layers full of thousands of neurons. Synapses are modelled by a weight such that each input is multiplied by it before entering the equivalent of the cell body. Here, the signals are summed together by simple arithmetic addition to supply a node activation and the total signal is compared with a threshold.

4.2.1 Architecture

A neural network consists of a pool of simple processing units which communicate by sending signals to each other over a large number of weighted connections.
The elements of a neural network are:

- a set of processing units called neurons,
- a state of activation y_k for every unit,
- connections between the units, and generally, each connection is defined by a weight w_{jk} which determines the effect that the signal of unit j has on unit k,
- an activation function $\mathcal{F}_k$, which determines the new level of activation based on the effective input $s_k(t)$ and the current activation $y_k(t)$,
- an external input, aka bias, θ_k for each unit,
- a method for information gathering, the learning rule with its learning rate parameter.

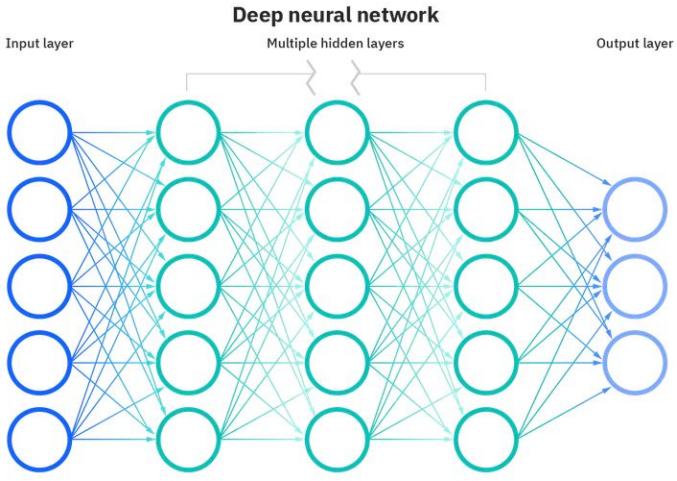


Figure 4.2: Deep neural network with an input layer, multiple hidden layers and an output layer.

Within a neural network, three types of units can be distinguished:

- input units that receive data from outside the neural network,
- output units that send data out of the network and compone the last layer,
- hidden units, whose inputs and outputs remain within the neural network.

Each neuron receives inputs from the previous layer and produces outputs that are the next layer's inputs. The total input to unit k is the weighted sum of the outputs from each of the connected units plus a bias term:

$$s_k(t) = \sum_j \omega_{jk}(t) y_j(t) + \theta_k(t)$$

Then an activation function gives the effect of the total input on the activation of the unit, it takes the total input $s_k(t)$ and the current activation $y_k(t)$ and produces a new value of the activation of the unit k:

$$y_k(t+1) = \mathcal{F}_k(y_k(t), s_k(t))$$

Often they are non-decreasing functions of the total input of the unit:

$$y_k(t+1) = \mathcal{F}_k(s_k(t)) = \mathcal{F}_k\Big(\sum_j \omega_{jk}(t) y_j(t) + \theta_k(t)\Big)$$

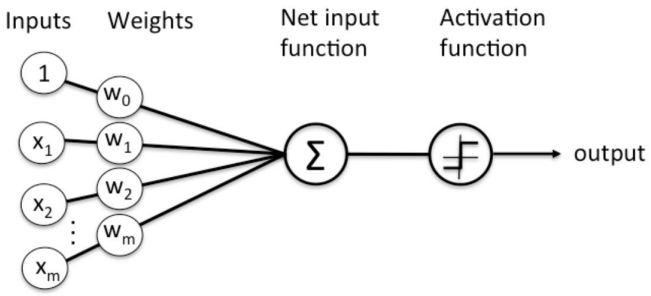


Figure 4.3: Activation function acting on a NN node.

We indicate a constant parameter whose value is set before the training process with the term *hyperparameter*. In neural networks, hyperparameters are the learning rate, the number of epochs, the batch size, the number of hidden layers, and the number of neurons for each layer.

4.2.2 Activation functions

Activation functions are an integral part of a neural network that introduces non-linearity and allows to perform more tasks. Without these functions, a NN is a simple linear regression model. An activation function decides whether a neuron should be activated, i.e. if the neuron's input is important or not in the process of prediction or classification. Its central role is transforming the weighted sum, input of the node, into a value, input of the following hidden or output layer.

A common activation function is the sigmoid one, whose curve looks like an S-shape. It is especially used in probability prediction since it exists between zero and one but never in hidden layers. It is defined as:

$$\Phi(x) = \frac{1}{1 + \exp^{-x}}$$

where x is the node's input. The sigmoid function is differentiable and monotonic, however, it may make the neural network stuck at the training time.

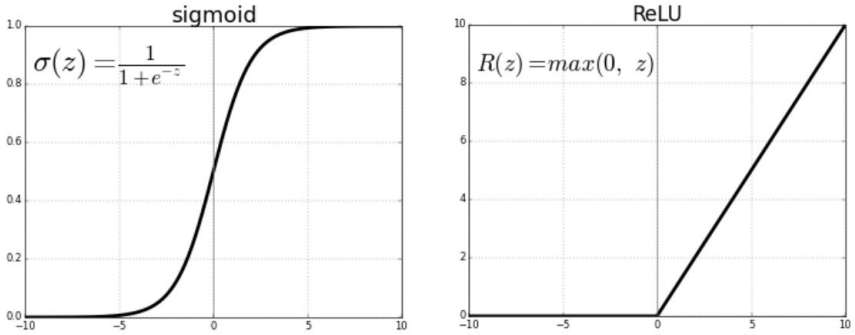


Figure 4.4: Comparison between the sigmoid and the ReLU activation function.

Another widely used activation function is the ReLU (Rectifier Linear Unit) function, used in the hidden layers and convolution networks. It is defined as:

$$f(x) = \max(0, x)$$

where x is the input of the neuron. It has the advantage of performing better gradient propagation with fewer vanishing gradients compared to the sigmoidal activation function that saturates in both directions. It is computationally efficient

and scales invariant. However, it is not differentiable at zero, not zero centred and not bound. It may suffer from the dying problem, which happens when neurons are pushed into an inactivation state such that no gradient flows backwards. They become stuck in a perpetually inactive state and die.

4.2.3 Learning process

A neural network must be tuned such that applying a set of inputs produces the desired set of outputs. One way is to set the weights explicitly using *a priori* knowledge, but usually, we do not have this information and have millions of parameters to set. The best way is to train the neural network by feeding it on teaching patterns and letting it change its weight according to some learning rule [14]. Then, it can be used to predict or classify the outcome of unseen data.

When we train the model, we evaluate its accuracy using a cost or loss function. One of the most used cost functions is the mean squared error (MSE), which we also used in our experiments, defined as:

$$MSE = \frac{1}{2m}\sum_{i=1}^{m}(\hat{y}_i - y_i)^2$$

where we consider the i-th sample, $\hat{y}_i$ is the predicted outcome and y_i is the actual outcome, m is the number of samples. We aim to minimise it to ensure the fit correctness for any given observation.

The cost function provides feedback to the model so that it can adjust its parameters and find the minimum. The iterative process stops when the cost function is equal to or close to zero.

The backpropagation algorithm adjusts the weights of the neural network through gradient descent, determining the direction to take in order to reduce errors and minimise the cost function. The parameters of the model are adjusted gradually to converge at the minimum.

The learning rate parameter determines the step's size at each iteration while moving toward a minimum of the loss function. It can be metaphorically thought of as the speed at which the model is learning.

A high learning rate shortens the training time but decreases the ultimate accuracy with the risk of overshooting the minimum. On the contrary, a low learning rate extends the training time and, thus, the computational cost, but with potentially more accuracy and precision.

A momentum can be used to allow the balance between the gradient and the last change such that the weight adjustments depend partially on the previous changes. If it is set close to zero, it emphasises the gradient, while a value close to one emphasises the last change.

The batch size is the number of samples passed to the network before updating the model's parameters. The term iterations is the number of batches needed to complete one epoch.
Batch normalisation is a technique for training neural networks that standardise the inputs to a layer for each batch, stabilising the learning process and reducing the number of epochs required to train the network [15]. It can be implemented during the training process by computing the neural network's weights through gradient descent, determining the direction to take to statistics just calculated. In common practice, two new parameters to learn are added to each layer: the input data's new mean and standard deviation, to make the scaling and shifting process automatic.

4.2.4 Backpropagation algorithms

Gradient descent (GD) is a first-order iterative optimisation algorithm where the gradient of the function to be minimised with respect to the parameters θ is computed, and a portion η of that gradient is subtracted from the parameters.
It is based on the idea that the direction to reach the minimum is the opposite direction of the function's gradient at the current point. In fact, this direction is the steepest descent. The gradient vector for each weight indicates by what amount the error would increase or decrease if a tiny amount increased the weight. The weight vector is then adjusted opposite of the gradient vector direction.
The algorithm aims to minimise a cost function built on the difference between the predicted and the desired output and the weights modifications are made to reduce this error. It requires two parameters: a direction and a learning rate.

Algorithm 1 Gradient descent [16]

$g_t \leftarrow \nabla_{\theta_{t-1}} f(\theta_{t-1})$
$\theta_t \leftarrow \theta_{t-1} - \eta g_t$

Stochastic gradient descent (SGD) is a procedure that shows an input vector for a few examples, computes the outputs and the errors, computes the average gradient for those examples and adjusts the weights accordingly. The process is repeated for any small sets of samples from the training set until the average of the objective function stops decreasing. The name stochastic derives from the fact that each small group gives a noisy estimate of the average gradient over all examples [17]. Then the performance of the system is measured on the test set, data that are not previously seen in order to test and measure the generalisation ability of the model.
The frequent updates are easier to store in memory and offer more detail and speed. Moreover, they result in noisy gradients, which are useful in escaping the local minimum and saddle points. SGD uses a single learning rate for all weights updates that does not change during training.

Two problems can be experienced in deeper neural networks: the vanishing gradient and the exploding gradients. In the first one, the gradient becomes smaller and smaller during backpropagation. Consequently, the earlier layers learn more slowly than the later layers, and the weight parameters update until zero. Therefore, the algorithm is no longer learning. In the exploding gradients, the gradient and the parameters become too large, causing model instability.

Stochastic gradient descent may have problems in finding the global minimum in ravines, which are areas where the surface curves much more steeply in one dimension than another, common near local minima of the cost function [18]. In these situations, SGD oscillates around the ravine's slope, making little progress towards the minimum. As shown in the figure below, momentum makes SGD accelerate in the right direction and reduces the oscillations.

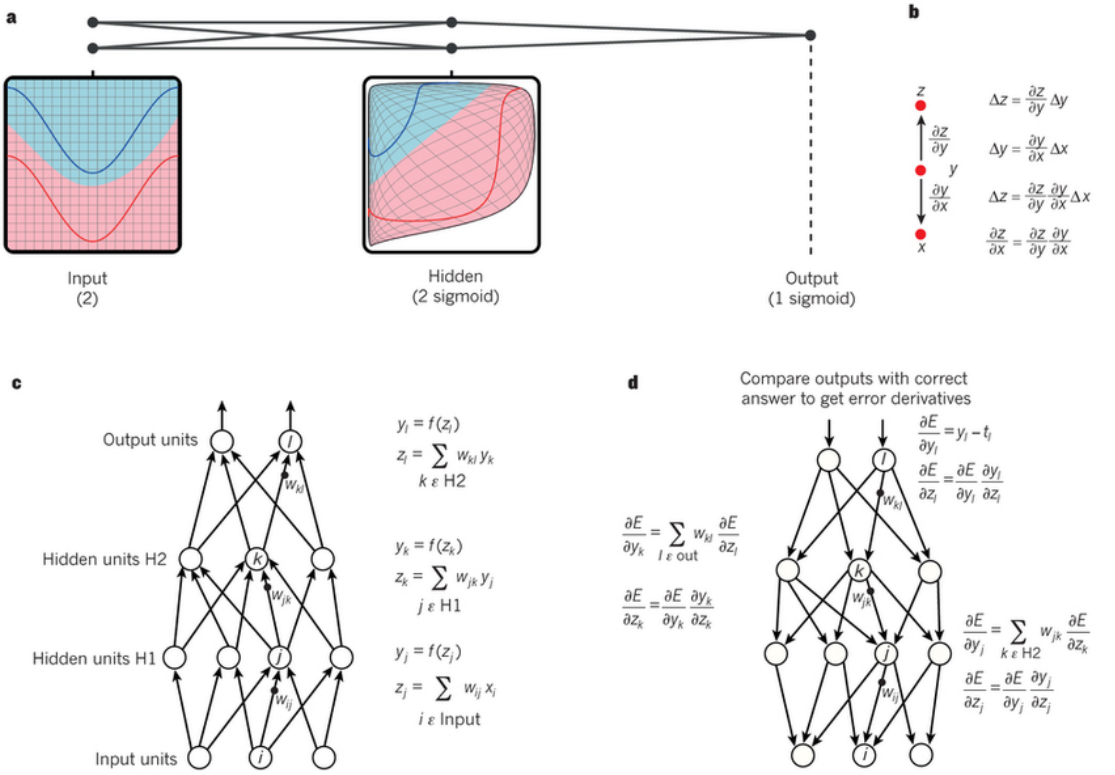


Figure 4.5: a) A neural network distorts the input space to make the data classes linearly separable, b) The chain rule of derivatives. A small change Δx in x is transformed into a small change Δy in y and then into a change Δz in z. c) The forward pass: at each layer, the total input z is calculated as a weighted sum of the outputs of the units of the previous layer. Then, a non-linear function $f(\cdot)$ is applied to obtain the layer's output. d) The backward pass: at each hidden layer, the error derivative with respect to the output is computed and then converted into the error derivative with respect to the input [17].

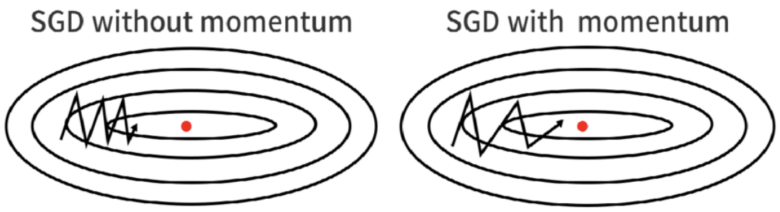
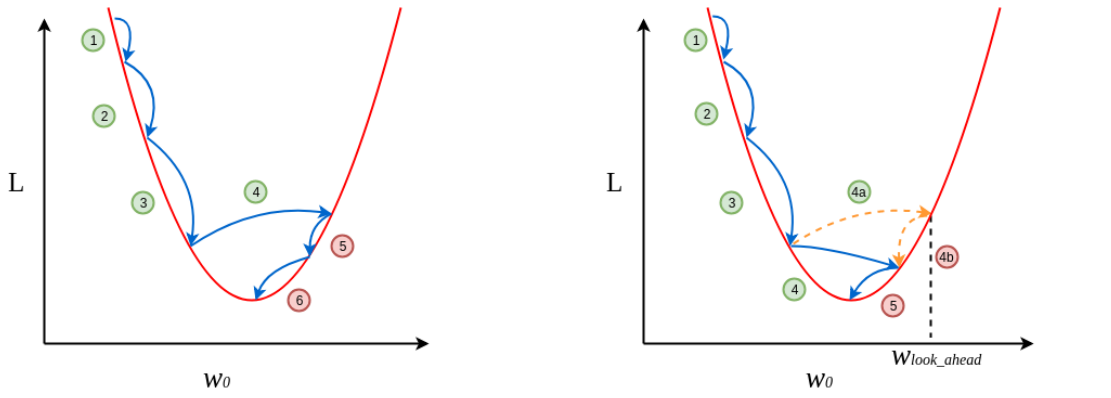


Figure 4.6: Comparison between SGD progress without momentum and with momentum.

Traditional momentum considers a variable that represents the last performed update. It accumulates a decaying sum with decay constant μ of past gradients and continues to move in their direction. The previous update is added into a momentum vector m and the hyperparameter η controls how much of the last change to add. We can compare it to a ball rolling downhill that accelerates in the same direction even in the presence of small hills [15].
SGD does not calculate the exact derivative of the cost function but estimates it on a small batch. Consequently, it may happen that the weights updates do not move in the right direction because of the noisy gradients. Using a momentum variable allows us to estimate the cost function's gradient better and move in the right direction.
However, traditional momentum suffers from the overshooting of the minima, which is a problem solved with the Nesterov approach.

Algorithm 2 SGD with traditional momentum

$g_t \leftarrow \nabla_{\theta_{t-1}} f(\theta_{t-1})$
$m_t \leftarrow \mu m_{t-1} + g_t$
$\theta_t \leftarrow \theta_{t-1} - \eta m_t$



(a) Momentum-Based Gradient Descent (b) Nesterov Accelerated Gradient Descent

$\bigcirc \Longrightarrow \frac{\partial L}{\partial w_0} = \frac{Negative(-)}{Positive(+)}$ $\bigcirc \Longrightarrow \frac{\partial L}{\partial w_0} = \frac{Negative(-)}{Negative(-)}$

Figure 4.7: Comparison between the tradition momentum and the NAG momentum.

An extension of the gradient descent optimisation algorithm is the **Nesterov accelerated gradient momentum** (NAG), which uses the partial derivative of the projected update rather than the derivative current variable value. At first, it calculates the projected position of the variable using the change from the last iteration and then uses its derivative for the calculation of the new position of the variable. The idea is to make a first big jump in the direction of the previously accumulated gradient and then measure it and make corrections.

The four steps of the algorithm are the followings:

- project the solution's position: projection$(t+1)$=x(t)+momentum*change(t),
- and calculate its gradient: gradient$(t+1)$=f'(projection$(t+1)$);
- calculate the change in the variable using the partial derivative: change$(t+1)$=momentum*change(t) - (step size)*gradient$(t+1)$;
- and update the variable: x$(t+1)$=x(t)+change$(t+1)$.

It has been demonstrated that NAG momentum improves the rate of convergence, i.e. the number of iterations required to find a solution is decreased.

Algorithm 3 Nesterov-accelerated gradient

$g_t \leftarrow \nabla_{\theta_{t-1}} f(\theta_{t-1} - \eta\mu m_{t-1})$
$m_t \leftarrow \mu m_{t-1} + g_t$
$\theta_t \leftarrow \theta_{t-1} - \eta m_t$

The challenge is finding the correct learning rate and momentum value in order to guarantee convergence. We can write NAG to be more straightforward and efficient in implementation as follows [16]:

Algorithm 4 NAG rewritten

$g_t \leftarrow \nabla_{\theta_{t-1}} f(\theta_{t-1})$
$m_t \leftarrow \mu_t m_{t-1} + g_t$
$\bar{m}_t \leftarrow g_t + \mu_{t+1} m_t$
$\theta_t \leftarrow \theta_{t-1} - \eta \bar{m}_t$

Choosing the optimisation algorithm for a neural network is crucial to time the solution's convergence. The **Adaptive Momentum** (Adam) optimisation algorithm is an extension of the stochastic gradient descent to update network weights iteratively. It has two main components: a momentum and an adaptive learning rate, and it combines the advantages of two techniques: the Adaptive Gradient Algorithm (AdaGrad) and the Root Mean Square Propagation.

AdaGrad adaptively scales the learning rate with respect to the accumulated squared gradient for each dimension at each iteration. It decreases the learning rate along dimensions that have already changed significantly and increases it along dimensions that have changed slightly.
At first, a step size for a given dimension is calculated by summing the partial derivatives for the examined parameter. Then, it is used to make a move in that dimension using the partial derivative.
The initial step size is divided by the square root of the sum of squared partial derivatives, such that the learning rate is shrunk according to the entire history of the squared gradient [19].

However, since it accumulates the gradients from the beginning of the training process, the norm vector may become very large. Consequently, the search progress slows, and the network no longer learns because the learning rate is almost zero. Moreover, its performance worsens when the loss function is non-convex and gradients are dense. This limitation is dealt with the RMSprop approach that replaces the sum of past gradients with a decaying mean, allowing the neural network to learn infinitely.

Algorithm 5 AdaGrad

$g_t \leftarrow \nabla_{\theta_{t-1}} f(\theta_{t-1})$
$n_t \leftarrow n_{t-1} + g_t^2$
$\theta_t \leftarrow \theta_{t-1} - \eta \frac{g_t}{\sqrt{n_t}+\epsilon}$

Algorithm 6 RMSprop

$g_t \leftarrow \nabla_{\theta_{t-1}} f(\theta_{t-1})$
$n_t \leftarrow \nu n_{t-1} + (1-\nu) g_t^2$
$\theta_t \leftarrow \theta_{t-1} - \eta \frac{g_t}{\sqrt{n_t}+\epsilon}$

RMSProp is a stochastic technique that introduces a moving average of squared gradients for each weight to normalise the gradient. It computes the moving average of the partial derivatives instead of the sum in calculating the new learning rate. The normalisation balances the step size because it decreases the momentum for large gradients to avoid exploding gradients and increases it for small gradients to prevent the problem of vanishing gradients. The usage of the moving average allows the neural network to forget early gradients, focusing on the most recently observed partial gradients, overcoming the limitation of AdaGrad.

Adam combines the classical momentum with RMSprop to improve performance. It is a first-order gradient-based algorithm based on adaptive estimates of lower-order moments. It needs three hyperparameters: the initial learning rate and the decay rates of the first and second-order moments.
The algorithm estimates the gradient's first-order moment (the gradient mean) and the second-order moment (element-wise squared gradient) using the exponential moving average and then corrects its bias. The first moment normalised by the second moment gives the direction of the update. An initialisation bias correction term overcomes the instability that the initialization of m and n to zero may create.

Algorithm 7 Adam

$g_t \leftarrow \nabla_{\theta_{t-1}} f_t(\theta_{t-1})$
$m_t \leftarrow \mu_t m_{t-1} + (1-\mu) g_t$
$n_t \leftarrow \nu n_{t-1} + (1-\nu) g_t^2$
$\hat{m} \leftarrow \frac{m_t}{1-\mu^t}$
$\hat{n} \leftarrow \frac{n_t}{1-\nu^t}$
$\theta_t \leftarrow \theta_{t-1} - \eta \frac{\hat{m}_t}{\sqrt{\hat{n}_t + \epsilon}}$

Each iteration of the algorithm performs the following steps:

- computation of the gradient and its element-wise square using the current parameters,
- updating the exponential moving average of the two moments,
- computation of the unbiased average of the moments,
- updating the weight by dividing the first order moment unbiased average by the square root of the second order moment unbiased average and then scaling by the learning rate,

- applying the update to the weights.

The advantages of Adam are several: it is straightforward to implement and computationally efficient, it requires little memory, it is invariant to diagonal rescale of the gradients, and it is appropriate initialisation very noisy or sparse gradients and with a lot of data and parameters. Moreover, its hyperparameters are intuitively interpretable and usually require little tuning.

Since Adam combines two algorithms that are beneficial for different reasons and Nesterov momentum is theoretically superior to traditional moment, we can think of combining Adam with NAG to obtain an algorithm that overcomes the limitations of the singular approaches [16].

Algorithm 8 Nesterov-accelerated Adaptive Momentum Estimation (Nadam)

Require: $\alpha_0, \ldots, \alpha_T$; $\mu_0, \ldots, \mu_T$; ν; ϵ: hyperparameters
  m_0; $n_0 \leftarrow$ first and second moment vectors
  while θ_t not converged **do**
    $g_t \leftarrow \nabla_{\theta_{t-1}} f_t(\theta_{t-1})$
    $m_t \leftarrow \mu_t m_{t-1} + (1 - \mu_t) g_t$
    $n_t \leftarrow \nu n_{t-1} + (1 - \nu) g_t^2$
    $\hat{g}_t \leftarrow \frac{g_t}{1 - \prod_{i=1}^{t} \mu_i}$
    $\hat{m} \leftarrow \frac{m_t}{1 - \prod_{i=1}^{t+1} \mu_i}$
    $\hat{n} \leftarrow \frac{n_t}{1 - \nu^t}$
    $\bar{m}_t \leftarrow (1 - \mu_t)\hat{g}_t + \mu_{t+1}\hat{m}_t$
    $\theta_t \leftarrow \theta_{t-1} - \eta \frac{\bar{m}_t}{\sqrt{\hat{n}_t} + \epsilon}$
  end while

4.3 Convolutional neural networks

Convolution neural networks (CNNs) are similar to feedforward networks, which comprehend an input layer, multiple hidden layers, and an output layer. However, CNNs are easier to train. In fact, the number of weights per layer is smaller, resulting in fewer parameters, which is an advantage in processing high-dimensional data such as images. CNNs are usually applied for image recognition and classification, pattern recognition and processing data organised in multiple arrays. Their capacity can be controlled by varying their depth, width and height.

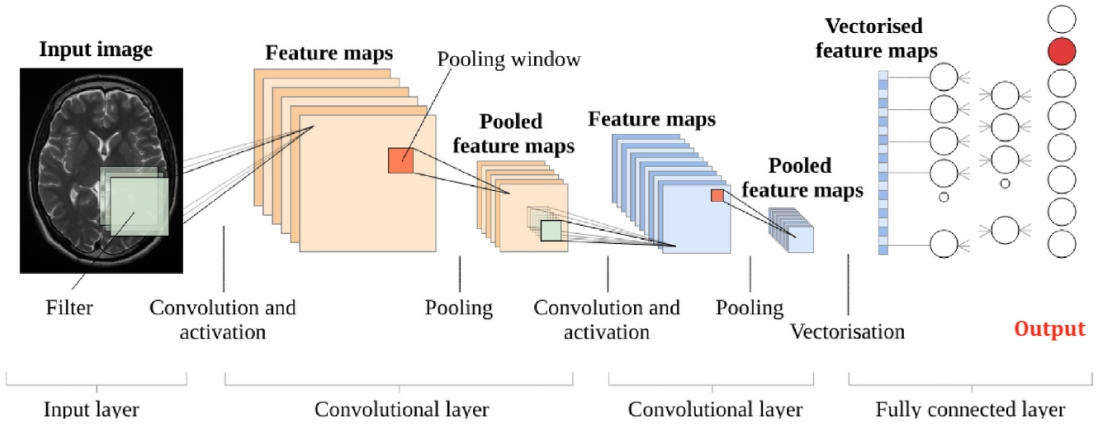


Figure 4.8: Example of a convolution network used for classification.

Convolutional networks usually have three types of layers:

- **convolutional layers**, that apply spatial filters, also called kernels, to input images to produce feature maps that indicate the presence of a feature in the image. The neural network learns the filter parameters in the training process in order to maximise the response to a local region of the input image. Especially, the first layers detect the low-level feature, while the later layers the high-level features.
  Each filter is convolved with the input image to compute an activation or feature map that detects the presence of a feature. The dot product between the pixels of the image and the value of the kernel is computed step by step at every spatial position. All units in a feature map share the same filter bank, but different feature maps in a layer use other filter banks.

Usually, convolutional filters are smaller than the input images, so each neuron is connected to only a tiny region of the image whose size is obviously equal to the filter size. Moreover, they are not fully connected, which means that not all input nodes affect all output nodes. Therefore, these layers have more flexibility in learning.
The convolution is very useful in images where local data groups are often highly correlated, as happens in images, forming neighbourhoods of points that can be easily detected and identified by the network.

- **pooling layers**, whose function is to merge similar features into one by summarising the presence of features in patches of the feature map. It is the so-called "downsampling" process [15] that reduces the spatial dimensionality and the size of the feature maps. These layers aim to create a lower resolution version of an input signal that still contains the essential features or elements but not the fine details that are usually unnecessary for the task.
  Pooling layers are generally added after convolutional layers and this structure can be repeated multiple times within a CNN, as we can see in Fig.4.8.
  Pooling requires the choice of a pooling operation that acts through a filter whose size is smaller than the size of the feature maps. In fact, it always reduces the dimension of the feature map, thus the number of pixels. Close pooling units are used to reduce the dimension of the representation and create invariance to small shifts, local translation, and distortions [17].
  A typically pooling function is the Max Pooling which returns the maximum or larger value of a local patch of points in a feature map. The output is the most present feature in that activation map.
  Another common function is the Average Pooling, which calculates the average value for each patch of the feature map, whose size is given by the kernel's size and returns the feature's average presence.

- **fully connected layers**, which correspond to the final layer in a CNN in which the neuron applies a linear transformation to the input vector through a weights matrix. Then a non-linear activation function is used to the product to get the output. In FC layers, all possible connections layer to layer are present, i.e. every input affects every output since every neuron is connected

to every neuron of the previous layer. However, not all weights affect all outputs and the output size can be arbitrarily chosen with the number of columns in the weights matrix.

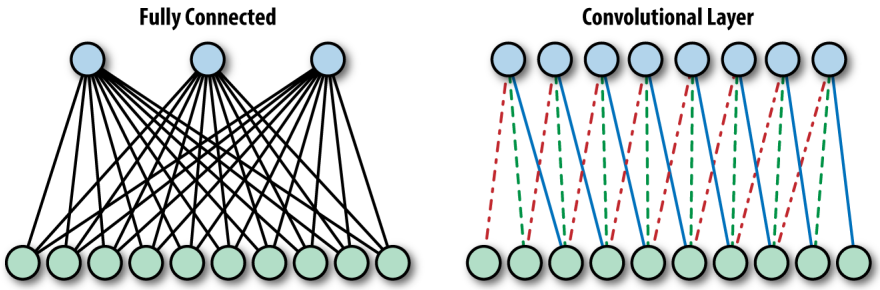


Figure 4.9: Comparison between a fully connected layer, where each neuron affects the output of each neuron in the following layer, and a convolutional layer, where a kernel selects the neurons that affect the output of the neurons in the next layer.

In Chapter 2, we have presented different papers in which neural networks were applied to retinal images to predict some outcomes with medical importance. Now we can provide more details of the used neural networks.

Poplin et al. [6] used the Inception-v3 neural network to predict cardiovascular factor from retinal images. Inception neural networks are convolutional networks that can detect the salient part of an image through the applications of different kernels with different sizes at the same level. In a traditional convolutional network, multiple deep layers may overfit the data. Moreover, the essential parts of an image that are captured by the network for classification may have different sizes. Therefore, choosing the right kernel size is crucial, and its tuning may not be so easy. The idea of the Inception networks is to apply filters of multiple sizes at the same level of the network to capture both the information globally distributed and the detailed information locally distributed in the input image. The result is a wider rather than deeper neural network.
Inception-v3 has 42 layers and includes convolution layers with kernel size $3\times$ 3, average pooling layers, Max pooling layers, organised in parallel blocks linked with concatenation layers, and lastly, a dropout, a fully connected and a softmax layer.

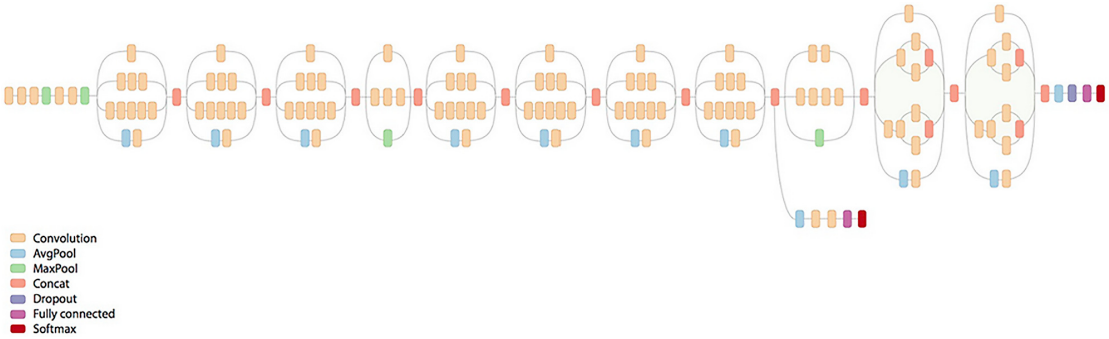


Figure 4.10: Architecture of the Inception-v3 network.

Kim et al. [7] used a convolutional network, particularly a ResNet-152 with a convolutional layer, followed by batch normalisation and ReLu activation function for 151 layers and a final fully connected layer. The kernel size of the first convolution layer was 7×7 with stride two and padding three, while the kernel of the other Conv layers was set to 3×3. The network was pre-trained in general image classification with the ImageNet database (a large database with more than 14 million images from 20.000 categories used in visual object recognition and classification tasks) and the found parameters were used as an initial guess in the training process.

Gerrits et al. [8] applied to retinal images a MobileNet-v2 pre-trained on ImageNet database to predict cardiovascular risk factors.
MobileNet is a class of CNNs that uses depthwise separable convolutions. Therefore, the number of parameters and the computational cost are significantly reduced in comparison with traditional convolution networks. Separable convolutions are made possible through two operations: a depthwise convolution and a pointwise convolution. MobileNets factorize the traditional convolution into a 3×3 depth-wise followed by a 1×1 pointwise convolution. The idea is to separate the dimensions of the kernels, i.e. separate the depth dimension from the horizontal one and then use a unitary filer to cover the third dimension.
MobilNet-v2 has 17 bottleneck residual blocks followed by a 1×1 traditional convolutional layer, a global average pooling and a classification layer. In each block, there is an expansion layer and a depthwise convolution layer, both followed by batch normalization and a ReLu activation function, then there is a projection layer followed by only batch normalization.

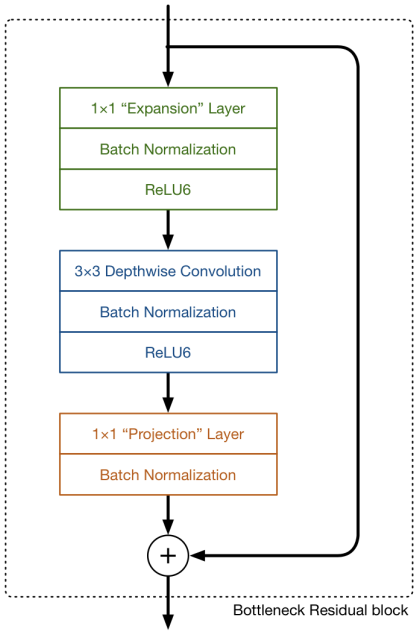


Figure 4.11: Architecture of the residual bottleneck block in MobileNet-v2.

Projection layers or bottleneck layers are added in convolutional blocks with the function of reducing the dimension of the data entering the layer. They are linear blocks since no activation function is applied to the output of these layers because non-linearity elements may destroy important information contained in the data. Expansion layers work as opposed to projection layers, their output dimension is bigger than the input dimension. Moreover, in each block, there is a residual connection which is a path that allows data to reach some layers of the neural network by skipping others. Residual connections help the gradients' backpropagation process making it converge faster and easier.

4.4 EfficientNets

Convolutional neural networks are usually developed according to a finite number of resources and then scaled up to obtain better performance if more resources are available.
Tan et al. [10] proposed an innovative scaling method, called *compound scaling method*, to balance the dimensions of convolutional neural networks. They obtained

a new type of convolutional neural network, the EfficientNets, which has been empirically shown to have higher performance with significantly fewer parameters. They tested these new networks on the ImageNet dataset and obtained a higher accuracy compared to the one obtained with ResNet or other ConvNets, as we can see below. They used the RMSprop optimiser.

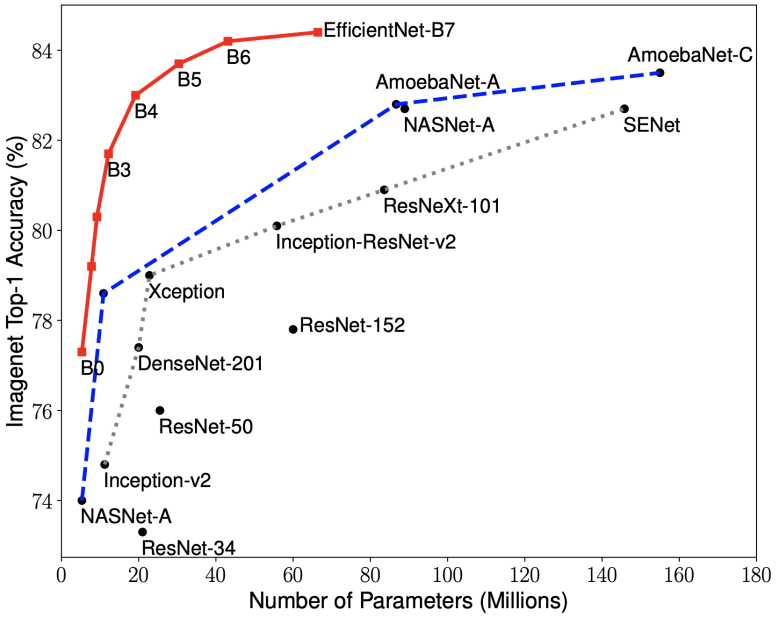


Figure 4.12: Number of parameters and accuracy of different convolutional neural networks applied to the ImageNet database. Efficient Nets have higher accuracy, reached with fewer millions of parameters.

The dimensions of a neural network are scaled independently, and often, only one of them is modified. The three possible dimensions to scale are:

- depth, which is typically increased since deeper networks can detect more complex features even if they are more difficult to train and may suffer from vanishing gradients problem;
- width, which is commonly modified in small models. Wider networks capture fine features and are easier to train. However, they have difficulty capturing high-level features if they are too wide.
- resolution, which is increased in ConvNets to detect more fine features and patterns.

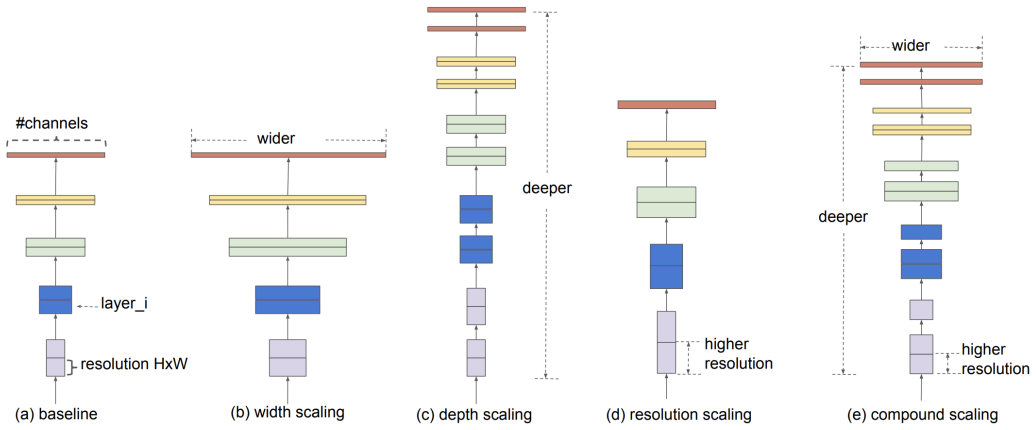


Figure 2. **Model Scaling.** (a) is a baseline network example; (b)-(d) are conventional scaling that only increases one dimension of network width, depth, or resolution. (e) is our proposed compound scaling method that uniformly scales all three dimensions with a fixed ratio.

Figure 4.13: a) baseline neural network; b,c,d) traditional scaling approaches according to which only one dimension is increased; e) compound scaling method in which the three dimensions are scaled uniformly with a fixed ratio.

The scaling process requires a lot of time and manual tuning in order to find the architecture that allows us to reach the best performance on our data. For example, ResNet can be scaled down or up to ResNet-200, the deepest ResNet network. However, we are not sure to have reached the optimal solution and often, we only find a sub-optimal tuning and thus a sub-optimal accuracy and efficiency. In fact, the three dimensions depend on one another. For this reason, *Tan at al.* thought to scale them with a common uniform criterion to balance them. The balance between depth, width and resolution is achieved by scaling them with a constant ratio, represented by the parameter Φ in the following way:

$$\begin{aligned} \text{Depth:} \quad & d = \alpha^{\Phi} \\ \text{Width:} \quad & w = \beta^{\Phi} \\ \text{Resolution:} \quad & r = \gamma^{\Phi} \\ \text{such that:} \quad & \alpha \cdot \beta^2 \cdot \gamma^2 \approx 2 \\ & \alpha \geq 1, \beta \geq 1, \gamma \geq 1 \end{aligned}$$

where α, β and γ are constant determined by a small grid search [10].

Starting from the baseline EfficientNet-B0, two steps have to be computed to obtain the other networks:

1. fixing $\Phi = 1$ and finding the optimal parameters α, β and γ for the EfficientNet-B0 network. In particular, they found $\alpha = 1.2$, $\beta = 1.1$ and $\gamma = 1.15$.

2. then fixing α, β and γ and finding Efficient Nets-B1 to B7 by using different values of the parameter Φ.

The architecture of these networks starts from a common structure of EffNet-B0 which is then built up to obtain networks from B0 (with 237 layers) to B7 (with 813 layers) [20]. There is a stem and a final layer that are equal in all eight networks, and they are represented in Fig.4.14.

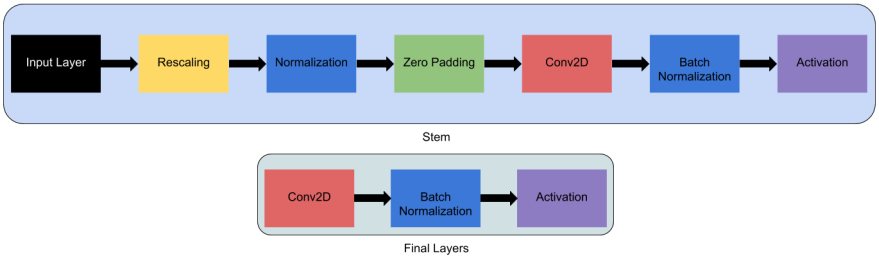


Figure 4.14: Common structure in all Efficient Nets.

The different layers are made up of 5 modules, represented in Fig.4.15. Then, these modules are combined to create sub-blocks which form the various layers.

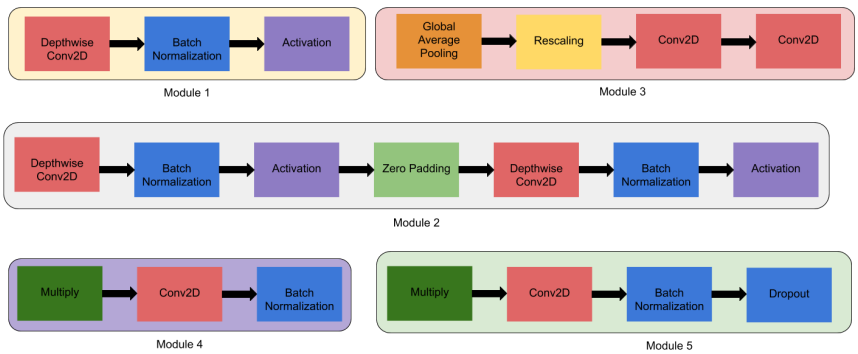


Figure 4.15: Modules that are used to build all the Efficient Nets.

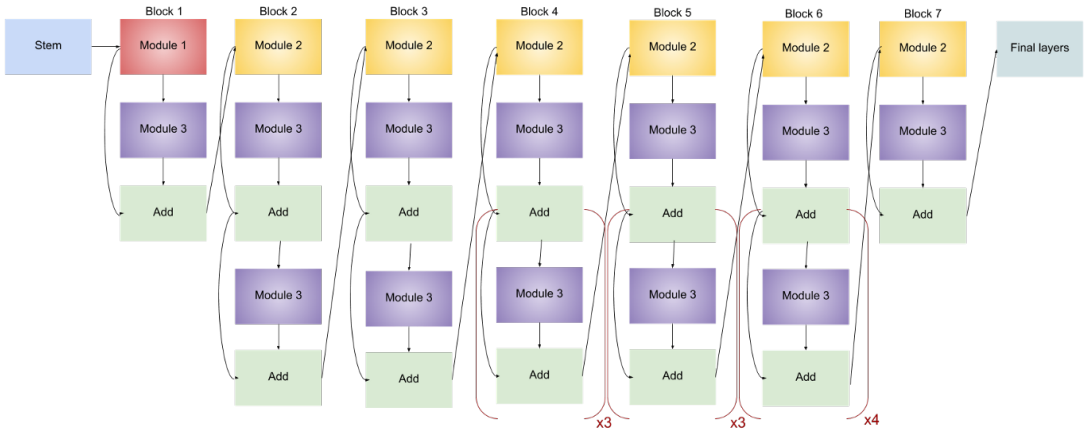


Figure 4.16: Architecture of EffNet-B2.

In our experiment, we used the same neural network used by *Ghouse et al.* [9] to predict the cumulative glycated haemoglobin from retinal images of diabetic Scottish patients. EfficientNet-B2 was implemented in Python, and Keras 2 and TensorFlow 1.9.0 packages were used to build, train and test the network.
Keras is an effective open-source neural network Application Programming Interface (API) written in Python and integrated with the TensorFlow library. TensorFlow is an end-to-end open-source deep learning framework used for developing neural networks.

In our Efficient Net B2, the fully connected layer was replaced with a global average pooling layer followed by a single output node with linear activation. The convolutional layers were not modified, and their weights were initialised by applying the network to the ImageNet database. The total number of trainable parameters was 7.7 million.

4.5 Conclusion

In this chapter, neural networks have been presented. In particular, we started with a biological overview to understand where the idea of implementing artificial neural networks was born. We explored their architecture, the possible layers that can be inserted and the activation functions that can be applied to each node of

the network. The role of activation functions is to introduce some non-linearities, and they are essential elements in making a neural network different from a simple linear regression model.

Then we explained the learning process, which aims to reduce the error between the real output of the data and the output predicted by the network. A cost function quantifies this error. With the term *learning*, we indicate the process in which the neural network learns the parameters of each layer by changing them and testing their performance on a labelled training set. The tunable parameters are the weights of the layers composing the network and they are changed in a backpropagation process that aims to minimise the cost function.

There are many algorithms which compute the backpropagation process by calculating the gradient of the error and then adjusting the network's weights according to it. The first proposed was the gradient descent algorithm, which has been improved in the stochastic gradient descent one. The advantages of adding a momentum to the process have been explained, and some efficient algorithms have been presented. We explored the Nesterov-accelerated gradient technique in more detail, which makes the convergence process faster and lets us obtain a better approximation of the calculated gradients.

Further steps have been made to arrive at the development of the Adam algorithm, which combines the advantages of two already existing techniques, AdaGrad and RMSProp. The combination between Adam and NAG is called the Nadam algorithm and it is used in our Efficient Net B2 to train the network and learn the optimal parameter for our task.

In the next chapter, we will present the experiments and the analysis carried out at the University of Dundee in Scotland by my colleague Quinto Andrea and me, under the supervision of Professor Trucco Manuel and with the precious collaboration of Doctor Doney Alex.

Chapter 5

Experiments and Results

5.1 About this chapter

The previous chapters presented the materials and methods needed to perform our experiments. In this chapter, we will explain in detail the experiments we performed and the results we obtained and achieved.

First of all, we tried to reproduce the results of *Boyle Liam* [21]. We performed a feature selection with Lasso regression and bootstrap analysis on a data set comprehending both clinical and retinal features. The dataset was a subset of the GoDarts database. The retinal features had been extracted from retinal images with the VAMPIRE software developed by the University of Dundee and the University of Edinburgh.

Then we moved inside the Safe Haven environment. The first experiment was made to familiarise ourselves with the data and the code and to reproduce the results obtained by *Ghouse et al.* [9]. The aim was to predict the age of diabetic patients from retinal images using an Efficient Net B2 network.

The following step was the prediction of the cumulative glycated haemoglobin using the same neural network and retinal images. We performed three different experiments and we elaborated on the results of the last one, the most correct and complete.

5.2 Feature selection with Lasso regression

In the first phase of our work, we tried to reproduce the results obtained by Boyle Liam [21], in particular, which are the most important features to predict a cardiovascular failure, indicated by the covariate *CVD_fail* in our dataset.

The dataset contains 4711 patients and 184 covariates between clinical and retinal features, extracted by retinal images with the software VAMPIRE, version 3.2.

The aim was to predict the covariate *CVD_fail* as a linear combination of the other features and then establish the most significant ones for the outcome prediction. The outcome equal to 1 means that the patient experienced a cardiovascular fail event, while equal to 0 means that cardiovascular disease is absent.

Cardiovascular disease is one of the leading causes of death and disability in the UK; only in 2019, an estimated 17.9 million people, which represents the 32% of global deaths, died from a cardiovascular disease [22]. However, this can be prevented by leading a healthy lifestyle. It is important to understand what environmental factors, lifestyle habits or clinical measurements are significant in developing a CVD to act on them and prevent the disease or slow its progress.

Our work has been organised in the following steps:

1. **pre-processing** of the data:

    - **correlation**: we investigate the correlation between the covariates and the relationship between them. We decide to delete all the covariates that were not important and misleading for our outcome prediction:
        - the id of the patient and the image size, not linked to a CVD,
        - the death and the date of death because the patient's death is a future event which has not still happened at the moment of the analysis and is not relevant to CVD prediction,
        - the *date_age55* which indicates the date at which the patient is 55 years old and is highly correlated with *dob*, the date of birth,
        - *e_date* which is the date at which the retinal image was taken, but it is highly correlated con *e_age*, which is the age of the patient at the moment of the image,

  - *dement_time* and *dement_date* which are highly correlated with *dement_time_age*,
  - *cvd_fail* because it is the outcome that we aim to predict and *cvd_time* because we have already in the dataset *cvd_time_age*.

- **splitting of the dataset**: we split our dataset into a training and a test set to train our model with the training set and then test its performance on the test set.
  Since the two classes are unbalanced because we have 3095 patients with *cvd_fail* equal to 0 and 1616 ones with *cvd_fail* equal to 1, we performed a stratified split to maintain the same class proportion of the original dataset in both the training and test sets.
  We have in the original dataset, in the training set (3533 subjects) and in the test set (1178 subjects) 34% of patients who experienced a CVD and 66% who did not.

- **imputation of the missing values**: we hypothesize that the missing values are missing values completely at random (MCAR), which means that the missing values do not depend on patient characteristics and on the missing data, but their occurrence is random. Under this hypothesis, we can perform their imputation and replace them.
  We use the k-nearest neighbours' algorithm with $k = 10$. For each missing data, it searches for the ten nearest values using a Euclidean metric, calculates their average value and replaces the missing value with it.

2. **model assessment and cross-validation**: we fit our training set with a logistic model and then apply cross-validation with folds number equal to 10 to find the optimal value of λ.
   We use the R package `glmnet`, which fits a data set with a generalised linear model (GLM) by penalising a maximum likelihood. The minimisation problem is the following:

$$\min_{\beta_0,\beta} \frac{1}{N} \sum_{i=1}^{N} \omega_i l_i(y_i, \beta_0 + \beta^T x_i) + \lambda \Big[\frac{1-\alpha}{2} ||\beta||_2^2 + \alpha ||\beta||_1 \Big]$$

By default, the parameter α is equal to 1. Hence, the GLM model becomes a Lasso (Least Absolute Shrinkage and Selection Operator) regression model. Lasso regression is a regularization technique used to increase the generalization ability of the model in predicting correctly the outcome of previously unseen data. It performs regularization, i.e. it shrinks towards zero the model coefficients, and feature selection by setting equal to zero the coefficient of covariates that are less significant for outcome prediction.

In the R function, we set a binomial `family` because we have two classes that are fitted with a logistic model. Thus, the minimisation problem becomes:

$$\min_{\beta_0,\beta} -\Big[\frac{1}{N}\sum_{i=1}^{N} y_i \cdot (\beta_0 + x_i^T\beta) - \log(1 + e^{(\beta_0 + x_i^T\beta)})\Big] + \lambda\Big[\frac{1-\alpha}{2}||\beta||_2^2 + \alpha||\beta||_1\Big]$$

The loss function used in the cross-validation is the misclassification error in the case of binomial classes.
We set the variable `standardise` = True to standardise the data of the training set and obtain columns with zero mean and unitary standard deviation. Covariates may have very different and large scales, so it is common practice to standardise them. Moreover, we weigh the weights of the model in order to take into account the fact that the two classes are unbalanced.

3. **bootstrap analysis**: at last, we perform a bootstrap analysis to assess the stability of feature selection.
   Feature selection methods may suffer from instability; in fact, variations of the same training set can lead to different results and selected features. Resampling methods such as cross-validation or bootstrap analysis are performed to make a feature selection more stable. Bootstrap analysis establishes that, at each iteration, we resample with replacement of the original training set to create an internal training set with a number of observations equal to the numerousness of the original dataset and an internal test set with the out-of-bag elements.
   At each iteration, we train the model on the internal training set and test its performance on the internal test set. Ultimately, we can average the results and assess the stability of feature selection.

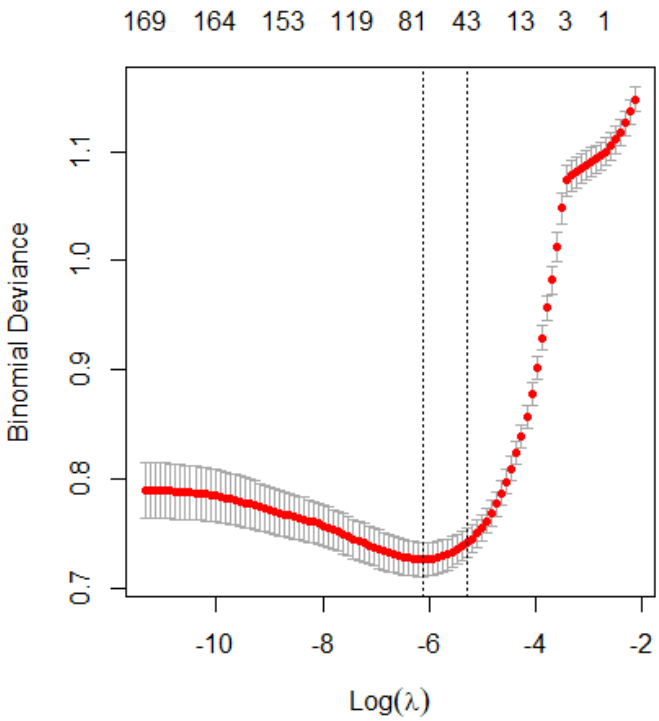


Figure 5.1: Curve obtained with the cross-validation which shows the deviance as a function of the number of features selected. The left vertical line represents the model with λ_{min}, the right line the model with λ_{1SE}.

	Feature	Counts	Percentage
1	leverhulme	500	100%
2	pre	500	100%
3	dement_time_age	500	100%
4	e_age	500	100%
5	odfovea	485	97%
6	ldrvspline	464	93%
7	bcvspline	450	90%
8	trig	440	88%
9	gradq3vspline	426	85%
10	tortimageg1amed	425	85%
11	dob	420	84%
12	tortq1g1amin	418	84%
13	gh	414	83%
14	gradq1ahermite	413	83%
15	odradiuspx	408	82%

Figure 5.2: Features that have been selected at most with bootstrap analysis.

As we can see above, the most important features in predicting cardiovascular disease are the clinical ones. In particular, the age when the image was taken (*e_age*), the age at diagnosis of dementia (*dement_time_age*), and the presence of past cardiovascular disease (*pre*) have been selected 100%. Some retinal features have been selected with a high percentage too. The variable *leverhulme* should be neglected because it indicates the source of the dataset, so it is not correlated with our outcome. We note also that the glycated haemoglobin (*gh*) has been selected with a high percentage equal to 83%, indicating its importance in predicting cardiovascular failure.

5.3 Age prediction with neural network

As the first step, in order to familiarize ourselves with the code, the Safe Haven environment and the data, we reproduced the results obtained by *Ghouse et al.* [9]. We applied the Efficient Net B2 network, described in more detail in Chapter 4, to the retinal images of the GoDarts database to predict the age of the patient.

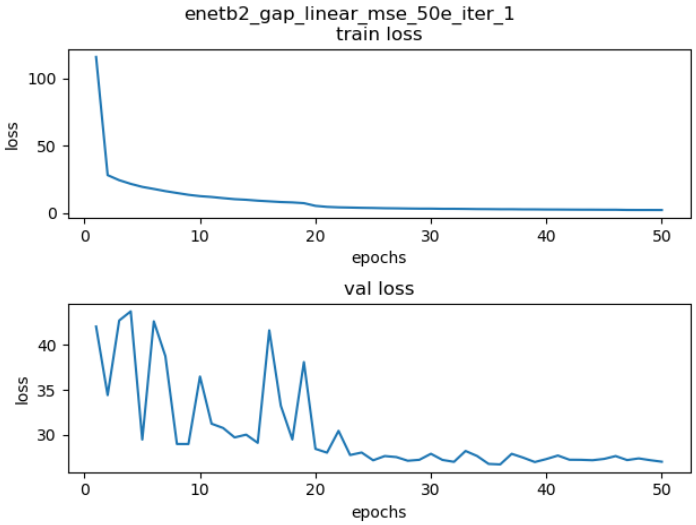


Figure 5.3: Training and validation loss as a function of the number of epochs.

The number of epochs was set to 50 and the batch size to 32.
The training process was stopped if the loss evaluated on the validation set did not

improve for 20 consecutive epochs. The learning rate was reduced by a factor of 0.1 if the validation loss did not improve for 10 successive epochs. The minimum learning rate was set to 10^{-7}. The training process took a total of 12.89 hours.

The original dataset was split into three datasets: 70% of images for the training set, 10% for the validation set and 20% for the test set. Our training set contained 57098 images, the validation set 8282 and the test set 16024 images.
The training set is needed to train the neural network and find the optimal parameters' values to get the best performance in our outcome prediction. The validation set is used to test the performance of the model fitted on the training set during the tuning of the parameters. Once the final model has been assessed, its final performance is tested on the test set.

After the training process of the neural network, the model performance was assessed on the test set and the patient's age was predicted for each retinal image. The average error, calculated on the absolute difference between the predicted age and the actual age, is 4.10 years. This result is consistent with the results of *Ghouse et all.* and the papers presented in Chapter 2 concerning age prediction.

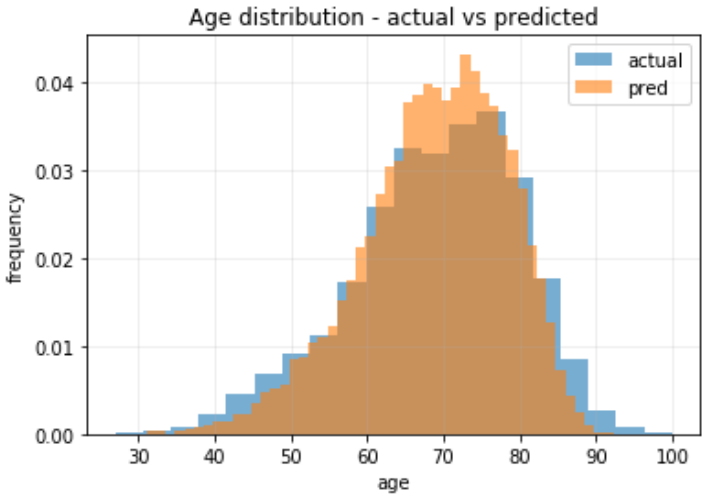
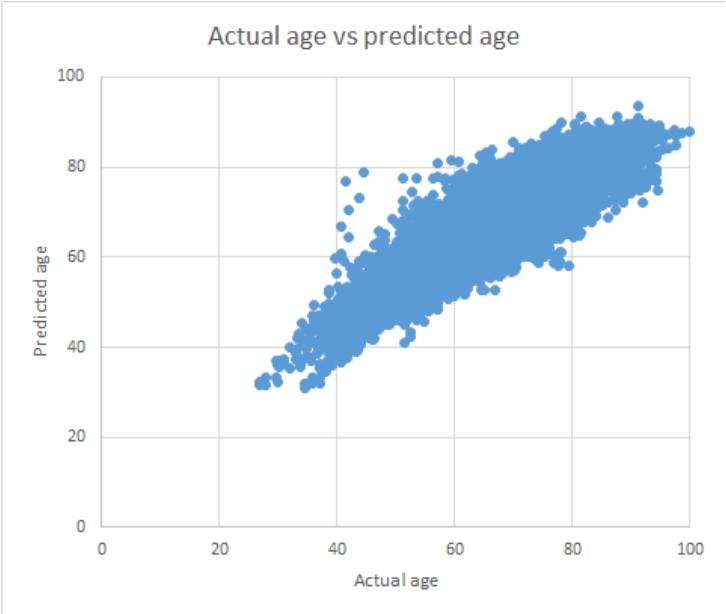


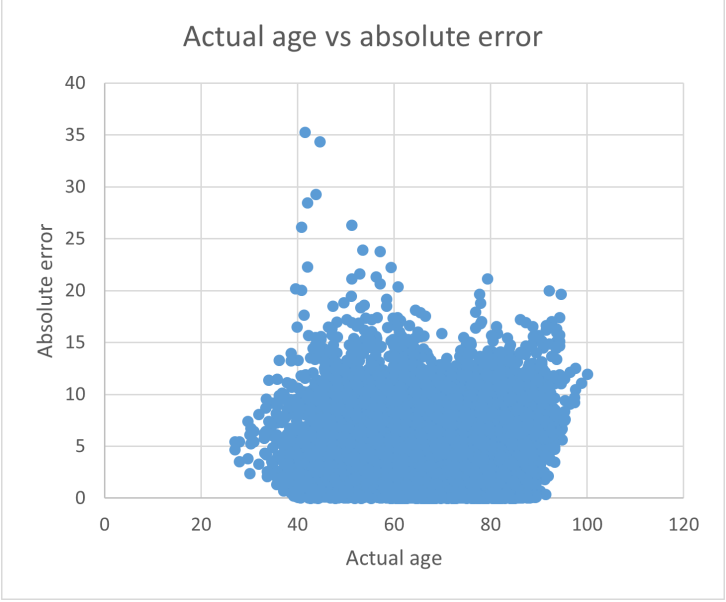
Figure 5.4: Distribution of the predicted and actual ages.

The figure above shows the distributions of the actual and predicted values which are almost equal. This means that the age has been accurately predicted.
In the figures below, we can see the actual age versus the predicted age and the

actual age versus the absolute error. The error is the absolute difference between the actual and the predicted age.



(a) Actual age vs predicted age.



(b) Actual age vs absolute error.

Figure 5.5: Plot of the actual slope versus the predicted age (a) and the absolute error (b).

As we can expect, there is a linear relationship between the actual age and the predicted age, in fact, the points lie around the bisector of the first and third quarters. On the contrary, we expect that the error is normally distributed and independent of the measurements, in fact, the points have a random distribution.

We can conclude that the Efficient Net B2 applied to retinal images can efficiently predict the age of a diabetic patient.

5.4 Haemoglobin prediction with neural network

Three experiments, whose main aim was predicting the exposure to glycated haemoglobin, were performed and are described in the following sections. The Efficient Net B2 network described in the previous chapter was used in all of them. The training conditions, such as the criterion for early stopping and the one for the reduction of the learning rate, were the same used in the experiment on age prediction, as well as the pre-processing process of the images.

The values of the cumulative glycated haemoglobin were provided by Dr Huan Wang, a statistician in the School of Medicine at the University of Dundee. A special thanks to him for his help and his provided data.

The cumulative value has been calculated as an approximation of the integral of all available values over time. The integral was approximated as a trapezoid sum. For those patients who do not have a haemoglobin measurement at the date of the first retinal image, we considered a constant value, thus the integral was approximated with a rectangle from the date of the first image to the date of the first haemoglobin measurement.

We started from a dataset with 100789 images and their corresponding cumulative haemoglobin value. All our experiments were carried out inside the Safe Haven environment to protect the patient's confidentiality and data's sensibility, and any of this information has been exported. All data exported and shown in the next sections do not contain sensitive data.

5.4.1 Experiment 1

This experiment aimed to predict the last cumulative glycated haemoglobin value from the first image available for the patient.
We aimed to understand if some regions detected by the neural network could predict the last cumulative haemoglobin value in the first retinal image taken for each patient. In other words, if we can predict a future cumulative value from a retinal image.

Despite the originality of the approach, we found that this is not possible. Retinal images do not store the information needed to calculate a future cumulative haemoglobin value.

This unsuccessful discovery was confirmed by Doctor Doney Alex, who confirmed that this approach contrasts with the approach of *Ghouse et al.*. In fact, they predicted the patient's age from his/her retinal image, where age is a piece of information dated at the exact moment of the image. On the contrary, we tried to use a retinal image to predict something that has not still happened but will happen in the future.

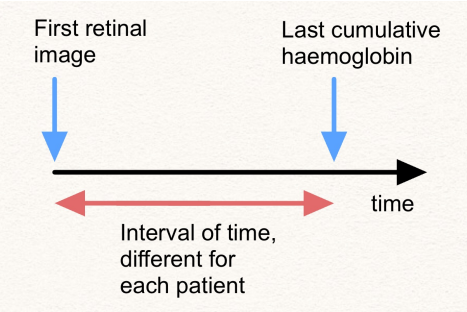


Figure 5.6: Schematization of the first experiment.

The main limitation of this experiment is that we do not even know when this potential cumulative haemoglobin value will happen in future. In our training set, we have different periods for each patient between the time when the first retinal image was taken and the time when the last cumulative haemoglobin value was calculated. Our neural network is impossible to learn to this interval because it does not depend on features or elements contained in the retinal image but relies on external motivations. This limitation is represented graphically above.

5.4.2 Experiment 2

The aim was to predict the cumulative baseline value of the glycated haemoglobin from the baseline retinal image.

Since we can not predict future values with our neural network, we performed a simplified experiment. We selected only the first image of each patient and trained the network to predict the corresponding cumulative haemoglobin value. In this case, we are trying to predict something that happens at the exact moment when the retinal picture is taken.

Once we understood that this experiment is correctly formulated, we moved to Experiment 3, in which we considered all available images. In fact, in Experiment 3 we did not reduce the size of the original dataset, as we did in Experiment 2.

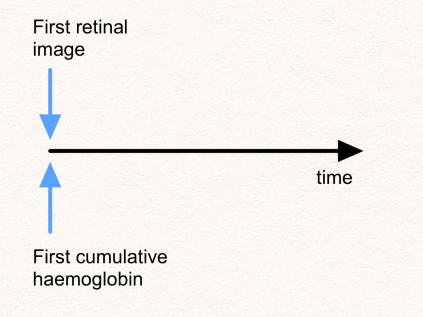


Figure 5.7: Schematization of the second experiment.

5.4.3 Experiment 3

The aim of this experiment was to predict the cumulative glycated haemoglobin value from all available images. We considered all available images, both left and right images, and tried to predict their corresponding cumulative haemoglobin value with our neural network.

The dataset was split into three datasets: the training set with 72388 images, the validation set with 7713 images and the test set with 20058 images.
This is the same approach and reasoning as Experiment 2. Still, rather than considering only the baseline image and value, we selected the couple image -

cumulative HbA1c value for each patient since we have the calculated cumulative glycated haemoglobin for each image.

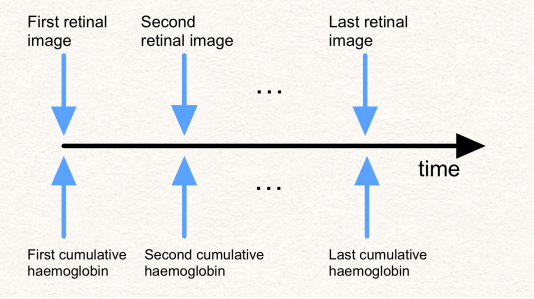


Figure 5.8: Schematization of the third experiment.

The number of epochs was set to 200. However, the early stopping criterion was met at epoch number 44. The training process took a total of 13.87 hours.

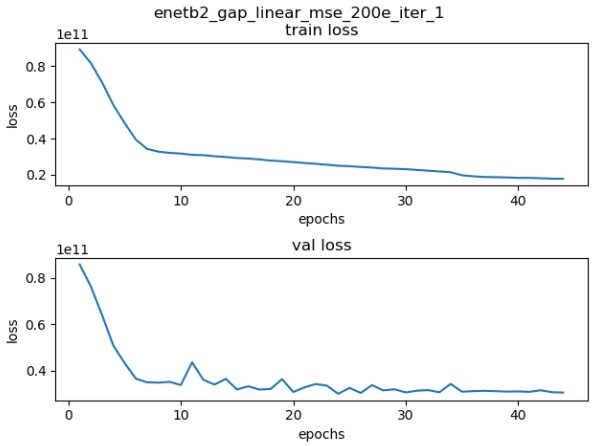
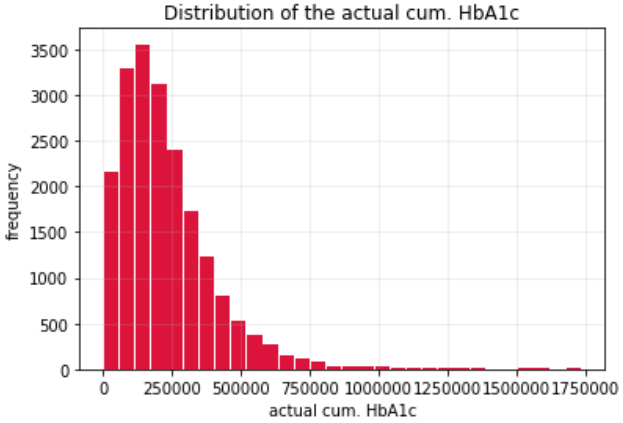


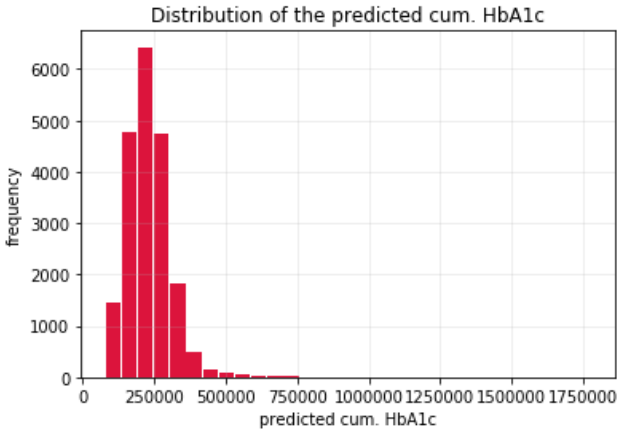
Figure 5.9: Training and validation loss as a function of the number of epochs.

Despite the promising results obtained with age prediction, the results of this experiment are not expected. The average absolute error, calculated as the absolute difference between the actual and the predicted value, is 117840.9 mmol/mol.

Fig.5.10 shows the distribution of the actual values and the one of the predicted values of cumulative HbA1c. The histogram of the predicted values is more compressed than the histogram of the actual values, indicating a substantial error in the prediction of the outcome.



(a) Distribution of the actual cumulative HbA1c.



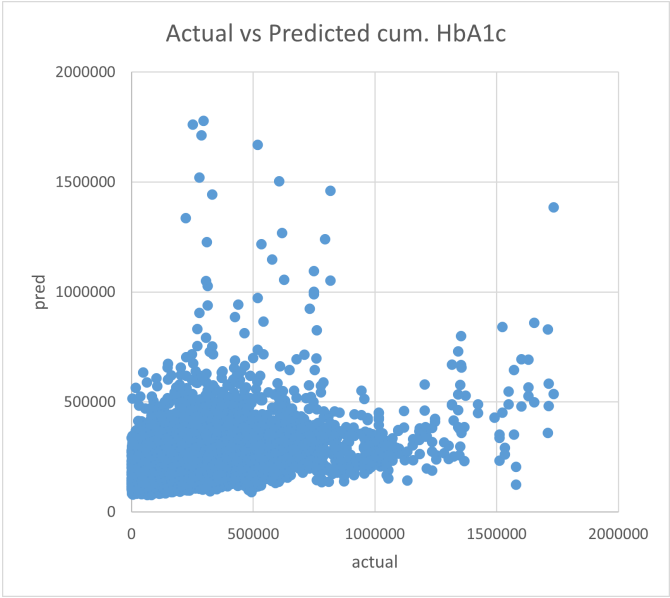
(b) Distribution of the predicted cumulative HbA1c.

Figure 5.10

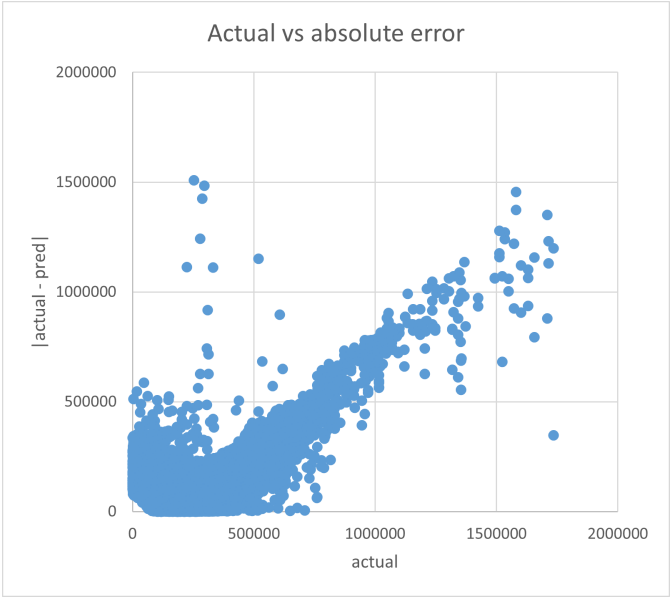
Fig.5.11 shows the predicted value and the absolute error versus the actual value. There is not a linear relationship qualitatively observable between the predicted and the actual values. However, we expected most of the points to lie around the first-third quarters' bisector.

From the second figure, it seems there is a linear relationship between the absolute error and the actual value of cumulative HbA1c. If so, this would imply that the error we made in the prediction depends on the amplitude of the actual cumulative haemoglobin, in contrast with what we supposed. In fact, we hypothesized that the error was normally distributed and independent of the measurements.

Looking at all graphs, we can note that there is, on average, a big error in the outcome predictions. We investigated it also by plotting some spaghetti plots.

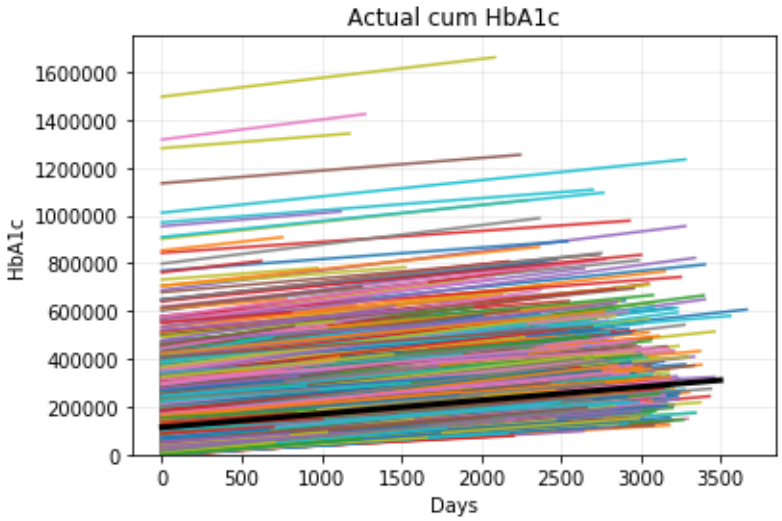


(a) Actual cumulative HbA1c vs predicted cumulative HbA1c.

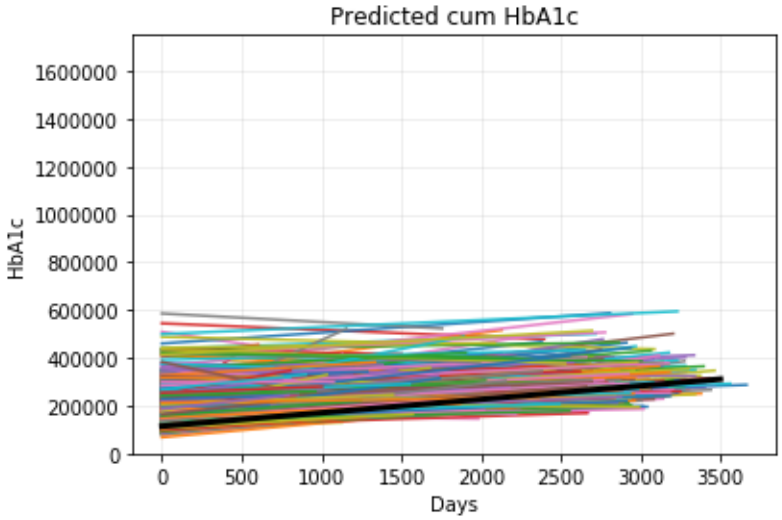


(b) Actual cumulative HbA1c vs absolute error.

Figure 5.11



(a)



(b)

Figure 5.12: Spaghetti plots of the actual (a) and predicted (b) values. Each line represents a different patient. The black line is the median line.

The different distribution of the predicted values can be easily detected from the "spaghetti plots" above. The predicted distribution is significantly more compressed than the actual one. One of the reasons could be a too wide dataset. In fact, the values of the cumulative HbA1c lie between some thousands and more than 1 million mmol/mol.

However, the median values are pretty similar and a big result is the increasing slope of the median predicted values. That means that, on average, the neural network was able to capture an increasing trend of the values of cumulative HbA1c.

We can conclude that this NN does not accurately estimate the single value of cumulative glycated haemoglobin but the increasing direction of the median trend's slope was correctly captured. The successive analysis described in Section 5.5 aims to investigate the trend of the predicted values for each patient and compare it with the trend of the actual values. We considered only patients who have more than three images, thus more than three values of cumulative HbA1c.

5.4.4 Left - Right division

For further information, we divided the dataset used in Experiment 3 into two datasets: the left one, which contains only the retinal images of the left eye, and the right one with only the right eye images. Then, we try to predict with the same neural network the value of the cumulative HbA1c for each image available in each dataset.

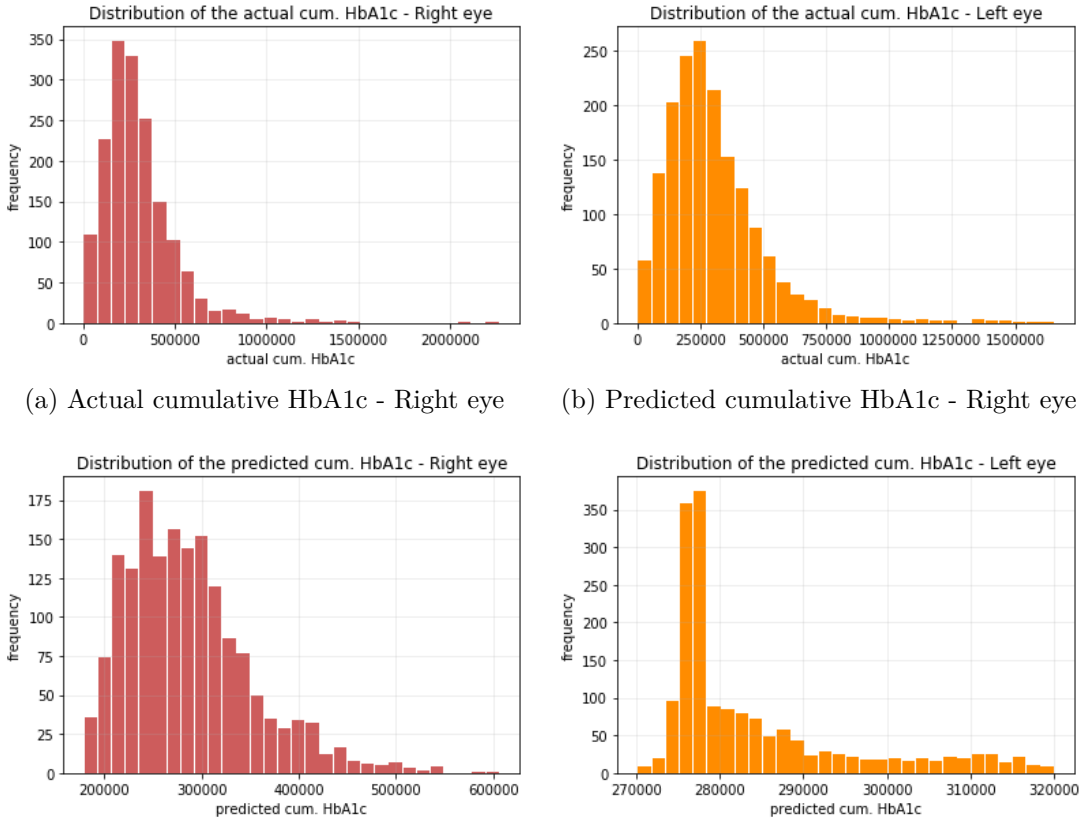
This left-right division aimed to understand if there were some significant differences between the left eye and the right eye. As we can see in the table below, the values are very similar and the average absolute error is slightly bigger than the average absolute error made in the predictions of the entire dataset. The higher error can be due to the smaller size of the dataset.

	Actual value $[mmol/mol]$	Predicted value $[mmol/mol]$	Average error $[mmol/mol]$
Left eye dataset	303075.2	284058.3	139975.2
Right eye dataset	301050.1	286834.2	137538.3

Table 5.1: Average values of the left eye and right eye experiments.

Fig.5.13 shows the distributions of the predicted values of the left and right eye and they are quite different from the actual ones. The bias, that was present in Experiment 3, is still present and the predicted values are quite far from the actual values.

We can conclude that there are no differences in using only right eyes or left eyes and we can give in input to our neural network both images indiscriminately.



(a) Actual cumulative HbA1c - Right eye

(b) Predicted cumulative HbA1c - Right eye

(c) Actual cumulative HbA1c - Left eye

(d) Predicted cumulative HbA1c - Left eye

Figure 5.13: Distributions of the actual and predicted values in the right-eye and left-eye datasets.

5.5 Trend of the cumulative HbA1c

In this step is to consider for each patient the set of the predicted values of cumulative HbA1c available for each image and calculate the trend of these values through a linear regression:

$$y = mx + q$$

where m is the slope of the interpolation line, q is the intercept with the y-axis, y is the predicted cumulative HbA1c, and x is the time evaluated in days between one measurement and the next.

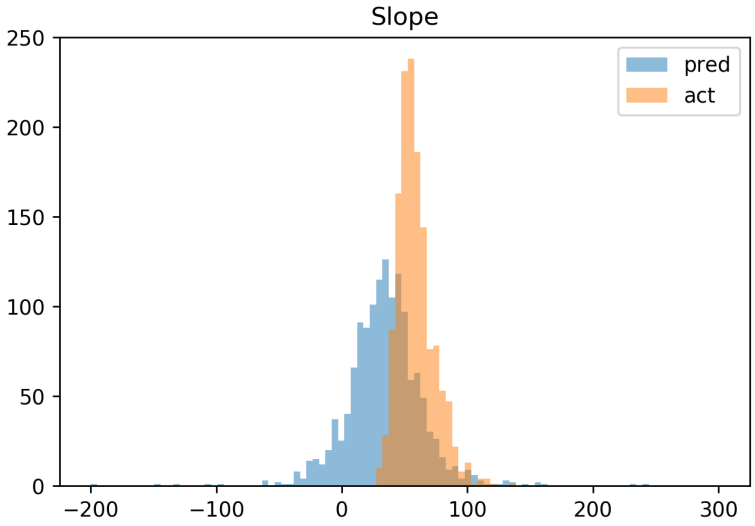
For each patient, we obtain the slope and the intercept of the line that interpolates the values of the predicted cumulative HbA1c and the slope and the intercept of the lines that interpolate the values of the true cumulative HbA1c. Thus, for each patient, we have two slopes, the actual and the predicted slopes, and two intercepts, the actual and the predicted ones, and we can compare them for further analysis.

The two slopes are very different, with the actual slope which is more than once and a half of the predicted slope. On the contrary, the predicted intercept is bigger than the actual one.

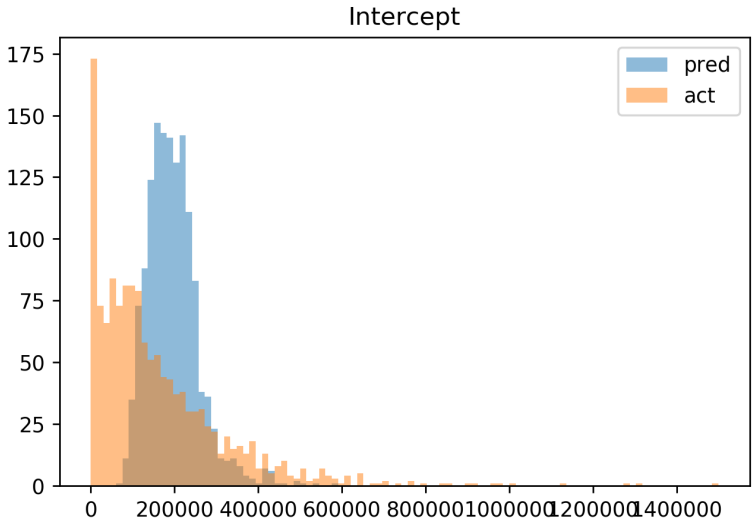
	slope $[mmol/mol/day]$	intercept $[mmol/mol]$
interpolation line of the actual values	58.86	158917.3
interpolation line of the predicted values	34.71	196364.3

Table 5.2: Average values of the slopes and intercepts of the interpolation lines.

In Fig.5.14, we can see the slopes' and intercepts' histograms of the actual and the predicted values. The histogram of the predicted slopes is wider, and its average value is lower than the average of the actual slopes, a signal of an underestimation of the daily increase. In fact, the slope of the interpolation line can be interpreted as the average increase of cumulative HbA1c per day. On the contrary, the histogram of the predicted intercepts is tighter, and its average value is right-shifted, which means that, on average, the initial value of the predicted line is higher than the initial value of the actual line.



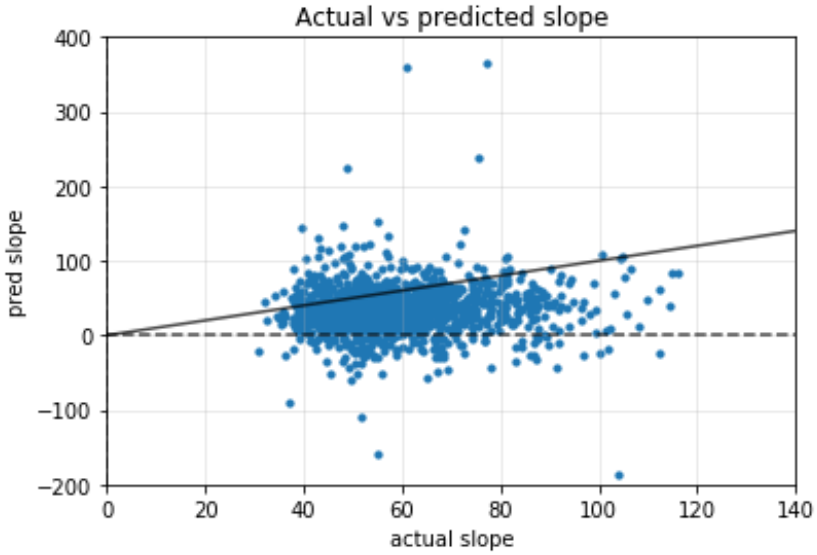
(a) Slopes of the actual and predicted cum. HbA1c.



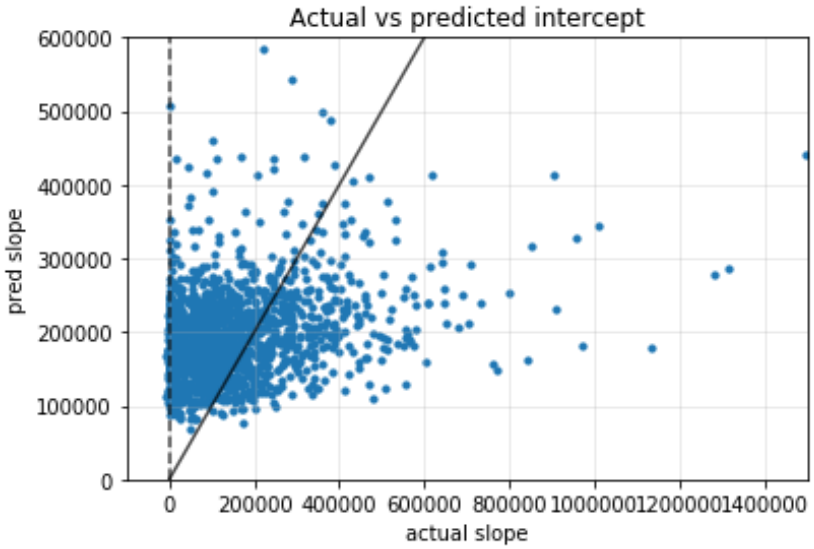
(b) Intercepts of the actual and predicted cum. HbA1c.

Figure 5.14: Distributions of the slopes and intercepts.

Note that with the expression *"predicted line"* we mean the line which interpolates the predicted values of cumulative glycated haemoglobin, while the *"actual line"* is the one which interpolates the actual values.



(a) Actual slope vs predicted slope.



(b) Actual intercept vs predicted intercept.

Figure 5.15: Plot of the actual slope and intercept versus the predicted slope and intercept. The black line is the bisector of the first-third quarter.

Fig.5.15 shows two plots: the actual slopes versus the predicted slopes and the actual intercepts vs the predicted intercepts. We expect most of the points to lie around the bisector of the first-third quarters, but, as we can see, this does not happen. Moreover, the predicted intercepts are always positive, while some of the actual intercepts are actually negative. This fact implies that the average predicted line starts from a higher initial value of cumulative HbA1c than the actual line but has a lower slope, so it rises slower.

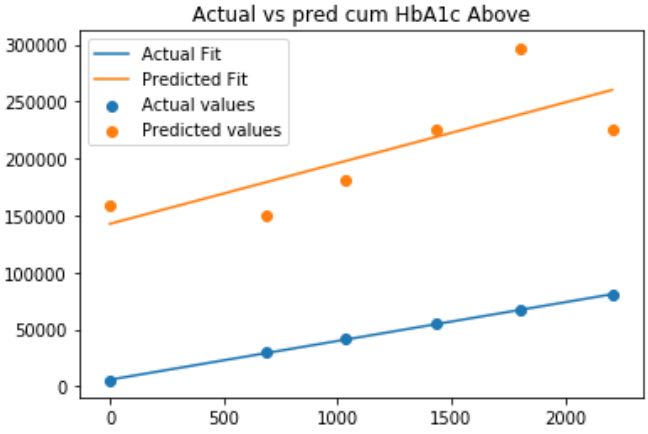
5.5.1 Classification of the position

The following step was to study if there is any association between a higher predicted trend of cumulative HbA1c and an outcome of medical interest, such as a MACE event.

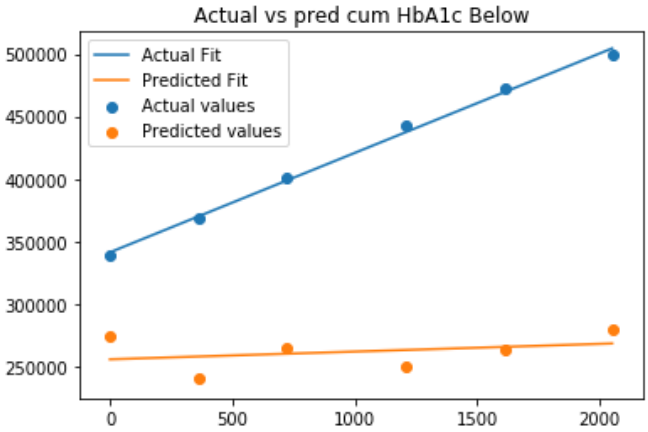
To study these potential associations we divided our dataset into three classes:

- above: the predicted interpolation line lies entirely above the actual interpolation line, such that for that patient the predicted trend is higher;
- below: the predicted interpolation line lies entirely below the actual interpolation line;
- intersection: there is an intersection between the two lines.

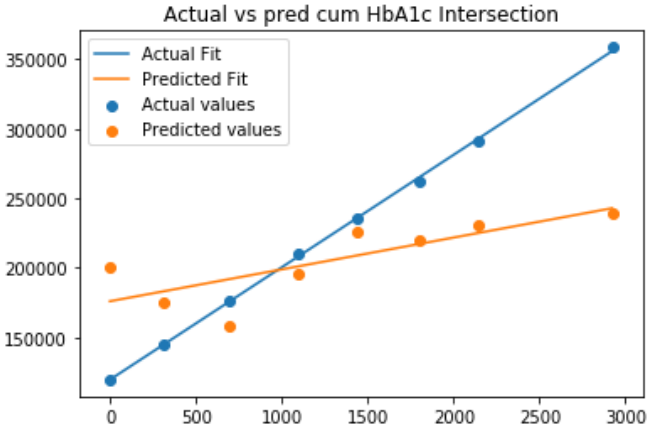
Three examples of the class divisions which has just been explained can be appreciated in Fig.5.16. For example, the first image is classified as *above* because the line that interpolates the predicted values (orange line) is above each point to the line that interpolates the actual values (blue line).



(a) Example of an above classification.



(b) Example of a below classification.



(c) Example of an intersection classification.

Figure 5.16: Example of the interpolation line's classification.

5.6 Risk assessment

This analysis aims to understand if there is an association between an above-predicted line and a higher probability of experiencing a MACE or death due to cardiovascular disease. The idea is that an above-predicted line is synonymous, for example, with a higher risk of mortality, i.e. those patients who will die have some characteristics in their retina which have been detected with the images and then are recognised by the neural network which predicts higher values as a warning. Starting from this idea, we study three possible associations described in the following sections: the risk of mortality, the risk of death due to a CVD and the risk of a MACE event.

5.6.1 Risk of mortality

We aim to understand if a higher predicted line implies a higher risk of mortality. We want to demonstrate that if a patient has a predicted line which is classified as above, he has a higher risk of dying than a patient whose predicted interpolation line is below the actual interpolation line.

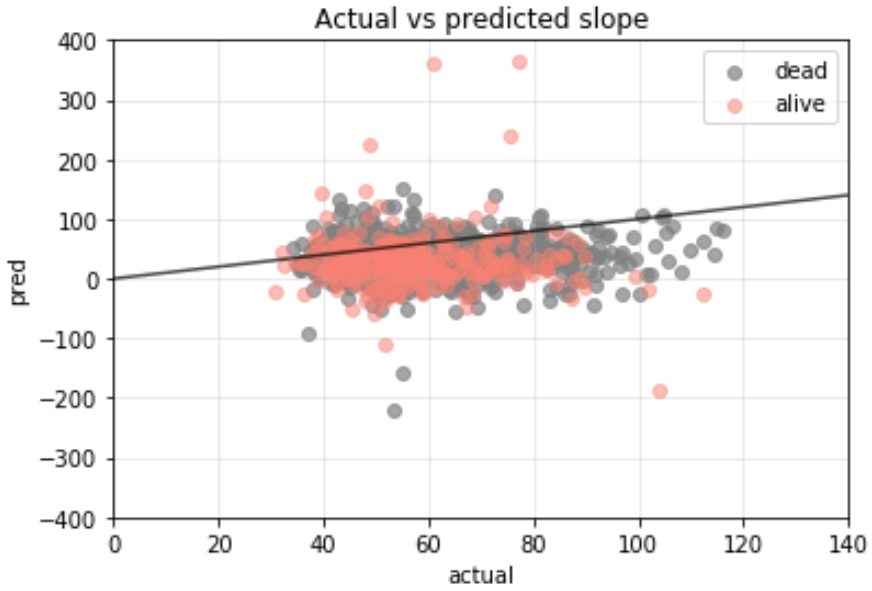


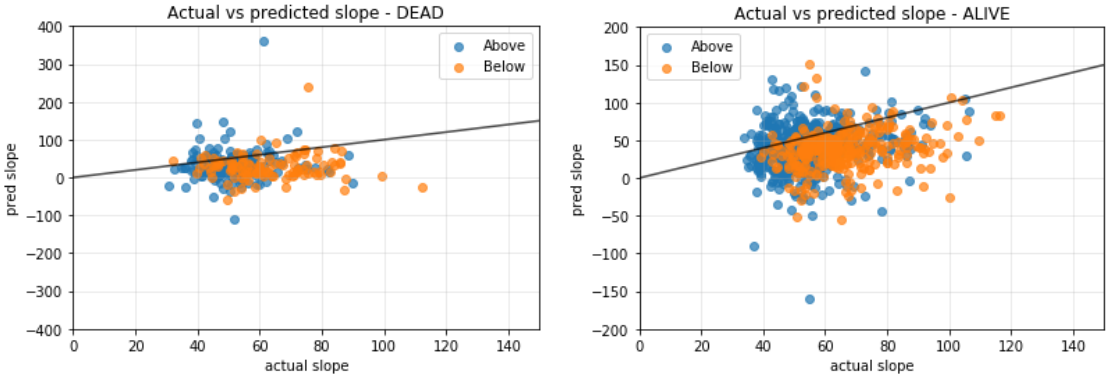
Figure 5.17: Actual versus predicted slope in the alive and dead datasets. The black line is the bisector of the first-third quarter.

Firstly, we tried to understand if people who died and those who did are distributed differently. In particular, we plot them with two different colours (salmon = alive patients, dark grey = dead patients) to see if there are some differences in the distributions of points (Fig.5.17).

To go into more detail, we consider two datasets: one with all the patients alive in our original datasets and one with the dead patients. Then we calculate how many dead patients have a predicted line above the actual line to see if this number is significant to the total number of dead.

	Above	Perc.	Below	Perc.	Intersection	Perc.	Total
Alive	585	54.1%	299	27.6%	198	18.3%	1082
Dead	155	49.5%	109	34.8%	49	15.7%	313

Table 5.3: Count of the above, below and intersecting lines in the alive and dead datasets.



(a) Actual vs predicted slope in dead people. (b) Actual vs predicted slope in alive people.

Figure 5.18: Plot of the actual versus the predicted slope in the dead-people and alive-people datasets.

Contrary to our hypothesis, most of the patients in both datasets have an above line, i.e. the interpolation line of the predicted values is at each point higher than the interpolation line of the actual values. Surprisingly, the percentage of alive people, whose predicted lines are above the actual lines, is slightly bigger than the one of dead people. However, this can be due to the small size of the datasets.

Also looking at their distribution, we can not appreciate any significant difference between the above and below lines. In fact, the distribution of the below slopes is a little right-shifted, meaning that, on average, high values of the actual slope are predicted smaller.

Therefore, we can not use the classification of the position of the interpolation line as an instrument to establish who has a higher risk of mortality. It seems that an above-predicted line is not correlated with death and, thus, the mortality risk.

For further information, we investigate the male and female distributions in the two datasets. We wonder if a patient is more likely to die than a patient of the opposite sex. Moreover, if males and females are equally distributed or if there is an under or overestimation for one of the two sex.

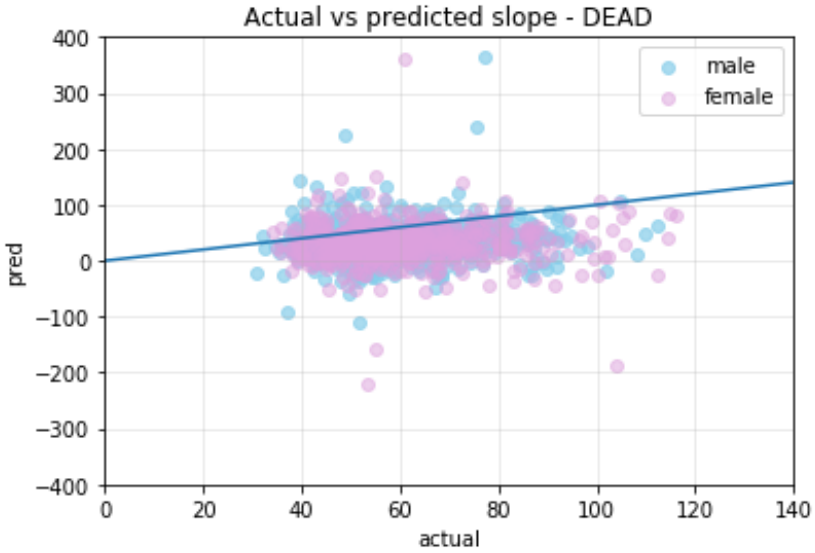


Figure 5.19: Actual versus predicted slope in males and females of the dead dataset. The blue line is the bisector of the first-third quarter.

	Males	Percentage	Females	Percentage	Total
Alive	574	75.2%	508	80.4%	1082
Dead	189	24.7%	124	19.6%	313
Total	763		632		1395

Table 5.4: Count of the males and females in the two datasets.

Males have a slightly higher probability of dying than females, however, the numerousness of the two datasets is too small to draw any conclusion.
Looking at Fig.5.19, we can not assess that sex is discriminating in dead people because males and females have pretty similar distributions.

The following step considers only the death due to cardiovascular disease. In fact, considering all the deaths, which can be happened to different causes, can be a too broad domain of investigation. Moreover, we know that the retina stores important information about the health of the cardiovascular system, so investigating the association between cumulative HbA1c and CVD death could be encouraging.

5.6.2 Risk of CV death

In our dataset, we have a variable whose name is *CVdeath* which indicates if one of the first three causes of death of a patient is a cardiovascular problem. We aim to study if CVdeath is associated with a higher predicted slope. In this study we are considering only died patients.
The first step is visualizing the distribution of people who died due to CVD and the one of people who died due to different causes.

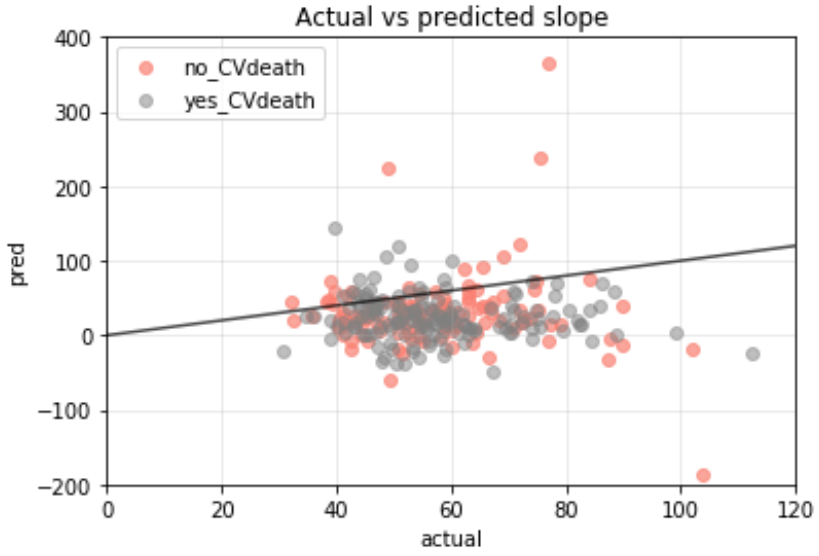
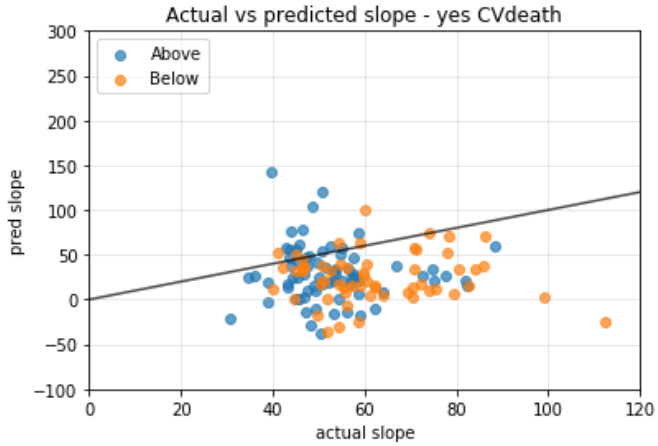


Figure 5.20: Actual versus predicted slope in two datasets: no death due to a CVD, yes death due to a CVD.

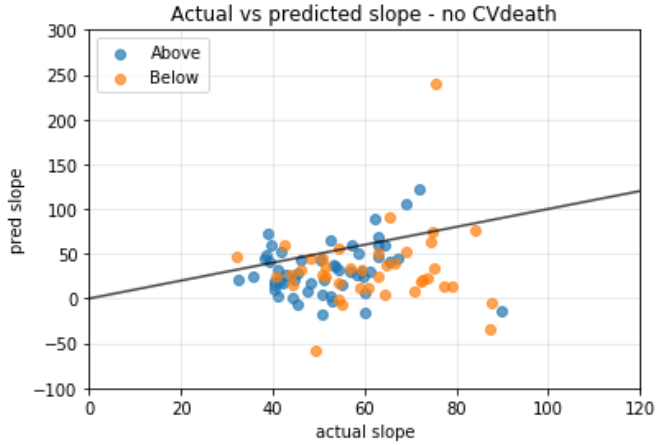
Then we divided our dataset into two subgroups: the patients whose one of the first causes of death is a cardiovascular disease and the patients who died due to different causes. Then we count how many predicted lines are classified above, below and intersecting the actual lines.

	Above	Perc.	Below	Perc.	Intersection	Perc.	Total
yes CVdeath	72	47.1%	58	37.9%	23	15.0%	153
no CVdeath	53	45.3%	40	34.2%	24	20.5%	117

Table 5.5: Count of the above, below and intersecting lines in the two datasets.



(a) Actual vs predicted slope in the yes-Cvdeath dataset.

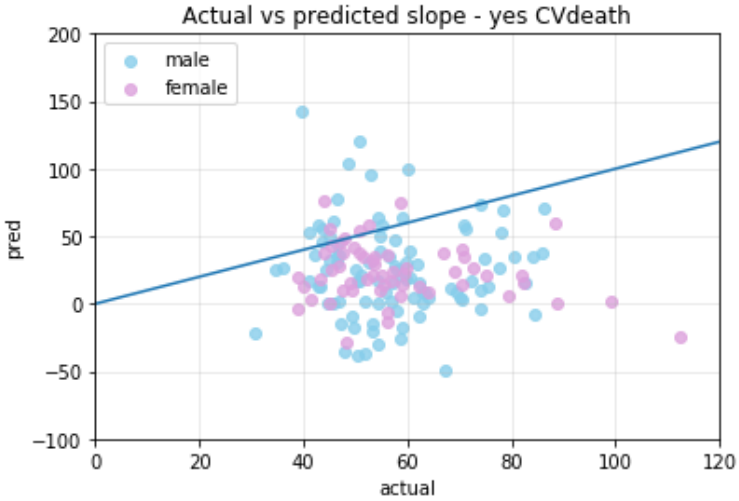


(b) Actual vs predicted slope in the no-CVdeath dataset.

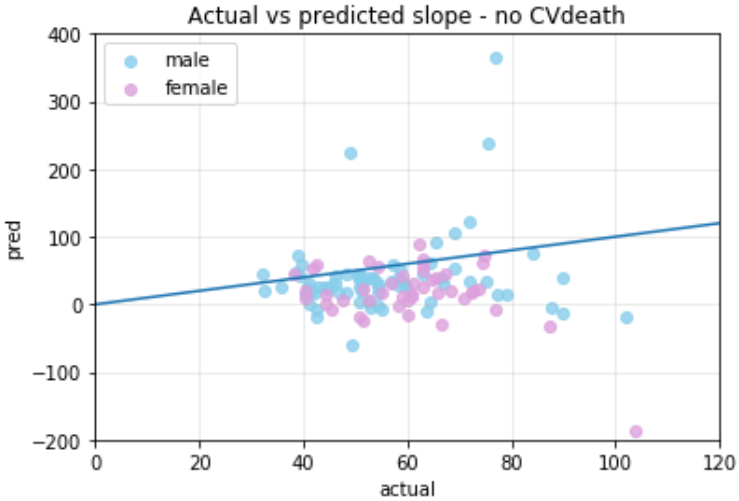
Figure 5.21: Plot of the actual versus predicted slope in the two datasets.

Both datasets have the same percentage of the above-predicted lines. Moreover, looking at their distribution, we can not assess that there are differences in the distributions of the above- and below-lines, as we can see in Fig.5.21.

As done before, we investigate the male and female distributions in the two datasets. We wonder if a patient is more likely to die due to cardiovascular disease than a patient of the opposite sex.



(a) Male and females distribution in the yes-Cvdeath dataset.



(b) Male and females distribution in the no-CVdeath dataset.

Figure 5.22: Plot of the actual versus predicted slope in the two datasets.

Looking at Fig.5.22, we can not assess that sex is discriminating both in the yes-CVdeath and no-CVdeath datasets because males and females have pretty similar distributions. In both plots, we can note some outliers blue points (male patients), however, there are too few points to draw any conclusion or association.

	Males	Percentage	Females	Percentage	Total
yes CVdeath	95	56.9%	58	56.3%	153
no CVdeath	72	43.1%	45	43.7%	117
total	167		103		270

Table 5.6: Count of the males and females in the two datasets.

Males have a higher probability than females of having CVD as one of the first three causes of death. However, even in this study, the numerousness of the two datasets is very small.

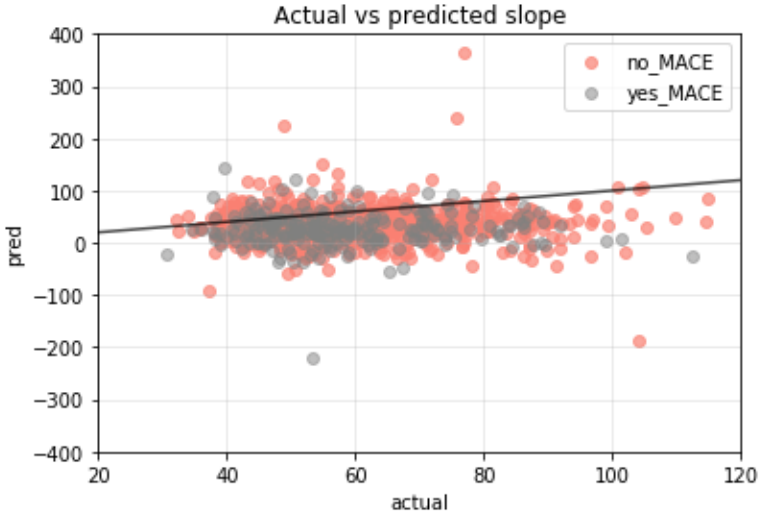
We can conclude the data we have do not show an association between an above-predicted line and death due to cardiovascular disease and that males and females seem to have the same probability of dying due to CVD. However, it is essential to note the numerousness of the dataset, which is very small.

Since we are considering broad causes of death, in fact, one of the first three is the cardiovascular disease but the other two can actually be very different, as the last step, we investigate a narrow outcome, which is the MACE. Thus, we repeated the studies that we have done till now about the risk of mortality and death due to CV, but now with the risk of experiencing a MACE.

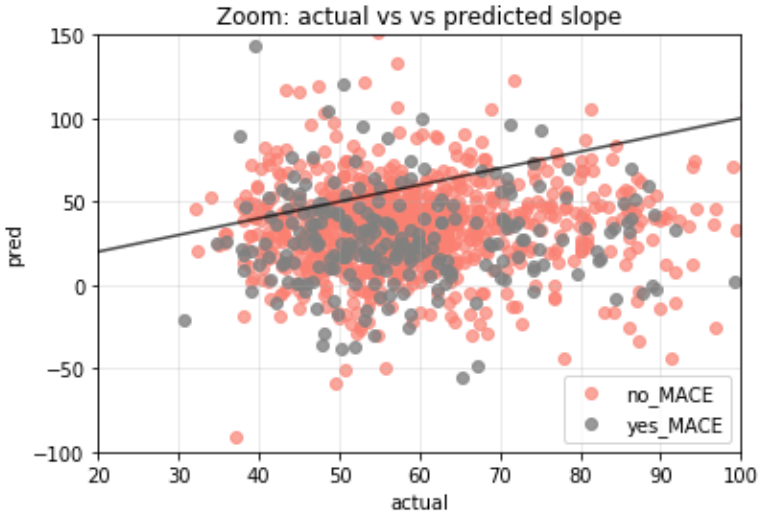
5.6.3 Risk of MACE

A major adverse cardiovascular event (MACE) is a cardiovascular event that can be a nonfatal stroke, a nonfatal myocardial infarction and cardiovascular death. It is one of the main outcomes of interest in the cardiovascular domain.

In this study, we aim to understand a potential association between an above-predicted line and the risk of experiencing a MACE. As we can see in Fig.5.23, we can not appreciate big differences between the two groups of patients.



(a)



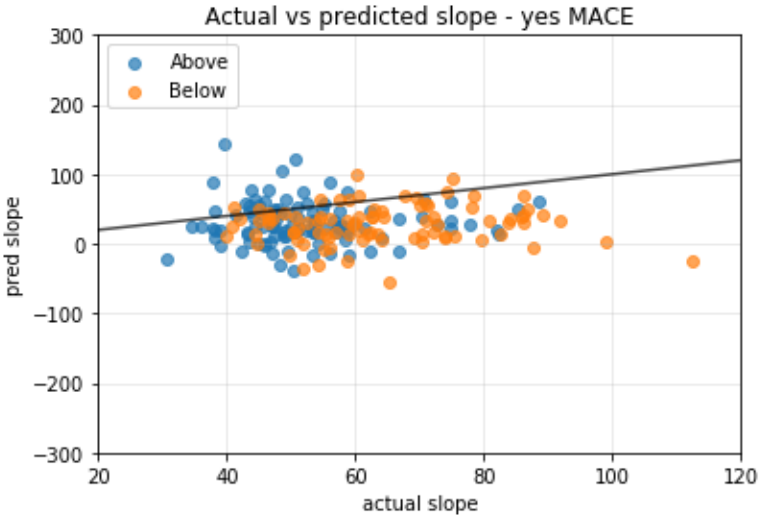
(b) Zoom of the previous plot.

Figure 5.23: Actual vs predicted slope highlighting with two different colours the two datasets.

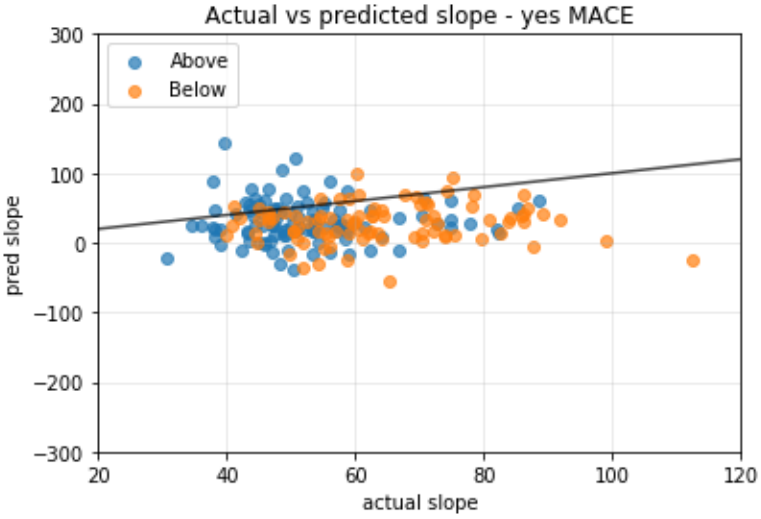
In Table 5.7, we can see the number of people with and without a MACE, whose interpolation predicted lines are classified as above, below or intersected. Contrary to our hypothesis, most of the people in the no-MACE dataset have an above-predicted line. However, the datasets are too small to draw any conclusion.

	Above	Perc.	Below	Perc.	Intersection	Perc.	Total
yes MACE	105	44.7%	89	37.8%	41	17.5%	235
no MACE	465	53.5%	240	27.6%	165	18.9%	870

Table 5.7: Count of the above, below and intersecting lines in the no-MACE and yes-MACE datasets.



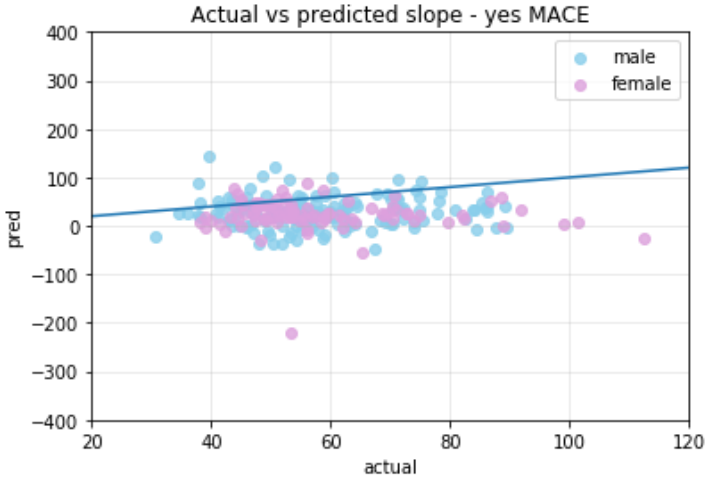
(a) Position of the interpolation line in people with no MACE.



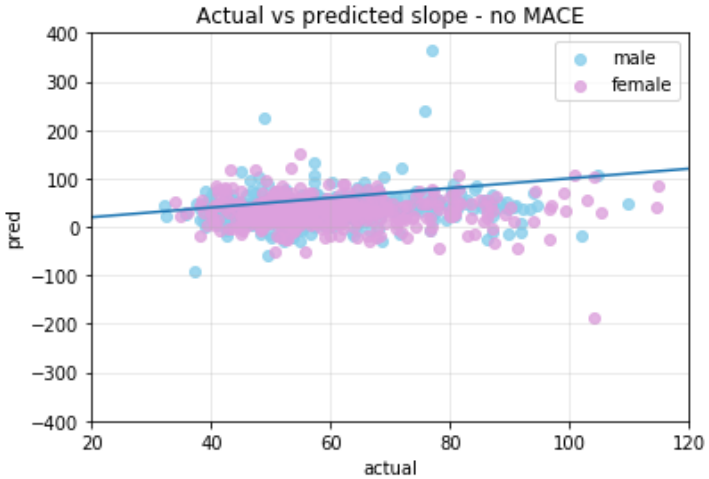
(b) Position of the interpolation line in people with a MACE.

Figure 5.24

Furthermore, we investigated if males and females have different distributions in the two datasets. We wondered if experiencing a MACE is more probable in a male or female patient and if sex discriminates in assessing a higher predicted slope in one of the two MACE groups. However, as shown in the table below, experiencing a MACE is equally probable in male and female patients. Males have a slightly higher probability, but the total numbers of patients are too small to conclude a higher risk of a MACE. The distributions of males and females in both datasets are similar and indistinguishable, as shown in Fig.5.25.



(a) Males and females distribution in the yes-MACE dataset.



(b) Males and females distribution in the no-MACE dataset.

Figure 5.25

	Males	Percentage	Females	Percentage	Total
yes MACE	141	23.4%	94	18.7%	235
no MACE	462	76.6%	408	81.3%	870
total	603		502		1105

Table 5.8: Count of the males and females in the two datasets.

We can conclude that the position of the predicted line with respect to the actual line and the sex of the patient are not significant in assessing a higher risk of experiencing a MACE.

5.7 Conclusion

In this chapter, we presented the experiments and the main achieved results. Firstly, we performed feature selection on a subset of the GoDARTS dataset to understand the most important features in predicting cardiovascular failure. We applied Lasso regression, reinforced by a bootstrap resampling to assess the stability of the selection. Some of the available clinical features, such as the patient's age, history of previous cardiovascular disease and the age of dementia diagnosis, were selected in each bootstrap iteration. These were the most significant features in outcome prediction. However, some retinal features, which had been extracted from retinal images with the software VAMPIRE, were selected with a percentage higher than 80%, suggesting that they are important too in outcome prediction.

The first step inside the Safe Haven environment aimed to predict the age of the patients from their retinal images. We wanted to reproduce the results achieved by *Ghouse et al.*[9] in order to familiarise ourselves with the environment, the data and the code. Moreover, we had the opportunity to check the functioning of the Efficient Net B2, the neural network we used in this experiment and the next ones. Age was accurately predicted with an average error of 4.10 years, a result consistent with the results obtained before in age prediction.

The following step aimed to predict the cumulative glycated haemoglobin with the Efficient Net B2 neural network from retinal images. We performed three different experiments to understand in depth which was the optimal strategy to proceed

and interpret our data and results. We considered the final results obtained in the third experiment, which aimed to predict the cumulative HbA1c for each image available. However, the neural network could not accurately predict the outcome of interest: the average error was around 117840 mmol/mol. We could see how much the distributions of actual and predicted values differentiate by looking at their plot.
Moreover, we proved that right and left eye images did not produce different predictions, thus, they can be used as input of our neural network indiscriminately.

Since a single-value estimation could not be obtained, we considered for each patient the set of images available and their predicted values. We interpolated them with linear regression, particularly considering the slope and the intercept of this interpolation line. We did the same for the actual values of cumulative HbA1c. We compared the two slopes and intercepts, but they are actually quite different.

We tried to elaborate on these results by interpreting an above-predicted line as a higher risk of developing a specific outcome. We investigated the risk of mortality, dying due to cardiovascular disease and experiencing a MACE. Despite the promising approach and hypothesis, we could not draw any conclusion, mainly due to the limited numerousness of available datasets.
When we studied the association between a higher predicted slope and the risk of mortality, we noted that most both alive and dead people have a predicted line which is classified above the actual line. The same result was observed when we studied the risk of dying due to cardiovascular disease and the risk of MACE. Moreover, no significant differences were evaluated between males and females, whose distributions were investigated in all the risk assessment tasks.

Therefore, we can conclude that our neural network can not accurately predict the cumulative glycated haemoglobin from retinal images. Moreover, that the predicted trend for each patient is quite different from his/her actual trend and that the position of the predicted line can not be exploited as an instrument in risk assessment. However, we proved that left and right eyes do not produce different estimates and that males and females of our dataset have the same probability of dying due to cardiovascular disease and experiencing a MACE.

Chapter 6

Conclusions

6.1 Our work

This research thesis concerns the applications of a neural network to retinal images to predict cumulative glycated haemoglobin. In this section, we will briefly summarize it.

We used the GoDARTS dataset, which contains patients who live in the Tayside region in Scotland and suffer from type 2 diabetes. Because of their disease and risks, they were offered retinal screening to monitor and diagnose diabetic retinopathy, a pathology affecting the eye's blood vessels that may cause irreversible blindness in diabetic people. The dataset contains 102,082 images taken between 2006 and 2016 and stored at the Scottish National Diabetes Eye Screening service. The images are colour fundus photography, which is a non-invasive and fast imaging technique that uses a fundus camera to record the interior surface of the eye. Before entering the neural network, the images are pre-processed in order to reduce the effect of image variations and create a set of images similar in size, colour and intensity.

During the first months spent at the University of Dundee, in order to familiarize ourselves with the dataset, we performed a feature selection on the GoDARTS dataset to understand which features were the most important in the prediction of cardiovascular failure. Our dataset contained both clinical features and retinal features, which have been extracted from the retinal images with the software VAMPIRE developed by the University of Dundee and the University of Edinburgh.

We confirmed the already existing results in the literature, id est that the most important features in the prediction of a CV failure are the clinical ones. However, retinal features have been selected with a high percentage in the bootstrap analysis, suggesting their importance in outcome prediction. After finishing this experiment, we moved inside the Safe Haven environment to implement the cornerstone of this thesis: the application of a neural network to retinal images of diabetic patients to predict the cumulative glycated haemoglobin.

We used the Efficient Net B2, a convolutional neural network widely used in image recognition and classification tasks. It allows reaching higher performance with a significantly lower number of parameters in comparison with the other ConvNets. Eight models of Efficient exist, from B0 to B7. We chose the B2 because it was used by *Ghouse et all.* in their work, with which they obtain excellent performance in predicting the age of a patient from retinal images of the same GoDARTS dataset. The training process was made with the Nesterov-accelerated Adaptive Momentum Estimation algorithm, an algorithm based on the stochastic gradient descent which adapts the learning rate at each parameters update and makes use of a momentum to reduce the oscillations and fasten the reaching of the solution.

All the experiments were carried out inside the Safe Haven environment, which is a virtual desktop environment created by the Health Informatics Centre of the University of Dundee. It allows the safe use of research data by regulating their use and protecting the patient's identity. Only data that do not contain sensitive information and patient-level data, therefore all the images and data shown in this thesis, can be extracted from the Safe Haven and used by the user in their report and further analysis.

In the first experiment, we tried to predict from the first retinal image available for each patient his/her last value of cumulative glycated haemoglobin. The idea was to use the first image to predict a future value which represents the exposure to HbA1c. However, that is not possible and this approach contrasts with the already existing approaches to the prediction of some medical outcomes of interest.

Therefore, we moved to Experiment 2, in which we predict the baseline value from the baseline image, and to Experiment 3. It concerned the prediction of the cumulative glycated haemoglobin from each available image. We trained the neural

network by proving the retinal image and the corresponding cumulative HbA1c value, which has been calculated as a trapezoidal sum as an approximation of the integral of the glycated haemoglobin curve.

6.2 Key achievements

The third experiment concerned the prediction of the cumulative HbA1c for each available image. However, the performance obtained with the Efficient Net-B2 on the test set is not the expected one. The distribution of the actual and predicted values are quite different, in particular, the range of the predicted values is significantly narrower than the actual range. Moreover, it seems that the error is linearly dependent on the actual measurement of HbA1c.

Nevertheless the discouraging results, we performed the same experiment on two different subsets of data: the left-eye and the right-eye datasets. We noted that the average predicted values for all right eyes and all left eyes are very similar (303075.2 versus 301050.1 mmol/mol). We demonstrated that in this experiment left eye and right eye images can be given indistinguishably as input of the Efficient Net-B2 in predicting the cumulative glycated haemoglobin.

The further analysis concerned the trend of the cumulative HbA1c for each patient. We considered the predicted values and the actual values for each patient and interpolated them with two lines. Then we considered their slopes and intercepts and compared them to see if the average trend for each patient was correctly captured by the neural network. The average predicted slope is smaller than the actual slope (34.71 versus 58.86 mmol/mol/day), indicating a lower increase in the cumulative HbA1c. However, the intercept is bigger (158917 versus 196364 mmol/mol) indicating a predicted baseline value higher than the actual one. We saw also that the distributions of the actual and predicted slopes and the ones of the actual and predicted intercepts are quite different.

As the last step, we considered the position of the predicted line with respect to the actual line. We aimed to study if an above-predicted line was associated with a higher risk of mortality, death due to cardiovascular disease or experiencing a MACE. In all our qualitative analyses, we did not notice significant differences

in the positions of the lines between the group that experienced the outcome of interest (e.g. yes MACE) and the control group (e.g. no MACE). Moreover, we investigated the distribution of the two groups with respect to the slope, i.e. if a group had on average a higher predicted slope than its control group, but we did not find significant differences.

For each task of risk assessment, we studied if there were differences between males and females. Thus, we investigated if males have a higher risk of mortality, death due to cardiovascular disease or experiencing a MACE than females and vice versa. We noted that the two sexes have the same distributions in the actual versus predicted slope plots in all three subtasks.

6.3 Limitations and future works

Despite the originality and innovation of this research work, it has some limitations that are presented in this section.

First of all, the GoDARTS dataset used in our experiments and analysis contains only patients suffering from type 2 diabetes. Diabetic patients have some characteristics which make them different from healthy people and they can be detectable from retinal images. In fact, their cardiovascular system and their retina too can be damaged by diabetes and the human retina has been proven to store important information about a patient's cardiovascular health. One possible future work can be repeating the prediction of the cumulative glycated haemoglobin from retinal images in a healthy population to evaluate if there are differences between non-diabetic and diabetic retinas. Moreover, the model performance can be assessed and compared with the aim to see if there is an improvement in the predictions.

The population of the GoDARTS dataset is a Scotland population living in the Tayside region. Habits and genetic factors, as well as geography and lifestyle, can influence the retinal images of our patients. Therefore, the ability to generalize this model can not be assessed now. Different populations should be studied and the neural network should be trained with their retinal images to assess the generalization ability of the model and investigate potential differences between various populations around the world.

Another limitation is the small numerousness of the dataset. Of the entire initial dataset, we choose 20% of the data for the test set, reducing significantly the initial numerousness. This choice was fundamental because we need to train the neural network on a large training set. As a consequence, we will assess the performance of the model on a small test set. All the analyses carried out in the previous chapter were performed on tiny datasets.

In our experiments, we used the Efficient Net B2 neural network. We choose it because *Ghouse et all.* [9] demonstrated on the same GoDARTS dataset that age can be accurately predicted from retinal images. Therefore, we use the same neural network with the same characteristic, e.g. the number of layers and number of neurons per layer, on the same retinal images to try to predict the cumulative glycated haemoglobin. Future improvements could concern the tuning and modification of the hyper-parameters of this neural network to improve the performance in the outcome prediction. The use of different neural networks could be explored too in order to find the optimal model for the prediction of the cumulative HbA1c.

Moreover, a Grad-CAM algorithm can be applied to our third experiment to investigate which are the regions of the retinal image on which the neural network focuses more during the training process. It would be very interesting to understand which parts of the human retina are more critical in predicting exposure to glycated haemoglobin. A further step could be studying if there is any association between these regions and the regions more affected by diabetic retinopathy.

Moreover, in the GoDARTS dataset, more than one image per patient is available. In the first phase of our experiment, when we predicted the cumulative HbA1c from retinal images, we gave one image at a time as input to our neural network, neglecting the fact that multiple images belonged to the same patient. Only when we calculated the actual and the predicted trend, we put together the values of cumulative glycated haemoglobin of the same patient. We did not let the neural network exploit this information during the training process. Future work can explore the possibility of using all the images available for the same patients as input, for example, by concatenating them or making a collage.

The team CVIP of the University of Dundee, with whom we worked for the development of this thesis, has been the first to explore neural networks applied to retinal images to predict the cumulative glycated haemoglobin. No one did this experiment before. Despite the limitations just presented, it is greatly innovative and deserves future works and implementations in order to find the optimal model able to predict accurately the cumulative HbA1c. It could become a beneficial instrument for the evaluation of the cardiovascular risk in non- and diabetic patients, useful for doctors and healthcare systems.

Bibliography

[1] Italian society of diabetology. [Online]. Available: http://www.siditalia.it

[2] The global diabetes community. [Online]. Available: https://www.diabetes.co.uk

[3] W. H. Organization, *Use of Glycated Haemoglobin (HbA1c) in the Diagnosis of Diabetes Mellitus*, 2011.

[4] J. Zhu, E. Zhang, and K. Del Rio-Tsonis, "Eye anatomy," November 2012.

[5] H. Kolb, R. Nelson, E. Fernandex, and B. Jones. The organization of the retina and visual system. [Online]. Available: https://webvision.med.utah.edu/book/part-i-foundations/simple-anatomy-of-the-retina/

[6] R. Poplin, A. Varadarajan, K. Blumer, Y. Liu, M. McConnell, G. Corrado, L. Peng, and D. Webster, "Prediction of cardiovascular risk factors from retinal fundus photographs via deep learning," *Nature biomedical enginnering*, 2018.

[7] Y. D. Kim, K. J. Noh, S. J. Byun, S. Lee, T. Kim, L. Sunwoo, K. J. Lee, S.-H. Kang, K. H. Park, and S. J. Park, "Effects of hypertension, diabetes and smoking on age and sex prediction from retinal fundus images," *Scientific reports*, 2020.

[8] N. Gerrits, E. Bart, T. Van Craenendonck, D. Triantafyllidou, I. Petropoulos, R. Malik, and P. De Boever, "Age and sex affect deep learning prediction of cardiometabolic risk factors from retinal images," *Scientific reports*, 2020.

[9] G. Syed, E. Trucco, C. C. Lang, Y. Huag, I. Mordi, and A. Doney, "Biological vascular age determined from retinal photographs used for diabetes retinal screening as a predictor for all-cause death and cardiovascular events."

[10] M. Tan and Q. V. Le, "Efficientnet: Rethinking model scaling for convolutional neural network," *International Conference on Machine Learning*, 2019.

[11] A. Fetit, A. Doney, S. Hogg, R. Wang, T. MacGillivray, J. Wardlaw, F. Doubal, G. McKay, S. McKenna, and E. Trucco, "A multimodal approach to cardiovascular risk stratification in patients with type 2 diabetes incorporating retinal, genomic and clinical features," *Scientific reports*, 2019.

[12] H. L. Hebert, B. Stepherd, K. Milburn, A. Veluchamy, W. Meng, F. Carr, L. D. Donnelly, R. Tavendale, G. Leese, H. M. Colhoun, E. Dow, A. D. Morris, A. S. Doney, C. C. Lang, E. R. Pearson, B. H. Smith, and C. N. Palmer, "Cohort profile: Genetics of diabetes audit and research in tayside scotland (godarts)," *International Journal of Epidemiology*, vol. 47, no. 2, 2018.

[13] M. R. Council, *MRC ethics series, Good research practice: principles and guidelines*, 2012.

[14] B. Krose and P. Van Der Smagt, *An introduction to neural networks*, 1996.

[15] J. Brownlee. [Online]. Available: https://machinelearningmastery.com

[16] T. Dozat, "Incorporating nesterov momentum into adama," 2016.

[17] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, 2015.

[18] S. Ruder, "An overview of gradient descent optimization algorithms," 2017.

[19] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning (Adaptive Computation and Machine learning series.* The MIT Press, 2016.

[20] V. Agarwal. (2020) Complete architectural details of all efficientnet models. [Online]. Available: https://towardsdatascience.com/complete-architectural-details-of-all-efficientnet-models-5fd5b736142

[21] L. Boyle, "Data analysis using godarts data sets and machine learning," Master's thesis, University of Dundee, 2020.

[22] World health organization. [Online]. Available: www.who.int