UNIVERSITÀ
DEGLI STUDI
DI PADOVA

DIPARTIMENTO
DI INGEGNERIA
DELL'INFORMAZIONE

DIPARTIMENTO DI
INGEGNERIA DELL'INFORMAZIONE

CORSO DI LAUREA MAGISTRALE IN
BIOINGEGNERIA

"ESTIMATING GLYCATED HEMOGLOBIN EXPOSURE FROM RETINAL SCANS: A DEEP LEARNING STUDY IN A DIABETIC POPULATION"

Relatore: Prof. Fabio Scarpa

Laureando: Andrea Quinto

Correlatore: Prof. Emanuele Trucco

ANNO ACCADEMICO 2021 – 2022

Data di laurea 13 Dicembre 2022

University
of Dundee

Abstract

Worldwide, 537 million people aged between 20 and 79 were living with diabetes in 2021 and this number is predicted to rise to 643 million by 2030 and 783 million by 2045. Diabetes is responsible for 6.7 million deaths in 2021 - 1 every 5 seconds. In order to contain the growth of these numbers and improve the life quality of patients, it is important to screen categories at risk, diagnosing diabetes at its early stage. To do so, one of the most popular tests is the glycated hemoglobin (HbA1C) test, that measures the percentage of blood sugar attached to hemoglobin and shows the average blood sugar level for the past 2 to 3 months.

This test requires a blood sample acquisition, which is an invasive procedure. For this reason, predicting HbA1c from a fundus camera image can be useful, and it also provides other retinal biomarkers. In this work, it has been tested whether this can be done with enough accuracy using convolutional neural networks. To do so, a dataset made of 100153 images coming from 16799 subjects has been split in training, validation and test set to train and test an EfficientNet B2 convolutional neural network. For each subject, many different images were taken over time for both eyes together with an HbA1c measurement.

Two different experiments have been performed: the first one trying to predict the latest HbA1c measurement using the first image collected for each subject and the second one trying to predict the HbA1c value corresponding to each image. After having found that the first experiment was unfeasible, due to the misconception of predicting future information from a tissue carrying information about the past history, the focus shifted to the second experiment. In this case, even if the results were not excellent, they have shown more promising perspectives and coherence of the predictions over time. An analysis of

the effect of sex and age on the cumulative HbA1c value as been performed, confirming that these variables do not affect it. Afterwards, an analysis of the time trend of the predictions has been performed, fitting them with a linear model and extracting its parameters, for each subject. Depending on the position of the predicted fit with the respect to the actual fit, the subjects have been given three categories. A χ^2 test has been performed to inspect whether there was an association or not between these categories and the death outcome. The same has been done for a risk-class category, build on whether the predicted slope was greater or smaller than the actual one.

Since this work is strongly experimental and one of the first of his kind, it has the aim to pave the path in this specific field and has big room for improvement. Most of the limitations of this work come from the cumulative nature of the data, but some suggestions on how to improve already exist and are presented in the dedicated chapter.

Dedication

Dedicated to my grandpa.

Acknowledgements

This work was made possible thanks to the collaboration between the University of Dundee and the University of Padova. In particular, thanks to the Vampire (Vascular Assessment and Measurement Platform for Images of the REtina) research group, specialized in Retinal Image Analysis (RIA), and the CVIP (Computer Vision and Image Processing) research group. The teams and their members played an important role in providing suggestions on how to overcome many issues. The most relevant contribution came from the supervisor Emanuele Trucco, who followed the work from start to finish, Dr Alex Doney, who checked medical and biological mistakes and provided interesting cues, Mohammad Ghouse Syed, who provided great support with the code and Huan Wang, who provided us with the dataset to work with. Moreover, this thesis was developed together with another research work carried out by my colleague Sara Poletto, from the University of Padua. These two dissertations should be considered a single research work, since they share parts and are complementary. Finally, thanks to the HIC support team, always providing fast replies and solutions to technical issues with the Safe Haven environment.

Contents

1. Introduction

1.1 Work motivation

The motivation for this work is strictly related to the high rate at which diabetes is spreading worldwide, as it will be explained in the dedicated following section 1.2. In order to slow down this spreading and hopefully stop it, it is crucial to detect patients from this condition at the early stages of the disease. To do so, it is of critical importance to have a reliable and accurate test. One of the most popular ones is the glycated haemoglobin (HbA1c) measurement, able to reveal the percentage of blood sugar attached to haemoglobin and showing the average blood sugar level of the past 2 to 3 months.

Since it requires the acquisition and the analysis of a blood sample from the patient, this procedure is slightly invasive. The importance of HBA1c in detecting diabetes patients and the invasiveness of this test explain the motivation of this work, in which it has been attempted to use a convolutional neural network to predict the cumulative value of the blood concentration of this molecule from fundus camera images, obtained from a simple and non-invasive procedure. The aim is to stratify patients depending on their exposure, which is measured as the integral under the curve of HbA1c measures taken over time.

This chapter aims to provide an introduction to the context in which this work has been conducted, starting from the description of diabetes and glycated haemoglobin, going through the anatomical structure of the eye and the retina, with the explanation of the main imaging techniques and their applications, arriving at the description of the dataset used. A brief summary of the work conducted and an overview of how the document is structured will be provided as well before moving on to the next chapter.

1.2 Diabetes

Diabetes is a chronic, metabolic disease caused by a modification of the insulin-glucose control system and characterized by elevated levels of blood glucose. Over time, it leads to serious damage to heart, blood vessels, eyes, kidneys and nerves. There are two different types of diabetes:

- Type 1: is an autoimmune pathology in which the β-cells of the pancreas, 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 does 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.

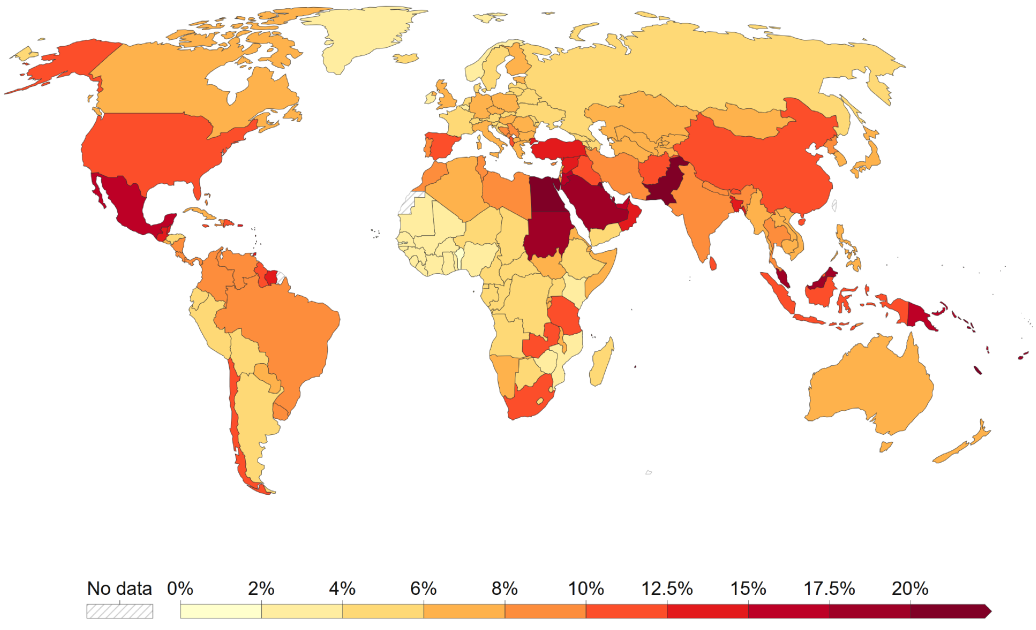

Figure 1.1: *Share of people between 20 and 79 who had diabetes in 2021 (International Diabetes Federation)*

The most common form is type 2 and during the past 3 decades its prevalence has risen dramatically in countries of all income levels. For people living with diabetes, access to affordable treatment, including insulin, is critical to their survival. There is a globally agreed target to halt the rise in diabetes and obesity by 2025 [8]. In 2021, about 537 million people worldwide had diabetes and 6.7 million died because of it. Both the number of cases and the prevalence of diabetes have been steadily increasing over the past few decades [14]. Concerning the UK, one in ten people over 40 has been diagnosed with diabetes, with a total of 3.8 million people, 90% of those were type 2. 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.

Glucose metabolism is a fundamental process in a living organism, since from the physiological point of view it provides the energy necessary for the different vital functions of cells and organs, while from the pathological point of view its malfunctioning may cause glucose intolerance and diabetes. The endocrine system is the most important in glucose regulation, since it produces two hormones: insulin (in pancreatic β-cells) and glucagon (in pancreatic α-cells). These two hormones are continuously secreted and act with opposing actions to maintain the glucose concentration into 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 both store glucose when its concentration is high and produce and release it in the bloodstream when the organism is in deficiency. For the subject's health, it is essential that the blood glucose concentration is maintained inside 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) [1]. 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 [3]. 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-resistance, which consists in cells' incapability of using insulin: they have fewer insulin receptors, making glucose not enter them and accumulate in the bloodstream. As the first reaction, the organism produces more insulin to maintain a low glucose concentration (hyperinsulinism). However, in the next stage, the increased insulin production does not maintain glycaemia to normal values, and hyperglycaemia happens.

The main stimulus for insulin release is glucose itself: when plasma glucose concentration exceeds 100mg/dL, it is absorbed from the small intestine and reaches β-cells, that can detect it with GLUT-2 transporters and release insulin in response. Glucose and insulin act through a closed-loop system: if plasma glucose concentration rises, β-cells secrete a larger amount of insulin, promoting its consumption in order to restore the basal level of glycaemia.

1.3 Glycated haemoglobin (HbA1c)

HbA1c refers to glycated haemoglobin, which is a molecule of haemoglobin bounded with glucose in the blood, thus becoming glycated. The measure of its concentration in the blood gives an overall picture of the average blood sugar levels in the last couple of months. For diabetic patients, this measure is very important since a high value of HbA1c is related to a greater risk of developing diabetes-related complications, like eye and kidney damage, dementia and cardiovascular problems [2].

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.

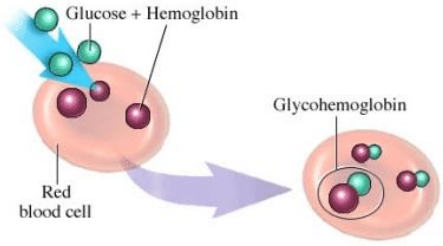

Figure 1.2: *Glycated haemoglobin*

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 long-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 any 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, haematologic and illness-related factors. Some of them are haemoglobinopathies, certain anaemia and disorders associated with accelerated red cell turnover, such as malaria [15]. Healthy people have HbA1c level below or equal to 5.6%, a prediabetic condition is identified within the range from 5.7% to 6.4%, and diabetic people have levels higher than 6.5%. The target for diabetic people is 6.5%, which corresponds to 48 mmol/mol [mmol of HbA1c per mol of haemoglobin].

1.4 Retina and retinal imaging

1.4.1 The anatomy of the eye

The protective bony socket where the eye is located is called orbit. Here, six muscles are attached to the eye, allowing all its movements. Other extraocular muscles are attached to the sclera, a strong layer of tissue that covers almost the entire surface of the eyeball. The membrane covering the eye surface and the inner surface of the eyelids is called conjunctiva. The surface of the eye is lubricated by three layers of tears that together compose the tear film: the mucous layer, made by the conjunctiva, the watery layer, secreted by the lacrimal gland, and the oily layer, secreted by the meibomian gland.

The frontal part of the eye, through which the light is focused, is called the cornea. Behind it, a space called anterior chamber is filled with a fluid called aqueous humor. A system of production and draining of this fluid is always active in order to keep a constant eye

pressure. Behind the anterior chamber there is the eye iris, with a dark hole in its middle, called pupil. The muscles contained in the iris allow changing the pupil size in order to control the amount of light reaching the back of the eye. Behind it, there is the lens, responsible for focusing the light passing through the pupil toward the back of the eye by changing shape depending on the distance of the object the eye is focusing on. The lens is surrounded by the lens capsule, and some fibers called zonules are attached to it, suspending it from the eye wall. By helping to focus light as it enters the eye, the cornea and the lens both play important roles in giving us clear vision. In fact, 70% of the eye's focusing power comes from the cornea and 30% from the lens.

The space between the lens and the back of the eye is called vitreous cavity and is filled with a fluid called vitreous humor. After passing through the pupil, being focused by the lens and passing through the vitreous humor, the light reaches the retina, the light sensitive tissue covering the inside of the back of the eye. A tiny but very specialized area of the retina, called macula, is responsible for giving us detailed central vision, while the peripheral retina provides us with side vision. The retina has photoreceptors, special cells converting light into an electric signal and transmitting it through the optic nerve to the visual cortex, the part of the brain responsible for sight. There are two types of photoreceptors: rods, perceiving black and white and enabling night vision, and cones, perceiving color and providing detailed central vision.

1.4.2 Inside the retina

With a deeper insight into the retina, we can see the optic nerve, a circular/oval white area containing the ganglion cells' axons running to the brain and, incoming blood vessels that open into the retina to vascularize its layers and neurons. Approximately two and a half disc diameters to the left of the disc there is an oval-shaped reddish spot free from blood vessels, the fovea, which is at the center of an area known as macula. A circular area of approximately 6 mm around the fovea is called central retina while the area beyond this is called peripheral retina, stretching to the ora serrata, 21 mm from the center of the retina. The total retina is a circular disc with a diameter varying between 30 and 40 millimeters. This concentric structure is shown in figure 1.3.

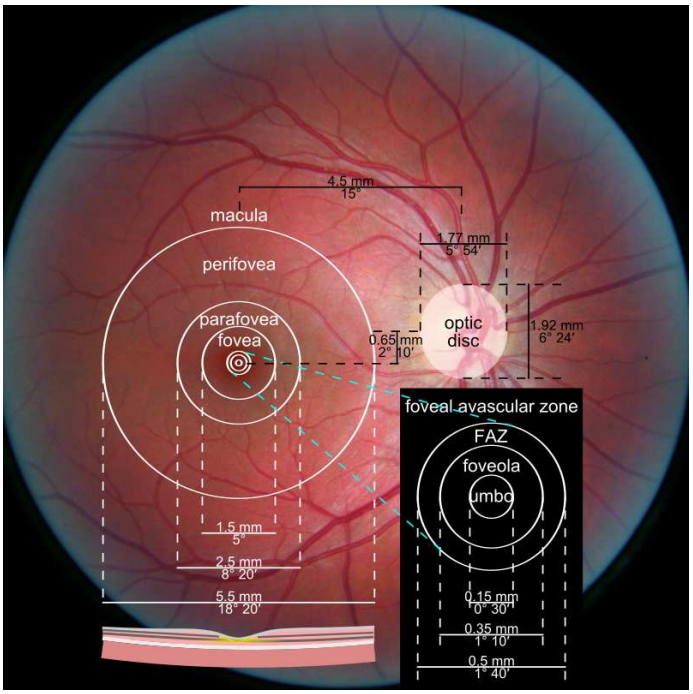

Figure 1.3: *Fundus camera image of the retinal, with details about its components*

The center of the fovea is known as the foveal pit, a highly specialized region of the retina different from the central and peripheral areas. It is a spot where rods are absent while cones are present at their maximum concentration density, achieved thanks to the hexagonal mosaic, their most efficient packing system. Up the foveal pit, along the foveal slope, there is the parafovea, and around it the perifovea. The whole foveal area - including foveal pit, foveal slope, parafovea and perifovea - is the macula. In the macula lutea, yellow screening pigments give rise to the yellow pigmentation. The macula lutea is thought to act as a short wavelength filter, additional to that provided by the lens. As the fovea is the most essential part of the retina for human vision, protective mechanisms for avoiding bright light and especially ultraviolet irradiation damage are essential. Indeed, the consequence of foveal cones being destroyed is blindness.

The retina is approximately 0.5 mm thick and lines the back of the eye. Its radial section reveals that the ganglion cells (the output neurons) lie innermost, closest to the lens and front of the eye, while rods and cones lie outermost, against the pigment epithelium and choroid. Therefore, light has to travel through the thickness of the retina before striking and activating the photoreceptors. After they absorb photons through the visual pigment,

light is translated into a biochemical message first and into an electrical one afters, able to stimulate all the succeeding neurons of the retina. The retinal message is finally transmitted to the brain from the spiking discharge pattern of the ganglion cells.

All vertebrate retinas are composed of three layers of nerve cell bodies and two layers of synapses (Fig. 5). The three nerve cell bodies are the outer nuclear layer, containing cell bodies of rods and cones, the inner nuclear layer, containing cell bodies of bipolar, horizontal and amacrine cells and the ganglion cell layer, containing cell bodies of ganglion cells and displaced amacrine cells. Dividing these nerve cell layers are two neuropils where synaptic contacts occur: the outer plexiform layer (OPL) and the inner plexiform layer (IPL). Concerning blood supply, there are two sources in the mammalian retina: the central retinal artery, receiving 20-30% of the flow and nourishing the inner retinal layers, and the choroidal blood vessels, receiving 65-85% of the flow and nourishing the outer layers.

1.4.3 Retinal imaging techniques

During the last century, retinal imaging underwent a rapid and consistent development. It is now a mainstay of the clinical care and management of patients with retinal as well as systemic diseases. The main retina imaging techniques and their applications are the following:

- Fluorescein Angiography (FA): is a diagnostic procedure that uses a special camera to record the blood flow in the retina without involving any direct contact with the eyes. Fluorescein dye is injected into a vein in the arm/hand, and it will fluoresce in the blood vessels and be recorded as gray or white light in the image. Photographs are taken as dye passes through the blood vessels of the eye, allowing abnormal blood vessels (displaying hypofluorescence or hyperfluorescence) or damage to the lining beneath the retina to be revealed. Fluorescein angiograms are often recommended to follow the progression of a disease and to monitor treatment results. It is particularly useful in the management of diabetic retinopathy and macular degeneration.

- Autofluorescence Imaging (FAF): is the concept of using naturally occurring fluorescence from the retina to provide an indicator of the retinal pigment epithelium

health. Illuminating the retina with blue light causes certain cellular components to "glow" without injecting any dye. The fluorescence returning from the retina can be used to create a black-and-white image which can be interpreted by recognizing characteristic patterns. Potential applications of FAF imaging have been explored in a variety of retinal diseases including: age-related macular degeneration, retinitis pigmentosa, central serous chorioretinopathy and macular dystrophies.

- Optical Coherence Tomography: is a noninvasive imaging technology used to obtain high resolution cross-sectional images of the retina. The layers within the retina can be differentiated and retinal thickness can be measured to aid in the early detection and diagnosis of retinal diseases and conditions. It uses rays of light to measure retinal thickness and since no radiation or X-rays are used in this test, it does not hurt, and it is not uncomfortable. Some applications are monitoring the progress of your disease, verifying or discounting suspected swelling of the retina, or checking OCT results against other results. This is done to determine the effectiveness of the current medication regime.

- Color Fundus Photography: is the technique that has been used to collect retina images in the GoDARTS dataset used for this work. It uses a fundus camera to record color images of the condition of the interior surface of the eye. This is done in order to document the presence of disorders and monitor their evolution over time. A fundus camera or retinal camera is a specialized low power microscope with an attached camera designed to photograph the interior surface of the eye, including the retina, retinal vasculature, optic disc, macula, and posterior pole (i.e. the fundus).

The retina is imaged to document conditions such as diabetic retinopathy, age related macular degeneration, macular edema and retinal detachment. Fundus photography is also used to help interpret fluorescein angiography, as certain retinal landmarks visible in fundus photography are not visible on a fluorescein angiogram. Your eyes will be dilated before the procedure. Widening (dilating) a patient's pupil increases the angle of observation. This allows the technicians to image a much wider area and have a clearer view of the back of the eye [13].

1.5 Dataset description

The DARTS - Diabetes Audit and Research in Tayside Scotland - study started in 1996 as a collaboration between the University of Dundee, the three Tayside Health Care Trusts: the Ninewells Hospital and Medical School, the Perth Royal Infirmary, the Stracathro Hospital and a group of Tayside general practitioners to identify all diabetic patients within the Tayside region and to improve health care [10]. The collected data includes hospital diabetes clinics, diabetes prescription database and all diabetes-related records. It is continually updated and forms an extremely valuable longitudinal dataset of clinical data.

In 1998 genetic data started to be collected through a blood sample for DNA extraction while phenotypic data was collected 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, progression and response to treatment.

GoDARTS was created in three phases. The first was the pilot phase, GoDARTS1, to test the recruitment processes and the ability to anonymously link patient clinical data from electronic records to the study. In this phase, when the patient was recruited, only blood samples were taken and no baseline data were recorded. In the second phase, GoDARTS2, two groups of patients were enrolled, a type 2 diabetes patients group and a control group. On average, there was one control per case of diabetes. Baseline clinical and lifestyle measurements were recorded for all patients. This included the smoking history, level of physical activity and menopause history for women, as well as height, weight, blood pressure and heart rate. 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. Only 1451 patients involved exclusively in the GoDARTS1 trial do not have these data.

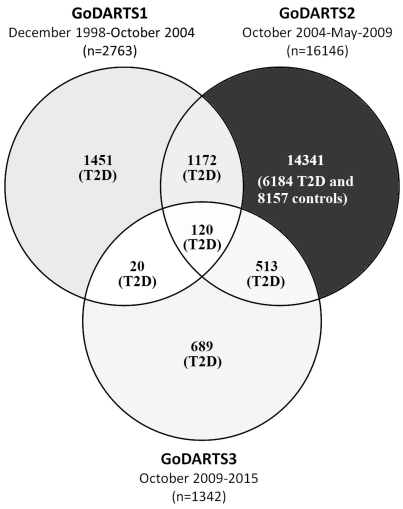

Figure 1.4: *Diagram of the overlap in patient recruitment in the three phases.*

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 opens the possibility to access longitudinal data related to routine diabetes management, such as the glycated haemoglobin data. 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. The index is then converted to a study pro-CHI so that patient identities are protected, but multiple datasets can still be linked.

	Case	Control	Overall
Gender (% male)	56.38	50.08	53.33
Age	67	60	64
HbA1c (%)	7.1	5.5	6.0

Table 1.1: *Median value*

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 that, being self-completed, may have some bias. In the following table, we can see the median value of the variables gender, age and glycated haemoglobin of the GoDARTS patients.

1.6 Summary of the work

The initial phase of the work consisted in feature selection from a GoDARTS dataset of retinal and clinical features, using a LASSO technique applied in a bootstrap resampling procedure. The aim is to familiarize with the GoDARTS data and machine learning techniques on big medical data and to confirm the results obtained in previous similar studies.

After this introductory work on feature selection, the two core experiments of the work have been conducted with the aim to discover whether it is possible or not to use temporal information present in retina images to predict the trend of cumulative glycated haemoglobin (HbA1c) in the future. The first experiment consists in predicting the last HbA1c measurement using the baseline image, while the second one in predicting the HbA1c measurement relative to each image.

Afterwards, an analysis of the effects of age and sex on cumulative glycated haemoglobin has been conducted, showing that in this data set of cumulative measures, they seem to have no particular effect, since the distribution is the same in males and females, and it does not have any trend with age.

The following part of the work dealt with the time trend of the actual and predicted data, fitting them with a linear model and extracting the parameters (slope and intercept) for each subject. On the base of the relative position of the two fits they have been classified as "Above", "Below" or "Interception" and on the base of the magnitude of the predicted slope with respect to the actual slope they have been given a "risk" score of 0 or 1. Statistical χ^2 test has been performed in order to establish whether a relationship between these 2 variables and the death outcome existed or not.

1.7 Structure of the document

In this introductory chapter, the most significant aspects of the context of the following work have been explained. The description started from the general understanding of diabetes and the relevance of HbA1c as a tool to detect it at its early stages. Then it moved to the general anatomy of the eye, digging into detail about what concerns the retina and describing the most popular imaging techniques of this tissue. In the end, an insight into the dataset used was also provided. The rest of the document is structured as follows: in the related work chapter, a series of analyses conducted for this work and others in the same field will be presented; in the methodology theory, a description of the mathematical and computational tools applied in this work will be explained in detail. Afterwards, in the fourth chapter, a more technical insight into the work will be provided, describing the work environment and some details about the algorithms used. The following chapter describes the experiments as well as how the data were manipulated to ask relevant questions and how the collected results were analyzed. To conclude, limitations will be discussed together with the possible ways to overcome them in the final chapter. In this section, a prospect about the future work on this topic will also be provided.

2. Related work

2.1 Introduction

The previous chapter provided a thorough introduction to the biological context of the work. In particular, it began diving into the description of diabetes, explaining its causes and effects, as well as the difference between type one and type two and the glucose-insulin system. Afterwards, it moved to glycated hemoglobin (HbA1c), explaining why this molecule is important in the context of diabetes and what kind of information it can provide. Then the anatomy of the eye, of the retina and the main imaging techniques to inspect them have been presented, with a quick overview of fluorescein angiography, autofluorescence, optical coherence tomography and color fundus photography. A description of the GoDARTS dataset followed, and a summary of the work and a description of the document's structure were provided to conclude.

As a follow-up to the introduction, this second chapter will provide a more technical description of the work's context, presenting some machine and deep learning applications to medical data and retinal imaging. In particular, the first section is about the work conducted on LASSO feature selection from a GoDARTS dataset of retinal and routine features. This was done in order to familiarize with GoDARTS data and features and to confirm the results previously obtained in other studies. In the following section, a Dundee CVIP's study using deep learning to predict age from fundus camera images will be described, since it played a crucial role for our work. Then, a brief summary of a series of studies about deep learning applied to retinal images, will be presented, since they have been useful to have a better understanding of the state of the art in this field.

2.2 LASSO features selection

Before dealing with deep learning and retinal images, the beginning of the work focused on establishing the usefulness of a machine learning algorithm for patient stratification and selecting the most relevant features to predict the presence or absence of MACE. Major Adverse Cardiovascular Events (MACE) are defined as nonfatal stroke, nonfatal myocardial infarction and cardiovascular death. The task has been carried out using the LASSO regression in a bootstrap resampling method and evaluating the performance using the area under the ROC curve (AUC). A GoDARTS dataset, made of 4711 subjects and 184 retinal and clinical features has been used. At the end, the results were compared with those obtained by L. Boyle in his dissertation [4].

LASSO stands for Least Absolute Shrinkage and Selection Operator and is a regularization method that encourages simple, sparse models by introducing some bias in the coefficient estimates but reducing their variance. This is done by adding the L1 norm of the coefficients as a penalty term to the residual sum of squares, weighted by the shrinkage parameter λ:

$$\sum_{i=1}^{n}(y_i - \sum_{j} x_{ij}\beta_j)^2 + \lambda \sum_{j=1}^{p} |\beta_j|, \tag{2.1}$$

where y_i is the outcome of the i-th subject, x_{ij} is the j-th feature of the i-th subject, β_j is the regression coefficient of the j-th feature. Data has been preprocessed before starting the bootstrap loop. Of all the features, some were removed since they were not informative (like image size, resolution or others having the same value for each subject). Others were removed because they were highly correlated (like the date of birth and age) so only one of each couple was kept. The feature cvd_fail, stating whether the subject had a cardiovascular disease or not, has been used as outcome. Then the data have been imputed using the nearest neighbor method and z-scored, subtracting the mean and dividing by the standard deviation.

Within each bootstrap loop, an internal training set of the same size as the original training set (made of 3533 subjects) has been sampled with repetition from it, and the out-of-bag subjects have been used as the test set. This internal dataset has been sampled with stratification, ensuring that the percentage of subjects with each outcome was the same as the original dataset. A number of features were selected at each iteration, corresponding to

a value of λ achieving the minimal value of the binomial deviance, and the area under the ROC curve (AUC) has been calculated. The bootstrap procedure allowed to extract confidence intervals of the AUC metric, showing how robust the model is. Over the entire procedure, it has been kept count of the number of times each variable was selected. The results are reported in table 2.2 and the plot of the binomial deviance trend for different values of λ at a certain iteration is reported in figure 2.3, together with the number of selected features. The box plot of the AUC values is reported in figure 2.1, and the first quartile, median and third quartile are reported in table 2.1.

Stat	AUC value (%)
0^{th} quartile	85.67%
1^{st} quartile	87.58%
Median	88.13%
3^{rd} quartile	88.75%
4^{th} quartile	91.08%

Table 2.1: *AUC distribution values*

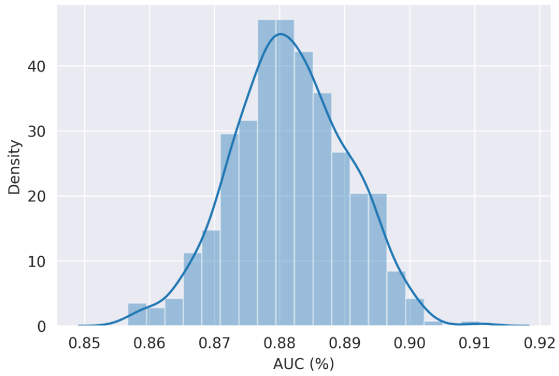

Figure 2.1: *Distribution of the AUC values throughout the 500 iterations.*

Rank	Feature	Counts	Type	Rank	Feature	Counts	Type
1	leverhulme	500	Ret	11	tortq1g1amin	420	Ret
2	pre	500	Cli	12	gradq1ahermite	414	Ret
3	dement_time_age	500	Cli	13	odradiuspx	410	Ret
4	e_age	500	Cli	14	gh	407	Cli
5	odfovea	482	Ret	15	bcvhermite	393	Ret
6	ldrvspline	456	Ret	16	torta	388	Ret
7	trig	448	Cli	17	afvspline	373	Ret
8	bcvspline	446	Ret	18	ltortv	369	Ret
9	tortimageg1amed	426	Ret	19	chol	366	Cli
10	gradq3vspline	423	Ret	20	c_sbp	345	Cli

Table 2.2: *Bootstrap feature selection results*

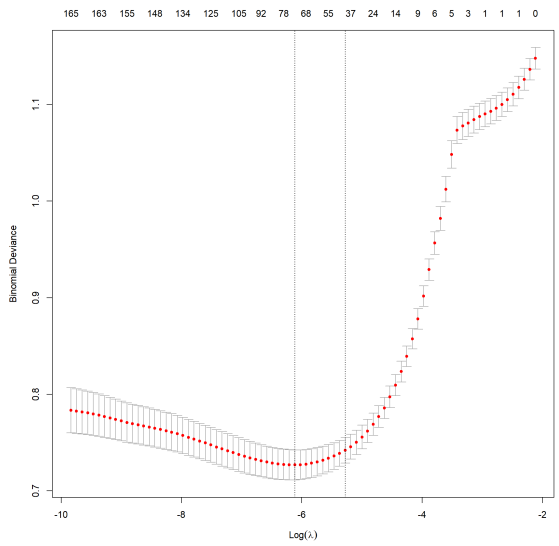

Figure 2.2: *Binomial deviance trend for different values of* λ.

The results are in line with the ones previously obtained in other studies, with the only exception that a Retinal feature has been selected 500 times over 500 iterations. Indeed, Retinal features usually contribute more modestly to the prediction, especially when combined with many other types of features.

2.3 Age prediction

Accelerate aging 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 recognized shreds of evidence of age-related tissue health, such as manifestations of white matter disease and other age-related structural changes. The retina has the same embryonic origin as the brain, the neural plate, and is a highly vascularized 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. We can think that an individual with a predicted age greater than their chronological age has a more significant risk of all-cause death.

In a previous study [16], DL was used to predict biological vascular age from retinal images 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 hospitalization. Images were pre-processed to reduce the effect of image variations, such as brightness, colour and focus, and they were resized to the standard size of 260x260 pixels, which is the one recommended improving accuracy in the used neural network [17].

Ghouse et al. used the EfficientNet-B2 network since it achieves excellent performance in imaging tasks. They modified the fully connected layer and replaced it with a global average pooling layer followed by a single output node with linear activation. Moreover, 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, they can be considered an important achievement in understanding how retinal images can be used in deep learning to predict a biological outcome.

2.4 Other work

The usage of deep learning algorithms to extract clinical information from medical images and signals began to spread widely along with the development of this new technology. For this aim, the usage of retinal images, fundus camera in particular, has become extremely popular. In the following, a brief summary of some studies that turned out useful for this work will be presented.

In a study by Gerrits et al. (2020) [9], the MobileNet-V2 architecture has been used to investigate whether fundus camera retinal images can predict cardiometabolic risk factors. The study stated that age and sex can be predicted with high performance, while systolic blood pressure, diastolic blood pressure, Haemoglobin A1c, relative fat mass and testosterone can be predicted with lower but still acceptable performance. In order to discover what areas of the image the network was paying attention to, they used the Grad-CAM

technique. It revealed that the network was focusing on the vasculature and on the optic disk when predicting age and sex, as well as when predicting the other features mentioned earlier.

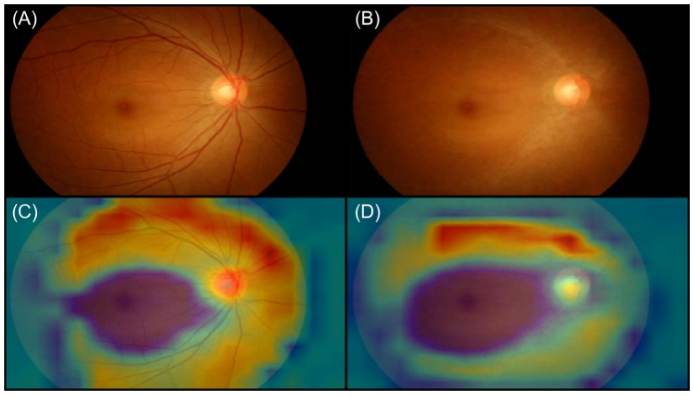

Figure 2.3: *GradCAM applied in the study by Kim et al. (2020) [11].*

Similarly, in a study by Y. D. Kim et al. (2020) [11], ResNet-152 has been used as a backbone network to predict age and sex and understand how hypertension, diabetes and smoking affect these predictions. They trained the network with fundus images coming from healthy patients and tested it on 4 different datasets: healthy set, hypertension set, diabetes set and smoking set. Briefly, they found that the healthy set reported the most accurate age prediction and that prediction accuracy decreases significantly with the increasing of the patient age. On the contrary, no relevant changes in accuracy were recorded when predicting sex, showing that vascular conditions only affect age prediction.

Another study by J. H. Cole et al. (2018) [6] investigated biomarker prediction using MRI brain images and machine learning. In particular, T1-weighted MRI scans were used to train a Gaussian process regressor to predict chronological age. The authors were able to show that this marker of brain aging is associated with a greater risk of death and poorer physical and cognitive fitness.

2.5 Conclusions

The field of biomarkers prediction using neural networks has been widely explored throughout this chapter. These studies revealed to be useful to understand the potential of neural network in this field, proving that they are a tool able to estimate many biomarkers reliably. The results of the work on feature selection with LASSO and bootstrap resampling method were presented as well, showing that clinical features still prevail over retinal ones in predicting cardiovascular events. After having provided some insight on this topic, the next chapter will move on to give a more technical and mathematical explanation of artificial intelligence. In particular, it will provide details about neural networks in general and on the architecture used in this work, EfficientNet.

3. Methodology

3.1 Introduction

After the summary of the main work in this field reported in the last chapter, this third chapter will introduce from a computational point of view the deep learning methods used. The first section will provide a general introduction about machine learning, explaining how it works and introducing its main frameworks. Then a quick overview of neural networks will cover from the basic neuron to the most popular architectures. To conclude the chapter, the final section will describe how the family of networks used for this work, EfficientNet, was conceived and its structure.

3.2 Machine learning fundamentals

Machine learning is a branch of artificial intelligence that consists in leveraging data to improve the performance of an algorithm carrying out a certain task. From the point of view of learning from data, there are four types of machine learning algorithms: supervised, semi-supervised, unsupervised and reinforcement.

In supervised learning, labelled data coming from a training set is used to learn the parameters of a model that is able to map certain input variables, called features, into some output/target variable (outcome). To assess the performance of the model during the training process, a loss function is used to measure how far the predictions of the model are from the true values. In order to improve the performance of the model, its parameters are updated in such a way that minimizes the loss function calculated over the training set. In general, based on the feature vector $\mathbf{x} \in \mathbb{R}^n$, the classifier has to predict the variable or

the correct class y, which is estimated by a function $\hat{y} = f(\mathbf{x}, \boldsymbol{\theta})$ that directly results in the classification or regression output $\hat{y}$. The classifier's parameter vector $\boldsymbol{\theta}$ is determined during the training phase and later evaluated on an independent test data set.

Unsupervised learning, on the other hand, makes use of unlabeled data to discover hidden patterns and data groupings without the need of human intervention. This makes it a useful technique to discover differences and similarities in information. The most popular algorithm is clustering, coming in many different variants (K-means, hierarchical, fuzzy and many others), but there are many more, like dimensionality reduction (PCA, SVD) and autoencoders.

Semi-supervised learning is instead an approach that performs the training process with a small amount of labeled data and a large amount of unlabeled data. It is an instance of weak-supervision used in cases where data labeling can be expensive. The use of a small amount of labelled data can improve the prediction accuracy considerably.

To conclude, reinforcement learning refers to a set of algorithms involving a learning agent that interacts with the environment to achieve a goal, typically the maximization of a reward provided by the environment. The learning agent must be able to detect the state of the environment and to take actions that affect the state of the environment. While in the other learning frameworks the model learns from datasets, in reinforcement learning the agent learns from its experience. No training data are provided to the system, since it generates the examples with its own activity. RL algorithms are based on the trial and error approach: to improve their ability to get rewards, agents must continuously test the effect of their actions to learn the action that maximizes the reward for each state of the environment (policy).

3.3 Deep learning

Deep learning is a branch of machine learning that tackles the problem of teaching a machine a typically human learning framework, learning by examples. It is based on artificial neural networks with representation learning, a set of techniques that allows a system to automatically discover the representations needed for feature detection or classification. Neural networks gained an important role in many computer science fields, ranging from computer vision to reinforcement learning. The fundamental unit of a neural network is a

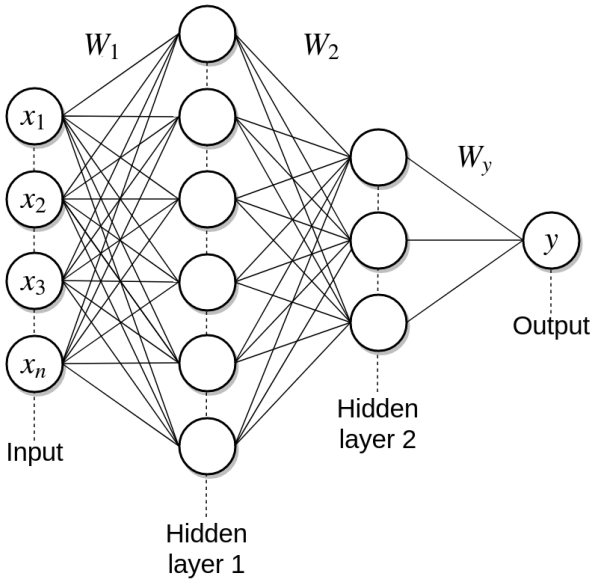

Figure 3.1: *Fully connected neural network with 2 hidden layers.*

neuron, that takes a bias w_0 and some weights $\boldsymbol{w} = (w_1, \ldots, w_n)$ to linearly combine the input $\boldsymbol{x}$ it receives as $\boldsymbol{w}^\intercal \boldsymbol{x} + w_0$, squeeze it into a non-linear activation function $h(\cdot)$ and output its own activation level. Different kinds of activation functions exist, and some of the most popular ones are the sigmoid and the hyperbolic tangent. A fully connected neural network is made up of many layers including many neurons each, in such a way that the inputs of the neuron in a layer are the outputs of the neurons in the previous layer and its output, together with the output of all the neurons in the same layer, are the inputs of all the neurons in the following layer. This architecture can be better understood looking at figure 3.3.

The parameters of the network comprise the connections between the neurons and the biases, and are learned with a gradient descent procedure performed on a function measuring the quality of the network parameters: the lower its value, the better the network performance. This function is known as loss function, since it's calculated as the average over the training set of the difference between the true and the predicted value or class probability. Depending on the application, this function can have different formulations. The gradient of the loss function with respect to each weight is calculated via the chain rule with a procedure called back-propagation, consisting in propagating the error made by the network on the training examples back from the output to the input.

To perform gradient descent, the parameters are initialized at a certain value. The initialization is crucial, since it can lead to problems like vanishing gradients or poor learning if not taken care of carefully. After initialization, the weights are iteratively updated with the following formula:

$$\mathbf{W}^{[t+1]} = \mathbf{W}^{[t]} - \alpha \nabla \mathcal{L}_{\mathbf{W}}(\mathbf{W}) \tag{3.1}$$

where $\mathbf{W}$ is the network's weights vector, α is the learning rate and $\nabla \mathcal{L}_{\mathbf{W}}(\mathbf{W})$ is the gradient of the loss function. The rationale behind it is that the parameters should be updated in a direction which is opposite to the gradient of the loss function, since it is the vector indicating its maximum growth direction in the parameters space. In the ideal scenario, the loss function is convex and has only a single absolute minima point. In practical applications, this is never the case, and the algorithm will hopefully reach a local minima point and converge, since there the gradient is close to zero and the weights update would not change much the result. This method has two main limitations:

- the choice of the learning rate (a too large one con lead to overshoot the local minima while a too small one would take too long for the algorithm to converge)
- the calculation of the gradient of the loss function over the entire training set can be computationally expensive, making the algorithm inefficient.

In order to solve the computation time issue, minibatch stochastic gradient descent (SGD) performs an update of the parameters for every mini-batch of m training examples, over which the gradient is calculated. The minibatch gradient at each step is calculated as

$$\mathbf{g}^{[t]} = \frac{1}{m} \sum_{j=i_1}^{i_m} \nabla \mathcal{L}_{\mathbf{W}}(\mathbf{x}_j, y_j, \mathbf{W}) \tag{3.2}$$

and the weights update formula becomes: $\mathbf{W}^{[t+1]} = \mathbf{W}^{[t]} - \alpha \mathbf{g}^{[t]}$. If this technique allows faster computation of the gradient, it may still take too long to converge since it uses just an approximation of the true gradient. Indeed, the steps moved are towards the local minima only on average, and the true "path" is highly oscillating. In order to reduce this source of noise, many methods have been developed, mainly based on the so-called momentum parameter. The most popular, among all, is called ADAM and updates the parameters with a normalized momentum:

$$\mathbf{W}^{[t+1]} = \mathbf{W}^{[t]} - \alpha \frac{\mathbf{p}^{[t]}}{\sqrt{\mathbf{s}^{[t]}}} \tag{3.3}$$

where:

- $\mathbf{p}^{[t]} = \beta_1 \mathbf{p}^{[t-1]} + (1 - \beta_1)\mathbf{g}^{[t]}$ is the momentum update rule,
- $\mathbf{s}^{[t]} = \beta_2 \mathbf{s}^{[t-1]} + (1 - \beta_2)(\mathbf{g}^{[t]})^2$ is the root-mean-square propagation (RMSprop) update rule.

To deal with the learning rate α, the most common and easy way is to set a decaying learning rate in order to speed up the descent with a large value at the early stages, when the algorithm is still far from the target, and a progressive smaller one when the local minima is being approached.

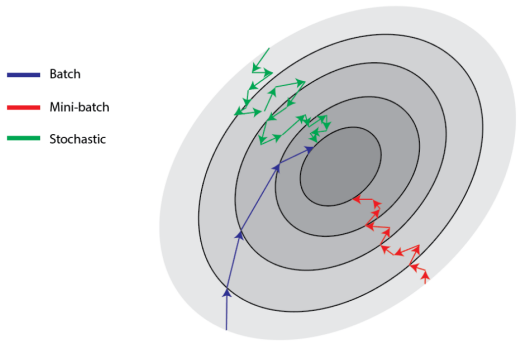

Figure 3.2: *Difference between gradient descent methods.*

3.4 Main neural networks architectures

The one described earlier is the simplest neural network architecture, connecting together basic layers of the same type. But it is also possible to connect layers of different types and give birth to more complex architectures suited for particular applications.

In the field of computer vision, where image processing is a routine task to perform, convolutional neural networks (CNN) dominate the scene thanks to their capability of extracting information from image data. The two main building blocks of this architecture are the convolutional layer and the max pooling layer. Each neuron in a convolutional layer is a filter that is convolved with the image to extract different features from it; the deeper the layer in the network, the more complex the information extracted. The convolution consists in an elementwise multiplication between the image and a filter strode along it. After the convolution with each filter in a layer, a bias is added to each feature map obtained and a non-linear function is applied to each element. The number of learnable parameters

in a convolutional layer is the product between the number of weights in each filter and the number of filters in the layer. Then the outputs are stacked one after the other and sent into a pooling layer, where a kernel is once again strode along the image, extracting typically the maximum value from the corresponding image area. This layer has no learnable parameters. A deep CNN is made of many convolutional and pooling layers stacked one after the other, connecting the last pooling one to a fully connected one.

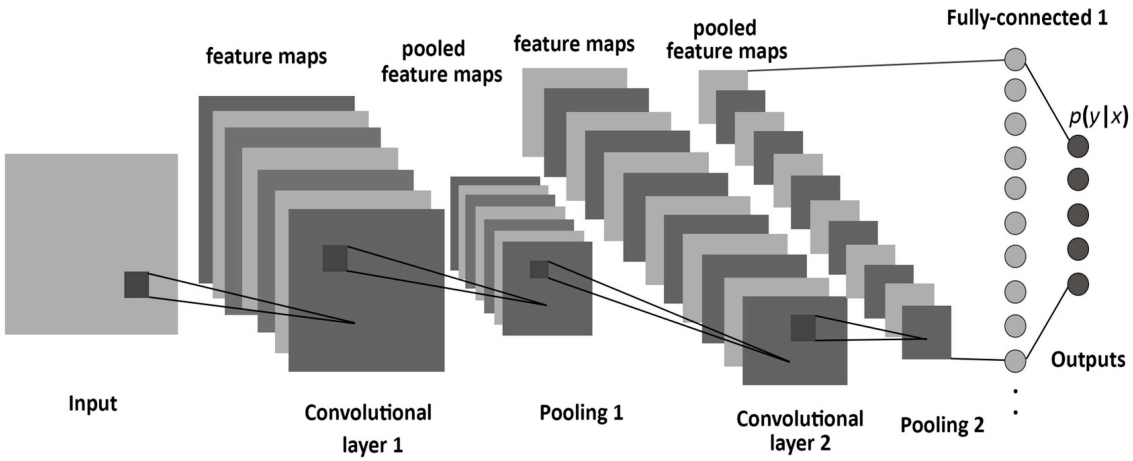

Figure 3.3: *Simple representation of a convolutional neural network.*

Recurrent neural networks are instead widespread in all the applications involving information persisting over time, like time series analysis, speech recognition and synthesis and many others. The temporal dynamic behaviour is achieved by allowing a neuron to take information not only from the former layer, but also from even prior ones [12]. There are different types of recurrent neural networks: one to one, one to many, many to one and many to many, depending on the size of the input and the output. Some of the most popular architectures are Bidirectional recurrent neural networks (BRNN), Long short-term memory (LSTM) and Gated recurrent units (GRUs).

3.5 EfficientNet

In common deep learning frameworks, convolutional neural networks are developed with a fixed amount of computational resources available and then scaled up by their dimensions to achieve better accuracy. In most cases, the scaling is performed by depth or width and more rarely by image resolution, but in any case this is done one dimension independently of the other.

EfficientNet is a family of convolutional neural networks developed from the same baseline architecture to achieve maximum accuracy using an optimal amount of resources (and so being the most efficient possible). This has been shown to be possible by scaling each dimension with a constant ratio with a set of fixed scaling coefficients. This method is called Compound Model Scaling and consists in increasing the width, depth and resolution of the network without changing its baseline architecture and considering that the optimal value of these parameters depend on each other. The network dimensions are scaled uniformly using a compound coefficient ϕ, controlling how many more resources are available:

- depth: $d = \alpha^{\phi}$,
- width: $w = \beta^{\phi}$,
- resolution: $r = \gamma^{\phi}$,

where α, β and γ are parameters determined by grid search and such that $\alpha \cdot \beta^2 \cdot \gamma^2 \approx 2$.

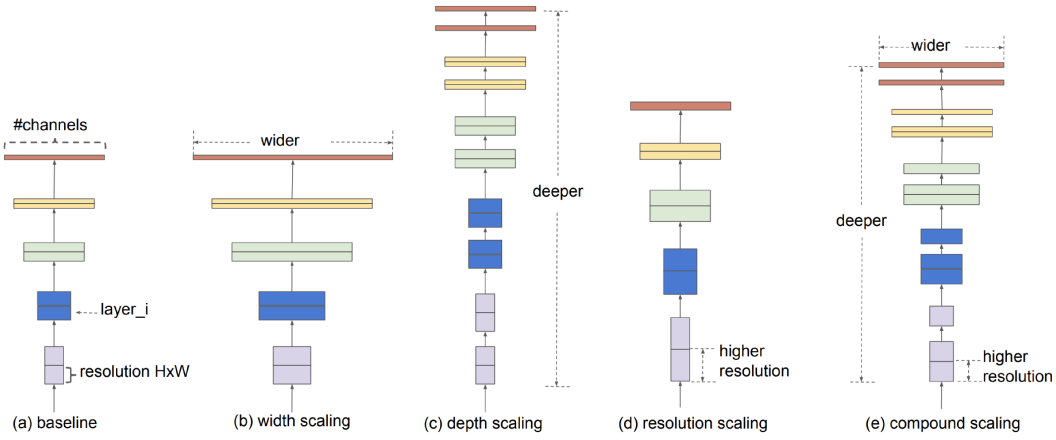

Figure 3.4: *Comparison of different scaling methods.*

The baseline network is for sure a critical choice, since this process does not modify the model architecture. To find it, a multi-objective neural architecture search was performed, optimizing both accuracy and FLOPS. It was called EfficientNet-B0 and the compound scaling method was applied to it in the following two steps to scale it up and obtain the bigger EfficientNet versions:

- step1: ϕ is fixed to one, and α, β and γ satisfying the constraints mentioned earlier are found by grid search,
- step2: α, β and γ are fixed, and the baseline net is scaled up using different values of ϕ, obtaining EfficientNet from B1 to B7.

The network used in our work is an EfficientNetB2; more details about the optimization procedure and training setup are presented in section 4.5.

In figure 3.5, the architecture of EfficientNet-B2 is represented with a block diagram. Its first module is the stem one, which is common in all the eight EfficientNet models (as well as the final layers). After the stem module, it contains 7 blocks (as well as all the other EfficientNet variants), on which all the experimenting with the architecture is performed. Each of these blocks has a varying number of sub-blocks itself, whose number is increased moving from EfficientNetB0 to EfficientNetB7. The total number of layers in EfficientNet-B0 is 237, while in EfficientNet-B7 is 813, but all these layers can be grouped in the stem module and the 5 modules shown in figures 3.7 and 3.8 respectively.

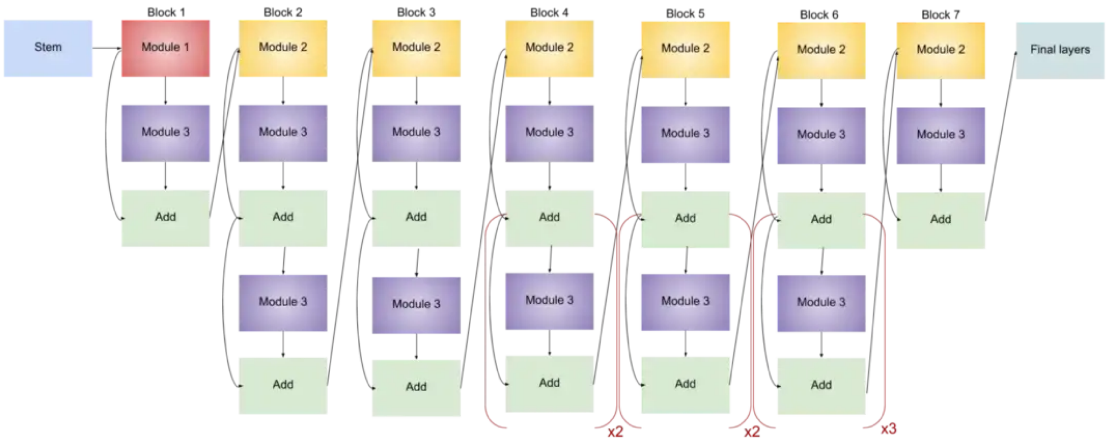

Figure 3.5: *EfficinetNet-B2 architecture composition.*

In figure 3.6, the three building sub-blocks of each of the seven blocks are represented:

- Sub-block 1 is only used as the first sub-block in the first block.
- Sub-block 2 is used as the first sub-block in all the other blocks.
- Sub-block 3 is used for any sub-block except the first one in all the blocks.

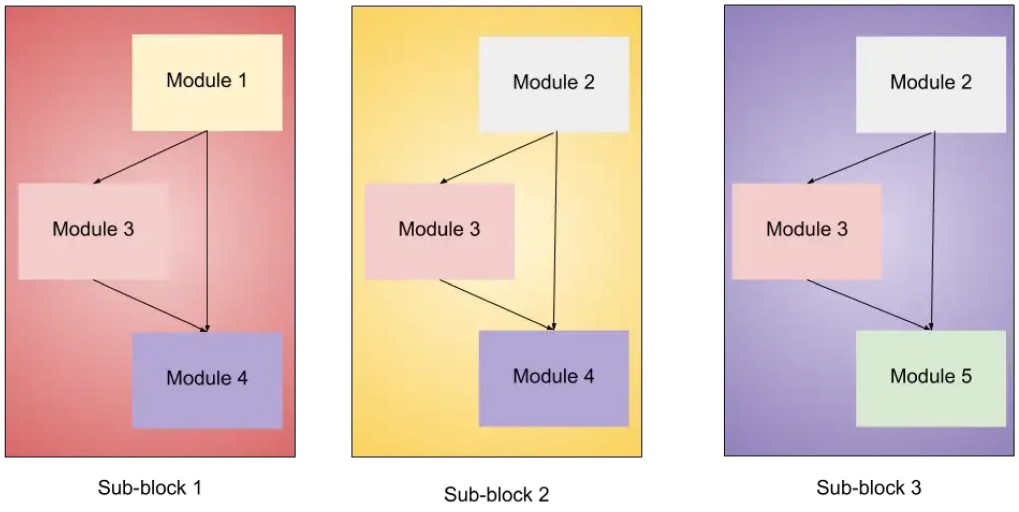

Figure 3.6: *Combination of modules to build the 3 sub-blocks.*

Each of these sub-blocks is obtained combining the five modules represented in figure 3.8 as building blocks. In particular:

- Module 1 is used as a starting point for sub-block 1.
- Module 2 is used as a starting point for sub-block 2 and 3.
- Module 3 is used as a skip connection in each of the three sub-blocks.
- Module 4 is used for combining the skip connection in sub-block 1 and 2.
- Module 5 is used for combining the skip connection in sub-block 3.

Each of these modules is made combining different types of layers.

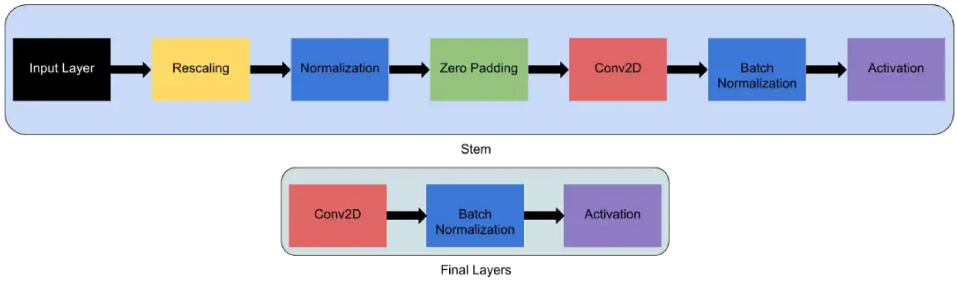

Figure 3.7: *Combination of layers building the stem module.*

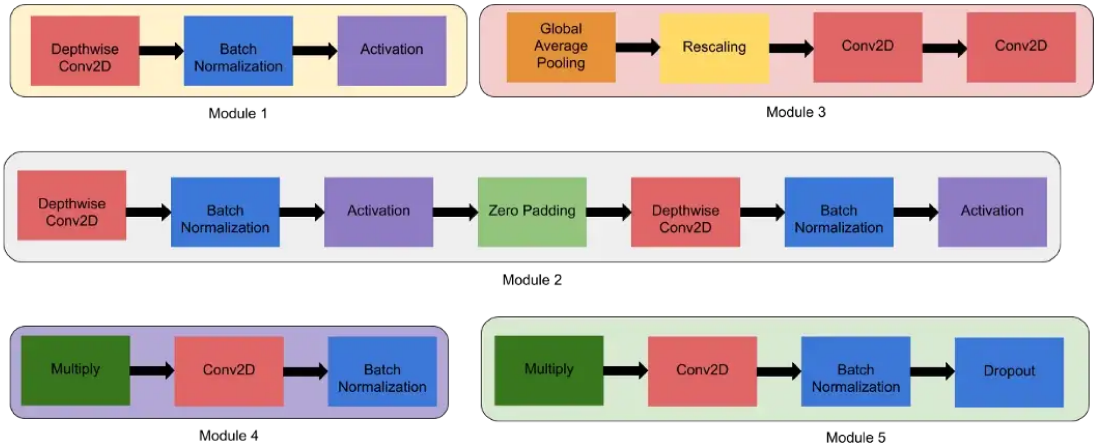

Figure 3.8: *Combination of layers building the 5 modules.*

3.6 Conclusions

This concludes the chapter covering the theory about the technologies used to carry out the work. In particular, a quick overview of intelligence has been provided, going from the explanation of how a single neuron works to the introduction to some more complicated architectures. The clever idea behind the conception of the EfficientNet family has also been provided. The next chapter will instead treat implementation aspects, such as the environment in which the work has been carried out, the preprocessing performed on the images, the details about the network parameters and some insight into the data management.

4. Work description and implementation

4.1 Introduction

The last chapter dived into detail about neural networks and about the specific family that has been used in this work, EfficientNet. In addition, it gave a general understanding of the reason why such technologies are so groundbreaking. This fourth chapter will instead move on to provide some details about the implementation, in particular about the deep learning framework that has been used. First, the environment in which the research was carried out will be presented. Then a description of how the data has been handled will follow, along with an explanation of how the images were pre-processed. To conclude the chapter, a description of the learning framework and the rationale behind the experiments will be provided.

4.2 Safe Haven: our work environment

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

Some restrictions are imposed to ensure the safe use of research data. Internet access and application installation are disabled in the Safe Haven environment, so only the applications installed by HIC service are accessible. Copying research data supplied by HIC out of this environment is not permitted either. After approval from the HIC Data Analyst, analysis results such as reports, summaries, and graphs with no patient-level data can be removed or exported.

To be granted access to the Safe Haven environment, it is required to attend the "Good research practice: principles and guidelines" course held by the Medical Research Council (MRC) [7] and pass the related exam. 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 efficient 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 high-quality research practice is fundamental and fostering a culture that supports it and aims to prevent research misconduct is a duty for research organizations.

Good research practice provides strong foundations for a research career, supporting high-quality education and training, it delivers assurance to those whose 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 guidelines for 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. They should also take appropriate steps to manage data appropriately, maintain confidentiality and minimize 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.

4.3 Data preparation and preprocessing

As it was mentioned in the dedicated section 1.5, the GoDARTS comprises 18306 subjects. To carry out the analysis on HbA1c exposure, a subset of the full dataset has been used, comprising 16799 subjects and 100153 images. Since our interest was in cumulative HbA1c as stated in section 1.1, for each image it has been calculated the respective cumulative value using the trapezoidal integration method on the HbA1c measures ordered in time. This dataset made of image names and cumulative HbA1c measurements has been then merged to another one containing other information like the patient age at image, the image date and so on.

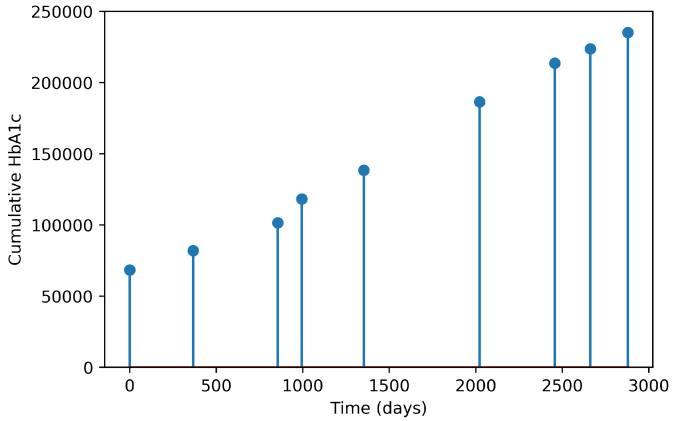

Figure 4.1: *Cumulative glycated haemoglobin measures of a patient over time*

In order to train the network effectively, a series of pre-processing operations have been performed on the retinal images. First, the excess black areas around the fundus have been removed. Then, since images were present in different resolution formats in the dataset, they have been resized to 260x260 pixels. This resolution appears to be the one achieving the highest performances for the EfficientNetB2 used. After, the images underwent contrast-limited adaptive histogram normalization on the three R, G and B channels separately. The last operation consisted in normalizing the pixel intensities to the range [0,1].

4.4 Description of the experiments

At the beginning of this study, the main goal was to focus on cumulative HbA1c as a time series and to use fundus images chronologically to learn whether changes in the vasculature could help forecast its trend evolution in the future. Before trying to do so, we conceived a preliminary experiment to understand if the first image in time was already carrying the information of the last cumulative HbA1c measurement available. To do so, a dataset where the first image (baseline) of each subject was associated with its last HbA1c measurement has been built. The results and discussion are reported in section 5.2.

Then, we moved on to a different experiment, in which the aim was to predict the cumulative HbA1c level taken at the time of imaging using the respective fundus photograph. The results and the discussion are in section 5.2. Since in this case actual values and predictions were both available over time for each subject, it has been investigated if the predictions over time were growing coherently with the actual values. To do so, both of them have been plotted against the time in days, calculated as the difference between image dates and taking the first image as 0. The analysis of the results and the discussion can be found in section 5.4.

4.5 Learning framework

All the code, from model development to results analysis, was written in Python 3.6. In particular, the libraries OpenCV and Scikit-learn were used for image processing, while Keras 2.2.2 and TensorFlow 1.9.0 for building, training and testing DL models. The Scikit-learn package has also been used to build linear models to fit values and predictions over time.

As mentioned earlier, the EfficientNetB2 network has been used to predict the cumulative HbA1c values in both the experiments from the retinal images. 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 unchanged, and their weights were initialized with weights pre-trained on ImageNet.

Since whenever images were collected from a patient they were taken from both eyes, the total dataset has been split into right and left eye datasets. The training process and the

predictions were made independently for each eye and then averaged, according to previous literature. To train each network, the datasets have been split into a training set (70%), a validation set (10%) and a test set (20%). It is relevant to underline that the split was made on individuals and not on images, in order to avoid information leakage: images of the same individuals were used either for training or testing, never for both. The composition of the datasets in the second experiment are reported in table 4.1, while the size of the datasets used in the first one are equal to the number of subjects present in the second, since only the baseline image of each subject has been used.

Dataset	Eye	Images	Males	Females	Subjects
Training	Left	36195	3412	2638	6050
Training	Right	36191	3413	2637	6050
Validation	Left	3854	378	288	666
Validation	Right	3857	378	290	668
Test	Left	10030	926	755	1681
Test	Right	10026	928	756	1684
Total		100153	9435	7364	16799

Table 4.1: *Composition of the datasets used in the learning framework.*

Mean squared error was used as loss function, while Adam optimization with Nesterov accelerated gradient momentum was used as gradient descent algorithm. The initial learning rate was set to 0.001, reduced by a factor 0.1 if the validation loss did not improve within 5 consecutive epochs. The minimum learning rate was set to 10^{-5}. The model was trained for 50 epochs, with batch size 32. The training set has also been augmented with random horizontal flips and random rotations. In the validation phase, the model training was stopped if there was no improvement in the validation loss for 20 consecutive epochs. The weight set associated with the best validation performance was saved for testing.

4.6 Conclusions

This chapter was strongly implementation oriented, with the aim of providing details about many aspects of this work. As mentioned earlier, in order to respect patients' privacy, the work has been conducted inside the Safe Haven. Then, it has been described how the data were obtained from the measurements and how the images were pre-processed. Afters, the aim of the work has been stated in detail, along with the description of the experiments performed in order to pursue it. Lastly, a description of the learning framework has been provided, with details about both the training of the network and its optimization algorithm. The next chapter will give an insight into the results of these experiments and discuss them critically.

5. Results and discussion

5.1 Introduction

After having explained the goal of the work and the experiments set up to achieve it in the previous chapter, this one will move on to the description and the discussion of the results. In particular, the first section will describe the general results, comparing the actual cumulative Hba1C values to the predicted ones, as well as analyzing the prediction error. Then, an analysis of the time trend will be provided, along with the linear fit of the measures. Slope and intercept of this fit will then be analyzed in the following section, before diving into the discussion of the limitations of this work.

5.2 General results

This section will illustrate and discuss the predictions obtained on glycated hemoglobin from fundus camera images. For both of the experiments conducted, the results have been analyzed qualitatively with the scatter plots of the predicted against actual cumulative HbA1c and of the absolute prediction error (actual value - predicted value) against the actual value, as well as with the comparison between the histograms of the actual and the predicted HbA1c values. On the other hand, by a quantitative point of view, the goodness of the predictions has been evaluated with the Pearson correlation coefficient, the mean absolute prediction error (MAE) and the median relative prediction error (RE). In the following, both plots and results' tables are reported.

Dataset	Pearson	MAE	RE (%)
Left	0.206	196482.96	60.47
Right	0.161	191764.84	58.94
Average	0.198	193799.32	59.41

Table 5.1: *Experiment one prediction performances.*

Dataset	Pearson	MAE	RE (%)
Left	0.345	118508.37	44.06
Right	0.312	125465.53	46.19
Average	0.355	119771.56	44.13

Table 5.2: *Experiment two prediction performances.*

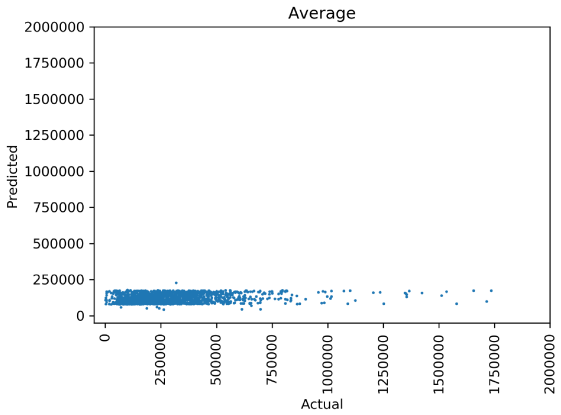

Figure 5.1: *Experiment 1, scatter plot of predicted versus actual cumulative HbA1c*

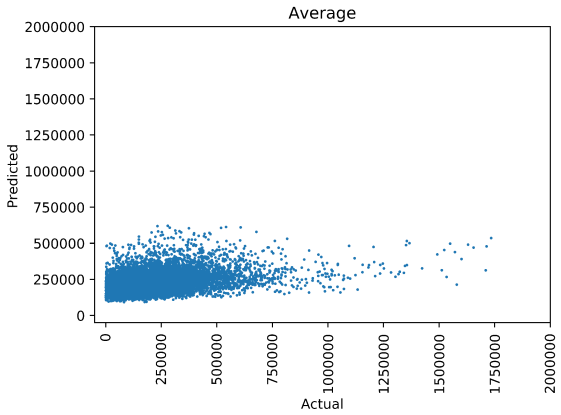

Figure 5.2: *Experiment 2, scatter plot of predicted versus actual cumulative HbA1c*

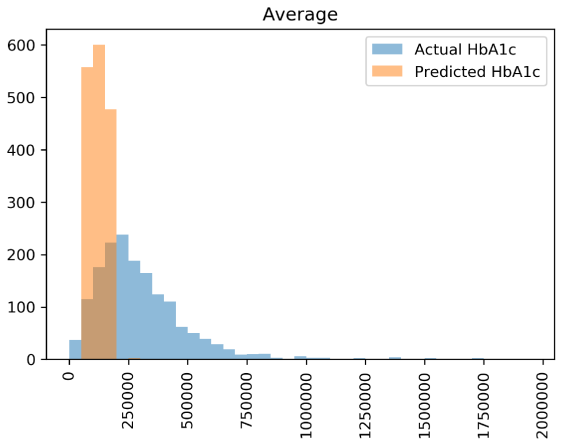

Figure 5.3: *Experiment 1, histogram of predicted versus actual cumulative HbA1c*

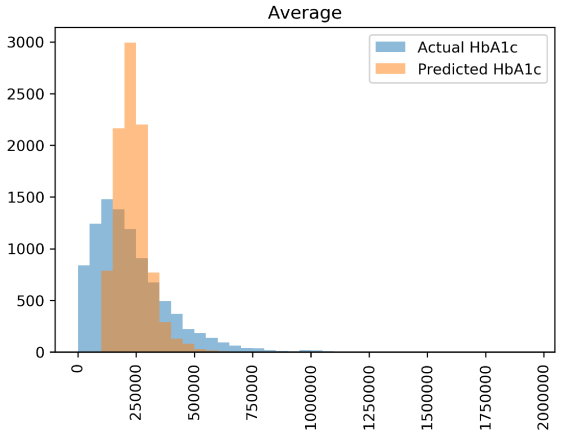

Figure 5.4: *Experiment 2, histogram predicted versus actual cumulative HbA1c*

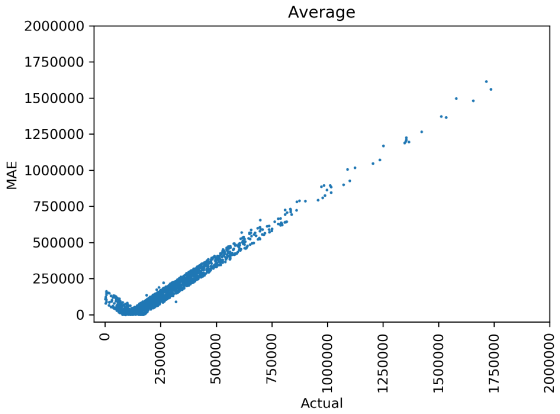

Figure 5.5: *Experiment 1, prediction error versus actual cumulative HbA1c*

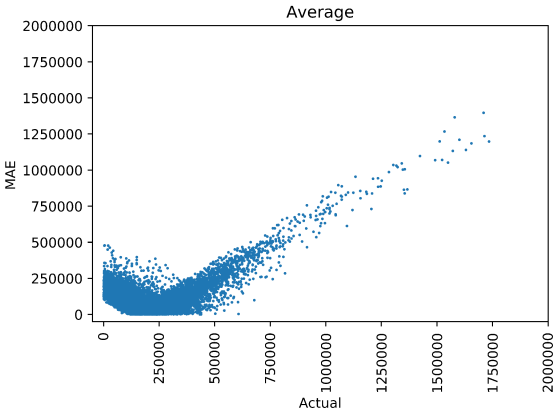

Figure 5.6: *Experiment 2, prediction error versus actual cumulative HbA1c*

As it emerges from the scatter plot in figure 5.1, in the first experiment there is an evident incapability of the network to estimate the true cumulative HbA1c value of the last measurement using the baseline image. In particular, the network seems to be assigning values from a uniform distribution in the range 75000 - 175000. Indeed, it's noticeable from both the scatter plot in figure 5.1 and the correlation values in table 5.1 that the association between the predicted and the true values is very low. The main reason and explanation for this fact is that the baseline image is not likely to contain information about the time progression of the HbA1c. Another factor to take into consideration is that using only baseline images, the number of training examples is greatly reduced. Looking instead at figure 5.5, where the prediction error is reported against the true value, it is clear that the error is increasing proportionally with the measure, revealing that the higher is the cumulative HbA1c value, the more difficult it is for the network to predict it accurately.

Similar considerations can be made for the second experiment. Indeed, even if the scatter plot in figure 5.2 is more promising, the correlation is still low. Also, the prediction error is still highly related to the actual cumulative HbA1c value, as shown in the plot in figure 5.6. After having tried in different ways to improve the quality of the prediction, it has been decided to discard the first experiment and to focus on the results of the second one. The further analysis conducted is reported in the following section.

5.3 Effect of sex and age

An initial analysis has been conducted, studying the effect of age and sex on cumulative HbA1c by a qualitative point of view. In figures 5.7 and 5.8, actual and predicted cumulative HbA1c respectively have been plotted against age at image, with different marker's color depending on sex. What emerges from these graphs is that the values are equally distributed in males and females, and the fact that the equal distribution is also present in the predicted values is a positive aspect.

Concerning instead the age, cumulative HbA1c does not seem to have a trend related to it, in neither males nor females. This consideration can be made for both the predicted and the actual values. This could be due to the cumulative nature of the variable under study: images from a younger patient at a late acquisition have a higher cumulative HbA1c value than images from an older patient at an early acquisition. This could hide the relationship between non-cumulative HbA1c and age: further investigation could be carried out on the effect of age on non-cumulative HbA1c.

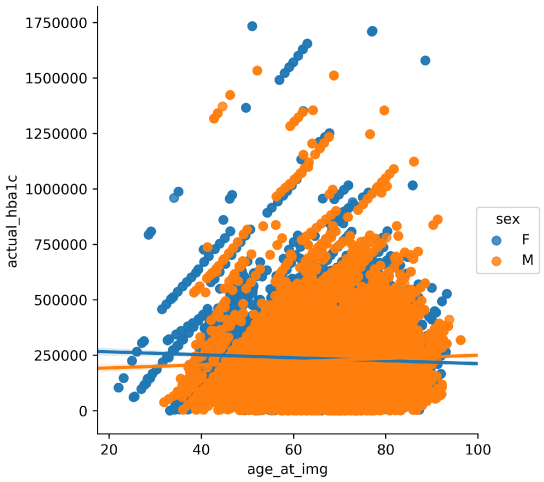

Figure 5.7: *Actual cumulative HbA1c against age.*

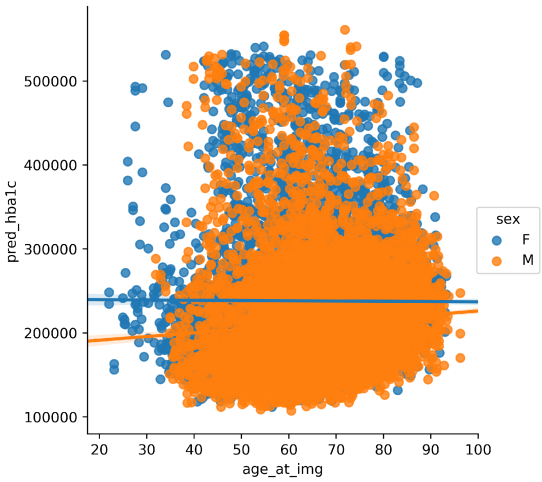

Figure 5.8: *Predicted cumulative HbA1c against age.*

5.4 Predictions over time

After the overall analysis of the predictions, the focus of the following phase has been put on time trend of the predictions. Since images and respective HbA1c values were collected over time, both the predictions and the true values have been plotted against time from the first image (baseline) in days. In order to have a clearer understanding of the increase over time of the cumulative HbA1c level, both the predicted and true values over time have been fitted with a linear model. In this way, the slope and the intercept calculated for each subject have been used for further analysis, reported in the next section. An example of this procedure has been reported for one subject in figure 5.9.

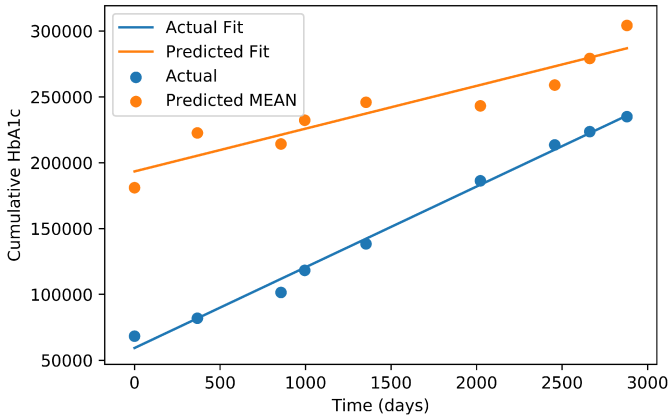

Figure 5.9: *Predicted and actual cumulative HbA1c values and relative fits.*

This situation reflects the most frequent prediction scenario, where both the actual and predicted time trends have a positive slope. Every subject has been classified with a label, indicating the position of the predicted fit with respect to the actual one. The three possible values for the label are: "Above" when the predicted value is greater than the actual one at each time point, "Below" on the opposite situation and "Interception" otherwise. This was done to investigate whether an always greater prediction could be associated with a greater risk of death. Similarly, it has been checked whether the predicted slope was greater than the actual one or not, to see if this could be another indicator of risk for the patient. In this way, a variable called "risk" has been created, with value 1 when this was the case and 0 when it was not. In the next section, this analysis is reported together with the discussion.

5.5 Slope and intercept analysis

Having obtained the three classes "Above", "Below" and "Interception" looking at the fits' relative position and a class as potential risk indicator comparing the predicted slope magnitude to the actual one, the following analysis concentrated the efforts into understanding if these two labels could tell something useful about the patient health status progression. In order to assess how close the parameters of the predicted fit are, a scatter plot of the predicted values versus the actual ones and a histogram of their distributions have been reported for the two fit parameters (intercept and slope), as it has been done for the cumulative HbA1c predictions. They are reported in figures 5.12 and 5.13 respectively.

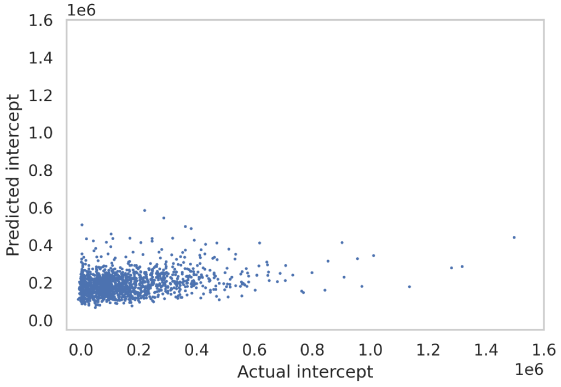

Figure 5.10: *Predicted versus actual fit's intercept scatter plot*

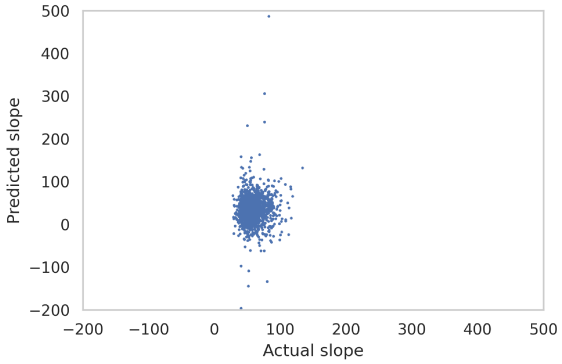

Figure 5.11: *Predicted versus actual fit's slope scatter plot*

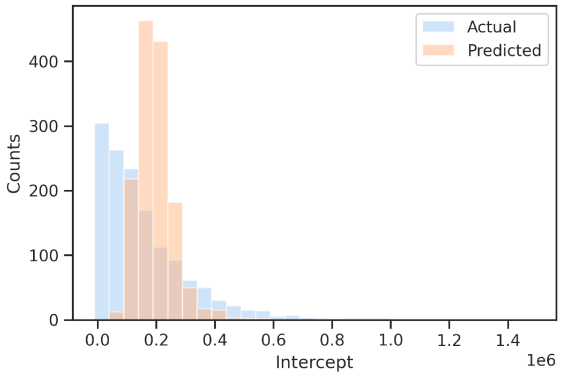

Figure 5.12: *Predicted and actual fit's intercept histogram*

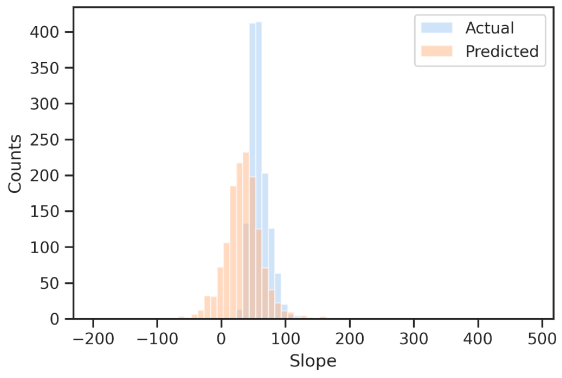

Figure 5.13: *Predicted and actual fit's slope histogram*

The values distributions of the intercepts reflect the distribution of the predictions reported in section 5.2, where the network was estimating the values in a tighter interval than the real one. On the contrary, concerning the slope, the situation is the opposite, since the predictions' distribution is wider than the one of the actual values. A positive aspect is that the center of the predicted values' distribution is located in the value of maximum frequency of the true values' distribution.

As already mentioned, the following analysis consists in discovering if the hypothesis of an association between the two fits' relative position or predicted slope magnitude with higher risk can be accepted or rejected. The focus was put into the slope analysis, since it is more physiologically relevant than the intercept. In order to do so, the death outcome available in the test set made of 1395 subjects has been inspected, and in particular it has been used to understand whether the frequency of its classes (0-1) was significantly different among the three position classes or between the two risk levels. The frequency investigation has been performed qualitatively, looking at the bar plot of the death outcome frequency inside the different classes. In order to test the relationship between these 2 couples of categorical variables, the Chi-squared test has been used, and the results are reported in the following.

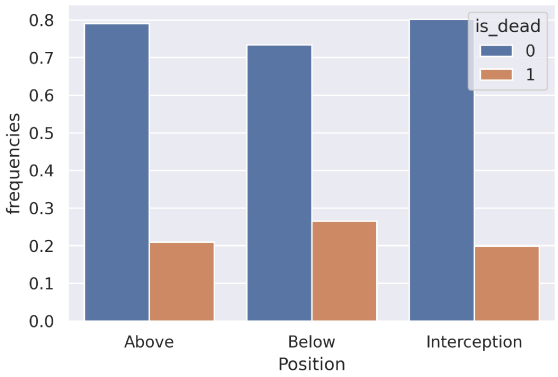

Figure 5.14: *Bar plot of the death outcome frequency among position classes*

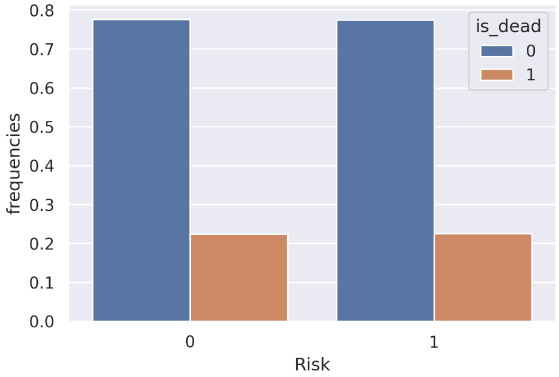

Figure 5.15: *Bar plot of the death outcome frequency between risk classes*

Position	# dead	# alive
Above	155	583
Below	109	301
Interception	45	198

Table 5.3: *Contingency table for death status and position.*

Risk	# dead	# alive
0	257	890
1	56	192

Table 5.4: *Contingency table for death status and risk.*

χ^2	p-value	Significant
5.88	0.052	NO

Table 5.5: χ^2 *test results for position.*

χ^2	p-value	Significant
0.0036	0.95	NO

Table 5.6: χ^2 *test results for risk.*

Concerning the analysis of the position, looking at the bar plot in figure 5.14, the frequency of the death outcome classes appear to be very similar among the three fit relative position classes. For this reason, the hypothesis that a predicted cumulative HbA1c always greater than the actual one could be a risk indicator seemed not to hold. In order to confirm this, a Chi-squared test (χ^2) was conducted, using the contingency table 5.3 and obtaining the results in table 5.5. A threshold of 0.05 was used for the p-value in order to decide whether to discard or not the null hypothesis that no relationship exists between the categorical variables. This led to state that there is no dependence between the two variables, since the obtained value is greater than the threshold. This outcome also means that the hypothesized explanation for the all-above or all-below prediction does not hold. For this reason, the situations just mentioned do not mean a greater or lower risk respectively.

Very similar considerations can be made for the risk variable, with value 1 when the predicted slope was grater than the actual one, and with value 0 otherwise. Also in this case the hypothesis that having this feature with value 1 could be associated with a higher death risk seems not to hold. Indeed, it emerges from the bar plot in figure 5.15 that the frequencies of death and survival are the same in the two class' values. Once again, the χ^2 test turned out useful to test the relationship between these 2 categorical variables and provide a quantitative result. Since the p-value is greater than threshold also in this situation, the hypothesis of the 2 variables not being related cannot be discarded.

5.6 Limitations of this work

As the results may suggest, this study comes with some limitations. First to mention is the nature of the data. As it has already been mentioned, the samples and the images were not taken at regular intervals and, more importantly, the intervals across subjects are not regular either. Also, the age at baseline is different basically for every patient. All this factors increase the data heterogeneity remarkably, increasing the difficulty for the network to predict the values accurately.

Another big limitation is the cumulative nature of the data. In previous studies, mentioned in section 2.4, the efforts always focused on the prediction of non-cumulative HbA1c, expressed in percentage value in respect to the total haemoglobin. In this case, instead, it was expressed in mmol/mol and it was cumulative: for this reason, values associated to an image were carrying information about former images as well. The network probably could not learn this, since it had no means to know whether an image was the first collected for a subject or not. In section 6.4 a brief discussion about how to overcome these limitations will take place.

5.7 Conclusions

This chapter presented, analyzed and discussed the results obtained from the experiment described in chapter 4. The first section presented the general results about cumulative HbA1c prediction, using scatter plots and histograms as visualization tools and Pearson correlation, mean absolute error and relative error as analytical tools. The following section quickly studied if sex and age had any effect on cumulative HbA1c with qualitative scatter plots, suggesting that in this dataset there is none. In the third section, the predictions and the actual values have been studied over time and fitted with a linear model. Two categorical variables have been created with the parameters of these fits, the "position" and the "risk" class. In the following section, the relationship between these classes and the death outcome has been studied with bar plots and χ^2 test. To conclude, the limitations of this study, mainly related to the nature and format of the data, have been presented. In the next and last chapter, final considerations about this work will be made and the future work and implementations will be described.

6. Conclusions and future work

6.1 Introduction

The previous chapter presented the results obtained with the 2 experiments and the analysis carried out on them. It analyzed the effect of age and sex on this variable, confirming they have no influence on it. Then it presented the time trend analysis, in which the predicted and the actual data have been fitted and the parameters extracted. Exploiting these parameters, two categorical variables have been created. The hypothesis of their association with the death outcome of the patients has been tested with a χ^2 test, that led to discard this hypothesis.

This chapter, being the last, will summarize the work and its main findings. It will then present the main limitations of this work, that are mainly related to the cumulative nature of the data and to their big range of variation. To conclude, it will provide an overview on how to overcome some of these limitations, in order to make further progress on this topic and other possible approaches to solve tackle this problem.

6.2 Work summary

To summarize, this work started with the aim to discover whether it is possible to use temporal information present in retina images to predict the trend of cumulative glycated haemoglobin (HbA1c) in the future. In order to do so, two experiments have been designed: the first one trying to predict the last HbA1c measurement using the baseline image and the second trying to predict the HbA1c measurement relative to each image. The results showed in section 5.2 revealed that the first experiment is not feasible, likely

because the baseline image does not possess any information of the future progression of cumulative HbA1c that the network can learn. The second experiment has shown more promising results, but there is still large room for improvement, for the many reasons concerning the data discussed in section 5.6.

Afterwards, an analysis of the effects of age and sex on cumulative glycated haemoglobin has been conducted. It has shown that in this data set of cumulative measure, they seem to have no particular effect, since the distribution is the same in males and female, and it does not have any trend with age.

The following part of the work dealt with the time trend of the actual and predicted data. Both the data points have been fitted with a linear model and the parameters (slope and intercept) have been extracted for each subject. On the base of the relative position of the two fits they have been classified as "Above", "Below" or "Interception" and on the base of the magnitude of the predicted slope with respect to the actual slope they have been given a "risk" score of 0 or 1. Statistical χ^2 test has been performed in order to establish whether a relationship between these 2 variables and the death outcome existed or not.

6.3 Main findings

The first experiment allowed confirming that the network is unable to predict the future cumulative HbA1c value from the baseline image. This is likely because the retina does not have enough information about it yet, and because the time gap between each baseline image and respective last HbA1c measurement is different in each subject. This is also in agreement with eye physiology and anatomy, since a single retinal image holds information about the past medical history rather than future outcome progression.

The second experiment showed that the prediction of glycated haemoglobin is a difficult task to perform, at least using EfficientNetB2. But the results are way more encouraging than the first experiment, since there is coherence in the time trend of the predictions. The statistical test and the bar plot proved that there is no relationship between the death outcome and the classes created using the parameters of the fits. This suggests that the "all-above" and "all-below" predictions with respect to actual values cannot be related to a capability of the network to discover a higher or lower risk. The same consideration can be made for the risk factor.

6.4 Improvements and future work

Being one of the first studies using retinal images to predict cumulative HbA1c, the ways to bring improvement to this work are many. The first way could be to use different networks instead of EfficientNetB2 and see whether this can lead to better predictions or not. But the mean absolute error and the relative error magnitude can suggest that the way to bring the most consistent improvement is in the data. Since cumulative haemoglobin can be calculated using the values of the non-cumulative one, integrating them over time, predicting non-cumulative HbA1c and calculating cumulative one using the predictions, could lead to more accurate results. Indeed, in previous studies reported in section 2.4, non-cumulative HbA1c has been predicted in percentage value with high performances.

Concerning instead the way to exploit longitudinal information, being in this case HbA1c measurements over time, an interesting source of possible solution could be the method purposed by J. Bridge et al. [5] (2020). In this study, the team developed a tool to prognosticate age-related macular degeneration using longitudinal images with a framework made up of three stages. In the first one, a convolutional neural network (Inception V3) is used to automatically perform feature extraction from the retinal images. Afterwards, each vector is scaled by a factor which is inversely proportional to the time distance between the image time point and the time point at which the prediction is wanted. In this way, images closer to the prediction time point are given more importance than the ones further in time. In the final stage, the feature vectors are concatenated in a matrix and fed into a recurrent neural network, a gated recurrent unit (GRU), that predicts whether the macular degeneration is progressing or not. This method could be adapted to predict a continuous variable such as HbA1c, but it has the limitation that the number of time points for each subject has to be the same.

References

[1] https://www.diabete.com.

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

[3] Italian society of diabetology. http://www.siditalia.it.

[4] Liam Boyle. Data Analysis using GoDARTS Data Sets and Machine Learning. 2020.

[5] Joshua Bridge, Simon Harding, and Yalin Zheng. Development and validation of a novel prognostic model for predicting amd progression using longitudinal fundus images. *BMJ open ophthalmology*, 5(1):e000569, 2020.

[6] James H Cole, Stuart J Ritchie, Mark E Bastin, Valdés Hernández, S Muñoz Maniega, Natalie Royle, Janie Corley, Alison Pattie, Sarah E Harris, Qian Zhang, et al. Brain age predicts mortality. *Molecular psychiatry*, 23(5):1385–1392, 2018.

[7] Medical Research Council. MRC Ethics Series, Good Research Practice: Principles and Guidelines, 2012.

[8] International Diabetes Federation. Diabetes. https://www.who.int/health-topics/diabetes#tab=tab_1.

[9] Nele Gerrits, Bart Elen, Toon Van Craenendonck, Danai Triantafyllidou, Ioannis N Petropoulos, Rayaz A Malik, and Patrick De Boever. Age and sex affect deep learning prediction of cardiometabolic risk factors from retinal images. *Scientific reports*, 10(1):1–9, 2020.

[10] Harry L Hebert, Bridget Stepherd, Keith Milburn, Abirami Veluchamy, Weihua Meng, Fiona Carr, Louise D Donnelly, Roger Tavendale, Graham Leese, Helen M Colhoun, Ellie Dow, Andrew D Morris, Alexander S Doney, Chim C Lang, Ewan R Pearson, Blair H Smith, and Colin NA Palmer. Cohort Profile: Genetics of Diabetes Audit and Research in Tayside Scotland (GoDARTS). *International Journal of Epidemiology*, 47(2), 2018.

[11] Yong Dae Kim, Kyoung Jin Noh, Seong Jun Byun, Soochahn Lee, Tackeun Kim, Leonard Sunwoo, Kyong Joon Lee, Si-Hyuck Kang, Kyu Hyung Park, and Sang Jun Park. Effects of hypertension, diabetes, and smoking on age and sex prediction from retinal fundus images. *Scientific reports*, 10(1):1–14, 2020.

[12] Andreas Maier, Christopher Syben, Tobias Lasser, and Christian Riess. A gentle introduction to deep learning in medical image processing. *Zeitschrift für Medizinische Physik*, 29(2):86–101, 2019.

[13] The University of British Columbia. Ophthalmic photography. `https://ophthalmology.med.ubc.ca/patient-care/ophthalmic-photography/`.

[14] World Health Organization. Diabetes around the world in 2021. `https://diabetesatlas.org/`.

[15] World Health Organization et al. Use of glycated haemoglobin (HbA1c) in diagnosis of diabetes mellitus: abbreviated report of a WHO consultation. Technical report, World Health Organization, 2011.

[16] Ghouse Syed, Emanuele Trucco, Chim C Lang, Yu Huag, Ify Mordi, and Alex Doney. Biological vascular age determined from retinal photographs used for diabetes retinal screening as a predictor for all-cause death and cardiovascular events.

[17] Mingxing Tan and Quoc Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International conference on machine learning*, pages 6105–6114. PMLR, 2019.