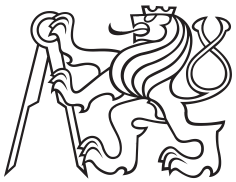
Bachelor's thesis



Czech Technical University in Prague

F3

Faculty of Electrical Engineering
Department of Cybernetics

Diabetic Retinopathy Detection Using Neural Networks

Vojtěch Pořízek

Supervisor: prof. Dr. Ing. Jan Kybic
August 2020

BACHELOR'S THESIS ASSIGNMENT

I. Personal and study details

Student's name: **Pořízek Vojtěch** Personal ID number: **474385**

Faculty / Institute: **Faculty of Electrical Engineering**

Department / Institute: **Department of Cybernetics**

Study program: **Cybernetics and Robotics**

II. Bachelor's thesis details

Bachelor's thesis title in English:

Diabetic Retinopathy Detection Using Neural Networks

Bachelor's thesis title in Czech:

Detekce diabetické retinopatie pomocí neuronových sítí

Guidelines:

Develop a method based on deep convolutional neural networks for solving the task of the detection of diabetic retinopathy from digital color fundus images of the retina, as defined by the Kaggle Diabetic Retinopathy Detection challenge. Get familiar with related work from the literature and develop a baseline solution based on standard techniques such as Inception or DenseNet networks. Investigate the possibility of performance improvement using one or several of the following techniques: (1) correlation between the two eyes of the same person, (2) using unlabeled data, (3) ensemble classification by several different networks, (4) using additional inputs such as blood vessel segmentation, (5) attention networks, (6) probabilistic neural networks.
Evaluate your results experimentally on the provided datasets and submit your solution to this or related online challenges to compare the performance of your method with state of the art.

Bibliography / sources:

[1] Kaggle Diabetic Retinopathy Detection Challenge https://www.kaggle.com/c/diabetic-retinopathy-detection
[2] E. J. Duh, J. K. Sun, and A. W. Stitt, "Diabetic retinopathy: current understanding, mechanisms, and treatment strategies", JCI Insight, vol. 2, no. 14, Jul. 2017
[3] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jonathon Shlens, Zbigniew Wojna "Rethinking the Inception Architecture for Computer Vision", CVPR 2016, pp. 2818-2826

Name and workplace of bachelor's thesis supervisor:

prof. Dr. Ing. Jan Kybic, Biomedical imaging algorithms, FEE

Name and workplace of second bachelor's thesis supervisor or consultant:

Date of bachelor's thesis assignment: **09.01.2020** Deadline for bachelor thesis submission: **14.08.2020**

Assignment valid until: **30.09.2021**

prof. Dr. Ing. Jan Kybic Supervisor's signature	doc. Ing. Tomáš Svoboda, Ph.D. Head of department's signature	prof. Mgr. Petr Páta, Ph.D. Dean's signature

III. Assignment receipt

The student acknowledges that the bachelor's thesis is an individual work. The student must produce his thesis without the assistance of others, with the exception of provided consultations. Within the bachelor's thesis, the author must state the names of consultants and include a list of references.

. | Date of assignment receipt

Student's signature

Acknowledgements

Firstly, I would like to express gratitude to the prof. Jan Kybic, supervisor of this thesis, for his endless support, helpful suggestions and kind attitude.

Secondly, I would like to thank prof. Jarmila Heissigerova, Head of Eye Clinic at General University Hospital in Prague for the medical consultations she provided me with.

Furthermore I would like to extend thanks to the following people for their consultations on selected methods, even though some of the discussed methods did not make it into the final version.

- Eric Arazo Sanchez, Dublin City University
- Dr. Boris Flach, Center for Machine Perception at CTU Prague
- Kamil Dedecius Ph.D, Institute of Information Theory and Automation in Prague
- Magdalena Kovacova MD, Eye Clinic at General University Hospital in Prague
- Xi Fang, Shanghai Jiao Tong University

Last but not least, I would like to thank my family and friends for their undying support and encouragement throughout my studies.

This thesis is dedicated to Ruzena, Jan and Rek, who sadly passed away this year.

Declaration

I hereby declare that the presented work was developed independently and that I have listed all sources of information used within it in accordance with the Methodical instructions for observing the ethical principles in the preparation of university thesis.

Prague, August 14, 2020

Abstract

Diabetic retinopathy is a severe eye disease that causes blindness and affects most of the diabetic patients. The disease onsets without any symptoms and when the patient notices minor difficulties with vision, the retina can already be irreversibly disrupted. Prevention can be achieved by adopting regular eye checks, which are time expensive. Various methods for detection of diabetic retinopathy from fundus photographs have emerged, with the best results achieved by the convolutional neural networks.

This bachelor's thesis tries to address the task of diabetic retinopathy detection using multiple approaches. A solution for a single image classification is developed and then improved by several methods: by using state of the art architectures and methods, by blending the predictions from both eyes of a person, by using unlabeled data, by using additional input in form of vessel segmentation masks and by ensemble classification.

The author discovered that the best results of single model can be achieved by using the additional input and training on unlabeled data. The best results of ensemble were obtained by using a shalow neural network architecture. The proposed ensemble solution achieved a QWK score of 0.851 in the Diabetic Retinopathy Detection Challenge 2015, which is comparable to the literature results.

Keywords: diabetic retinopathy, convolutional neural networks, early detection, blood vessel segmentation, ensemble, pseudolabelling

Supervisor: prof. Dr. Ing. Jan Kybic
Biomedical imaging algorithms, FEE

Abstrakt

Diabetická retinopatie je závažné onemocnění oka, které způsobuje slepotu a zasahuje většinu populace diabetiků. Navenek se projevuje většinou až v pokročilých stádiích, kdy už je oko nevratně poškozeno. Prevencí jsou pravidelné kontroly, které jsou však náročné na čas. Proto se v posledních letech začaly objevovat systémy pro automatickou detekci diabetické retinopatie ze snímků sítnice. Nejlepších výsledků dosáhly metody používající konvoluční neuronové sítě.

Tato bakalářská práce prozkoumává metody pro detekci diabetické retinopatie za použití konvolučních neuronových sítí. Nejprve je představeno základní řešení a to je poté vylepšeno za pomoci nejmodernějších architektur a metod; za pomoci využití korelací z obou očí jednoho pacienta; za pomoci využití neoznačených dat; za pomoci dodatečného vstupu ve formě osegmentovaných masek cévního systému a za pomoci souboru více sítí.

Autor zjistil, že nejlepších výsledků při využití jednoho modelu je dosaženo za použití dodatečného vstupu ve formě osegmentovaných masek a neoznačených dat. Přesnost detekce je dále možné zlepšit využitím mělké neuronové sítě, která spojuje predikce z více modelů. Navržené řešení dosahuje v soutěži Diabetic Retinopathy Challenge 2015 skóre QWK 0.851, což je srovnatelné s výsledky popisovanými v literatuře.

Klíčová slova: diabetická retinopatie, konvoluční neuronové sítě, včasná detekce, segmentace cév, ensemble, data bez označení

Překlad názvu: Detekce diabetické retinopatie pomocí neuronových sítí

Contents

Chapter 1

Introduction

Diabetic retinopathy is the leading cause of blindness in the working population of the developed world. Linked closely to diabetes mellitus, it causes structural changes in the retina, that lead to irreversible damage. As the medical retinal exam has to be performed regularly among diabetic patients, it is costly in terms of time and human resources.

Therefore, there has been a call for an automatic detection system, that would automatically evaluate the retinal images and refer to the ophthalmologist only the affected patients. This could result in a faster diagnostics of the patients and reduced workload of the medical staff.

The task of this thesis is to develop a method based on deep convolutional neural networks for solving the task of diabetic retinopathy detection from digital colour fundus images of the retina, as defined by the Kaggle Diabetic Retinopathy Detection challenge. A baseline solution is created and then improved by the following techniques:

- using state of the art neural networks, optimisers and loss functions
- correlation between the two eyes of the same person
- using unlabeled data
- using additional inputs such as blood vessel segmentation
- ensemble classification

The final solution is then evaluated in the related competitions.

The structure of the thesis is as follows: In Chapter 2, the medical background is introduced, describing the fundamentals of diabetes mellitus, the eye and diabetic retinopathy. In Chapter 3, the related competitions and dataset are introduced. Chapter 4 discusses the related work and state of the art solutions. In Chapter 5, the theoretical fundamentals of the proposed

techniques are introduced. Chapter 6 contains the proposed solution with its description. In Chapter 7, the results of the proposed solution are presented. In Chapter 8, the results and the proposed methods are discussed in more detail. In Chapter 9, the author summarises the methods and suggests future improvements.

Chapter 2

Medical Background

This chapter contains the medical theory behind Diabetic retinopathy. In Section 2.1, the Diabetes mellitus is described along with the pathogenesis and classification. After that, Section 2.2 outlines the general anatomy of the eye and Section 2.3 provides a more focused summary of the retina. Section 2.4 then describes the Diabetic retinopathy, its epidemiology, signs, symptoms, classification, diagnostic tools, treatment options and prevention.

2.1 Diabetes mellitus

2.1.1 Pathogenesis

Glucose is a fundamental energy source in the human body. It is used by all the organs, muscles, immune cells, or it may be transformed into more complex structures by body metabolism.

Healthy adult utilises a collection of regulatory mechanisms that maintains the levels of glucose in arterial blood on average between 4 – 5 mmol/l during regular activity. The levels peak after food intake reaching up to 10 mmol/l and can fall to the levels around 3 mmol/l during physical activity or starvation.

One of the essential glucose regulatory mechanisms is the insulin mechanism. Insulin is a protein produced by the beta-cells of islets of Langerhans in the pancreas. Releasing it into the blood causes glucose to enter the cells, lowering the glucose concentration in blood. When the islets of Langerhans are disrupted, the body cells lack their primary energy source – glucose - and the body begins to be controlled by catabolic processes. These processes can cause conditions like hyperglycemia, dehydration and acid-base imbalance (shift of pH out of the normal range).

Diabetes mellitus is a chronic metabolic disorder characterised by hyperglycemia caused by defects in insulin secretion and perception. Diabetes mellitus is on a very steep rise, predicted to reach from an estimated 382 million in 2013 to 592 million by 2035 [1]. Diabetes may include many complications, including macro- (stroke, infarct) and micro- (retinopathy, neuropathy and nephropathy) -vascular diseases [2].

2.1.2 Classification

- Type 1 is characterised by selective destruction of beta-cells, that leads to a complete insulin deficiency and life-long insulin treatment dependency. Pancreas fails to produce any insulin at all, causing glucose to remain in the bloodstream. The cause of beta-cells destruction is of autoimmune origin and usually triggered in genetically susceptible individuals by a viral infection (such as influenza viruses, rubella or herpes) [3].

- Type 2 is the most common. It is primarily characterised by insulin resistance but may involve a partial decrease in insulin production by the beta-cells. Body cells cannot adequately respond to insulin, and glucose remains in the bloodstream. The primary cause of type 2 diabetes is excessive calories intake, improper diet, little or no physical activity, stress and smoking. In most cases, type 2 manifests at around 40 years of age and the disease usually onsets slowly without any significant symptoms. Patients may or may not have to take medical insulin treatment [3], [4].

2.2 The eye

The human eye is a complex organ that allows vision. It measures approximately 22 to 27 mm in diameter and 69 to 85 mm in circumference [5]. The organ itself is protected by the orbits. Orbits are pair structures in the skull that protect the eye itself as well as the optical nerve, the ocular muscles and the lacrimal apparatus (tear production). Each orbit is comprised of seven bones. [6]. The motion of the eye is facilitated using six extraocular muscles, and the focusing is provided by three intrinsic muscles. The blood supply is provided by a branch of the internal carotid artery [5]. The eye anatomy is depicted on Figure 2.1 and can be divided into three layers:

2.2.1 Outer layer: protective function

- sclera - Comprised of collagen fibres, it is a non-transparent white protective outer layer [6].

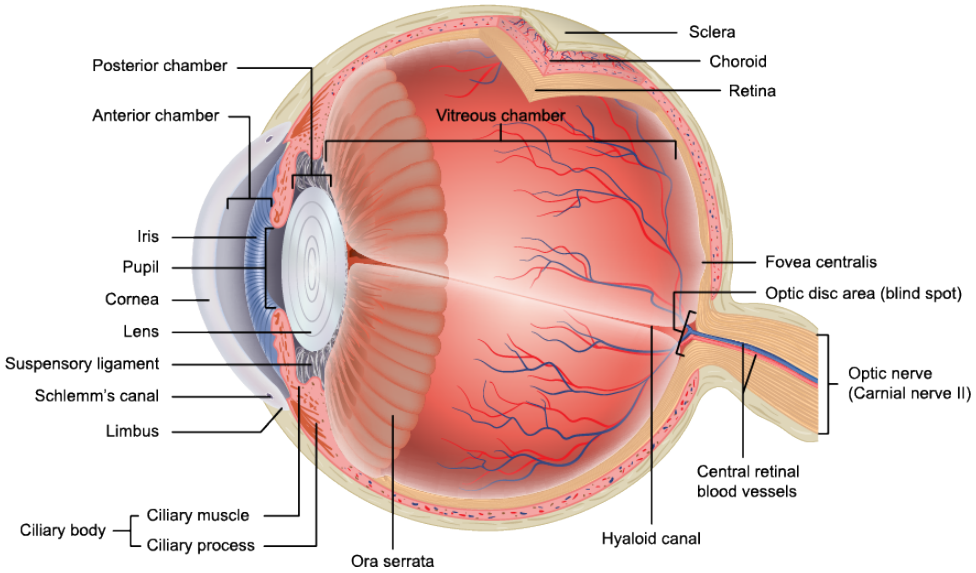


Figure 2.1: Eye anatomy, source [7].

- cornea - The cornea is a thin transparent fibrous complex structure. Aside from its protective function, the main function is to refract light [6], and it cooperates with the lens to produce a reduced inverted image [8].

2.2.2 Medium layer: nutrient function

- iris - The coloured part of the eye, it controls the amount of light entering the eye – if there is too much light, the iris closes the pupil [9].
- ciliary body - Contains ciliary muscle, that directly controls the lens.
- lens - The lens is a transparent optical element that allows focusing light onto the retina.
- choroid - A thin layer that contains most of the vessels inside the eye.

2.2.3 Inner layer:

- vitreous humour - Clear, gelatinous fluid filling the eye cavity [9].
- optic nerve - A bundle of more than a million nerve fibres that allow transmitting neurological signals from the retina to the brain [9].
- and retina, with its sensory function.

2.3 Retina

The retina is a thin transparent structure, which function is primarily to provide the photosensitive receptors. It is composed of 10 layers [6] and its thickness differ from 0.1 mm to 0.5 mm. It is adhesive to the inner eyewall (choroid and retinal pigment epithelium) [10].

There are two types of photosensitive receptors:

- cone cells - Are responsible for colour vision, need bright light, there are about 6 million of them [11].
- rod cells - Are used for recognising contrasts, work better in dim light – 92 million cells, 100 times more sensitive to a single photon than cones [12].

When looking through at the retina using the ophthalmoscope (Figure 2.2), we can distinguish the following structures [6]:

- optic disc - A bright spot, where the arteries enter, and veins and nerves leave the retina.
- macular area with fovea and foveola - It is the area of the high-acuity – the supportive vision cells (bipolar cells and ganglia cells) are deflected, letting the light to reach directly the cone cells allowing for very sharp and focused vision [8].
- nasal and temporal blood vessel arcades - Thinner being the arteries, wider being the veins.

2.3.1 Diabetic Eye Diseases

Diabetes mellitus can cause various kinds of complications in the eye, collectively called the diabetic ophthalmopathy [13] Under the term complications, there are diabetic retinopathy, diabetic maculopathy, rubeosis of the iris, secondary glaucoma, complicated cataract, diabetic neuropathy of cerebral nerves supporting ocular muscles, diabetic neuropathy of optic nerves. From the above mentioned, the most dangerous is diabetic retinopathy [13].

2.4 Diabetic retinopathy

Diabetic retinopathy (DR) is the most common microvascular complication of diabetes [14].

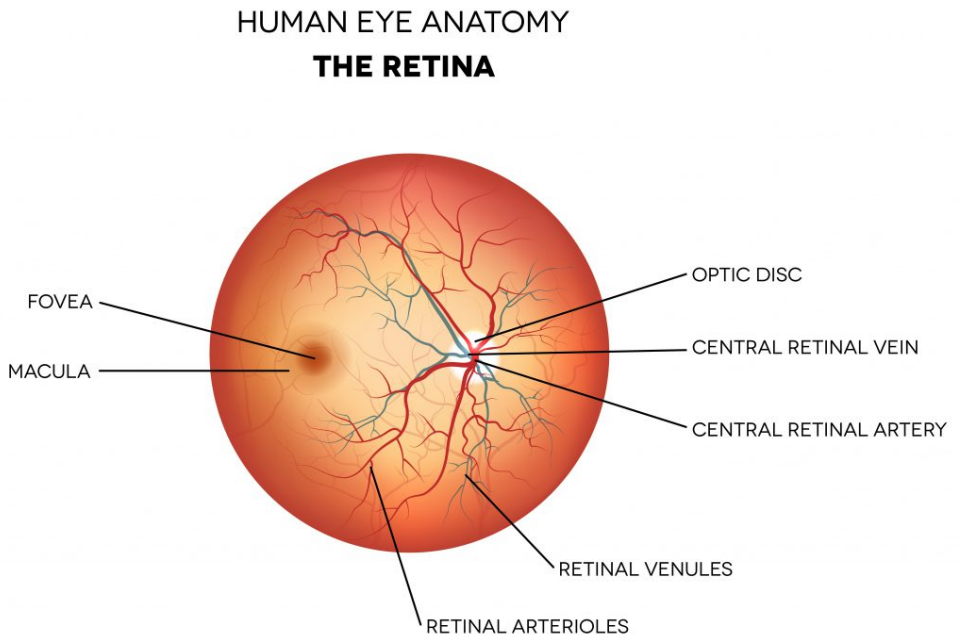


Figure 2.2: Schematic sight through the opthalmoscope, source [**?**].

It is a typical chronic progressive microvascular diabetes complication and the most common cause of blindness of patients in the productive age in the developed world [6]. It is caused by long-term hyperglycemic levels in small retinal vessels and capillaries and triggers structural and functional changes in the retinal cells [15].

2.4.1 Epidemiology

The number of diabetic patients is on a sharp rise [6]. In the Czech Republic, more than 858 thousand people were suffering from diabetes mellitus, from which 95 100 people with diabetic retinopathy, and from which 2 267 were blind because of the illness, as of 2016 [16]. Around the world, it is affecting approximately 4.2 million people worldwide. The number of Americans suffering from DR is estimated to reach 16.0 million by 2050, with vision-threatening diabetic retinopathy affecting an estimated 3.4 million of them. DR is a significant public health burden with direct medical costs accounting for 492 million USD, in addition to lost time and wages associated with receiving care [8].

2.4.2 Symptoms

Patients report various levels of symptoms while suffering from DR. This is caused by the fact, that unless the macula is degraded, the patient might not notice any subjective symptoms. The most common symptoms are [16]:

- blurred or distorted vision, that complicates reading, watching TV
- problems with balance
- increased sensitivity on light
- troubles seeing in the dark

2.4.3 Clinical signs

Clinical signs (depicted on Figure 2.3) include:

- microaneurysms: Small circular red lesions in the retina, represent vascular leakage, early signs of DR, not visually threatening, generally resolve within 3 to 4 months.
- retinal haemorages: Heamorages in different layers of retina, can be of different sizes and shapes.
- hard (lipid) exudates: Yellow irregular shaped lesions with sharp edges, represent lipids and protein depositions. When accompanied by retinal thickening, represent a feature of diabetic macular edema, usually around the macula.
- cotton-wool spots (soft exudates): White lesions with faded edges, represent nerve fibre layer infarct, usually around the optic disc, the sign of retinal ischemia [8].
- Intraretinal microvascular abnormalities (IRMA): Abnormal branching or dilation of existing blood vessels, can either be caused by the generation of new vessels or by remodelling of the existing ones [17].
- macular edema: Disturbance of the blood-retinal boundary produces leakage of plasma into the retinal tissue and swelling. It can set in in any DR stage and may result in the detachment of the retinal wall [8].
- venous beading: Dilated, irregular, tortuous veins, a non-specific sign of retinal ischemia.
- neovascularisation: An abnormal proliferation of new blood vessels. These vessels can grow into the vitreous cavity inflicting obscured vision and may result in vitreous haemorrhages.

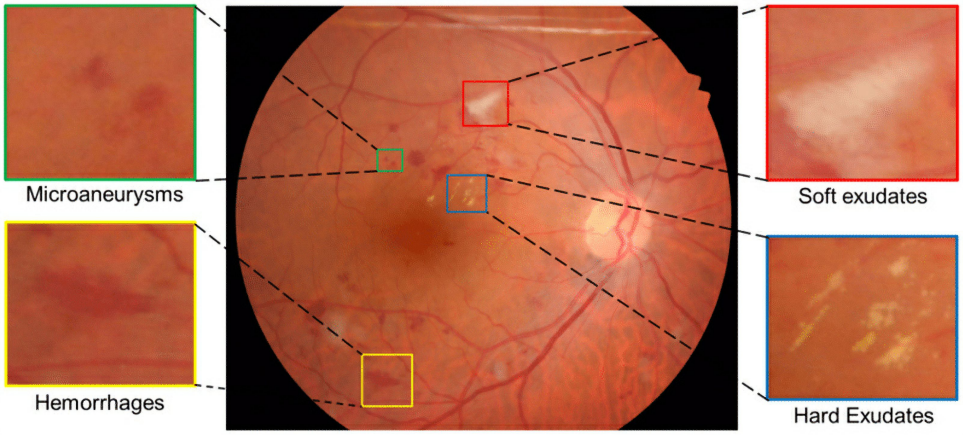


Figure 2.3: Some clinical signs of DR, source [19].

- vitreous haemorrhage: Bleeding into the vitreous cavity [18].
- fibrosis and retinal traction chages: Irreversible changes in the retinal structure, may result in a complete retinal detachment [6].

2.4.4 Retinopathy Trends

Type 1 Diabetes

DR is affecting 80 – 100 % patients. It is usually not detected when the patient is diagnosed with Type 1 DM, but after 20 years, 99 % of patients have the DR diagnosis.

Patients with Type 1 diabetes tend to present with capillary narrowing causes ischemia and hypoxia of the retina, followed by the neovascularisation. These new vessels are fragile and can easily burst and bleed (vitreous haemorrhage), causing blurred vision and eventually blindness due to fibrosis and retinal detachment.

Type 2 Diabetes

DR can be present at the time of Type 2 DM diagnosis. After 15 years, around 58 – 85 % of patients develop DR. [3, 4]

Instead of ischemia, type two diabetes is mostly seen occurring in the retinal background, causing microaneurysms and creation of macular edema, which is the most frequent cause of vision impairment (see Figure 2.4) among Type 2 diabetic patients [15].

Figure 2.4: Comparision of human vision of a healthy person and a person suffering from diabetic retinopathy, source: [20].

2.4.5 Hypertensive retinopathy

- DR is very often mistaken for a similar disease Hypertensive retinopathy (HR). As with DR, HR is identifiable by the cotton-wool spots, retinal haemorrhages, and vessel abnormalities. On the other hand, HR does not display microaneurysms and hard lipid exudates and can be distinguished from DR by the Arteriovenous nicking (a phenomenon where the small retinal veins and arteries begin to cross and bulge) [6].

2.4.6 DR Classification

According to the international classification, we classify DR into two stages – non-proliferate and proliferate, with each stage having its subclasses according to the severity and seriousness. The critical distinction separating NPDR and proliferative DR (PDR) is the presence of ocular neovascularisation. [8] Diabetic macular edema can be present at both of the stages and have its type.

- Non-proliferate diabetic retinopathy (NPDR):
  - mild NPDR – microaneurysms, retinal haemorrhage, venous beading
  - moderate NPDR - + hard exudates, cotton-wool spots, venous changes in the macula
  - severe NPDR - + IRMA, advanced retinal ischemia, haemorrhage in deeper retinal layers
- Proliferate/proliferative diabetic retinopathy (PDR):
  - mild PDR – neovascularisation of the retina and optic disc
  - high-risk PDR – neovascularisation on the iris

- advanced PDR – detachment of the retina, intravitreal haemorrhage, neovascular glaucoma [6].

2.4.7 Diagnostics

Early diagnostics is crucial for slowing down the disease and preservation of the vision and even though vision loss avoidance relies on early detection [8]. There are the following options for DR diagnostics:

The eye exam for diabetic retinopathy is to be always done in artificial mydriasis (applying eye drops to widen the pupils)[6], as is it proven to reduce the proportion of ungradable photographs from 26% to 5% [21]. There are several exams:

Fundus photography

Baseline technique, a specialised slit lamp is used to take a picture of the fundus. It consists of optical and illumination part. The optical part enables the doctors to see the retinal tissue with up to 40x magnification. The lamp also features a camera, which enables to také a retinal image. [22]. Currently, there are state of the art ultrawide-field colour fundus cameras, that can capture over 80% of retinal surface in a single image. [2]

Optical coherence tomography (OCT)

This method uses low-coherence light to provide 1D, 2D or 3D macular cross-section image. It is commonly used to analyse the thickness of the retina to diagnose DME or retinal detachments [8].

Fluorescein angiography

A technique that provides a view at the retinal vessels, enabling microaneurysm detection and neovascularisation [8].

B-scan ultrasonography

A rapid and non-invasive technique used with preretinal or vitreous haemorrhage, where the retina cannot be visually examined [23].

2.4.8 Treatment

Generalised treatment consists of hyperglycemia, hypertension (high blood pressure) and hyperlipidemias (high lipid levels) compensation [24]. When initiated early on, it can serve as long-term protection for all diabetic patients. When the disease progresses, there are several options available:

Pharmacological approach

Currently focuses on the development of drugs that prevent or delay the appearance of diabetic retinopathy. Many of the approaches provide intensive control of glycemia levels in patients with a risk of DR development. [25] Modern clinical trials also show the benefit of using anti-VEGF (anti vascular endothelial growth factor) injections at treating the advanced DR stages, where vision loss is already present [2].

Laser therapy

The golden standard in managing DR. The therapy is focused on photocoagulation using a laser beam. Most commonly, an argon laser with wavelengths around 500 nm is used. A coherent laser beam is absorbed by pigment cells in the retina and cause the coagulation. Typically it is used outside of macula to eliminate neovascularisation [26].

Chirurgical therapy

The procedure called “vitrectomy“ allows for the removal of vitreous haemorrhages, laser coagulation and macular edema treatment. It is often used when the laser therapy fails or in the advanced stages of proliferate DR. [26]

2.4.9 Prevention

Optimal care can be achieved by a systemative active screening, multidisciplinary cooperation and early treatment. Currently, in the Czech Republic, asymptomatic diabetic children are screened every year and when the disease onsets, screening period shortens to 3-6 months [6]. However, in other countries like India, most general ophthalmologists lack the necessary equipment to detect diabetic retinopathy in its crucial early stages, when the disease is most sensitive to treatment [27], and there is a growing need for an automated screening [28]. Such screening reduces the grading workload of the staff [29], expands the number of people in the screening programmes,

and outperforms the specificity and sensitivity for moderate and worse stages of DR [30].

Chapter 3

Problem statement

This chapter first introduces the Kaggle competitions in general and then focuses on the related challenges. There are three related competitions, the Diabetic Retinopathy Challenge 2015 is in Section 3.1, the APTOS 2019 Blindness Detection is in Section 3.2 and the ISBI 2nd Diabetic Retinopathy Challenge in Section 3.3.

Each of these sections describes first the competition, then the dataset provided and then it provides the summary of the participants' methods.

Section 3.4 describes an extra dataset that was collected by the author of this thesis.

Kaggle competitions

Kaggle is a platform for data science competitions [31]. One of them, Diabetic Retinopathy Challenge [32] is the base of this thesis. However, there are also other related competitions.

3.1 Diabetic Retinopathy Challenge - 2015

This competition ran from 17.2.2015 to 27.7.2015. The task was to create an automated system for DR detection. Among other rules, the use of external data was not permitted. The only permitted dataset were retinal images provided by EyePACS, a free platform for retinopathy screening.

Team size was not limited. Photos have been mostly obtained with undilated pupils and provide different views at the retina. The main arguments favouring undilated pupils are the reduction of time required and greater patient acceptability [33]. There is one image per eye and two images per

Severity	Label	Number of training images	Number of testing images
No DR	0	25810 (73.5%)	39533 (73.8%)
Mild	1	2443 (7.0%)	3762 (%)
Moderate	2	5292 (15.0%)	7861 (14.7%)
Severe	3	873 (2.5%)	1214 (2.3%)
Proliferate	4	708 (2.0%)	1206 (2.3%)

Table 3.1: Distribution of the EyePACS dataset

patient (left and right eye).

In total, 660 teams have participated, and score leaderboard has been published. Public leaderboard, that evaluated performance on 20% of the test data, was available during the competition. Private leaderboard, that evaluated performance on 80% of the test data, was available only after the submission deadline. The final positions were based on the private leaderboard.

In the private leaderboard, the winner obtained QWK score of 0.849570. The total average score of all the participants was 0.159833. As a lot of the submissions reached very low QWK, a filter has been added to account only the participants with QWK above 0.10. This criterion was fulfilled by 236 teams and the average score of these teams was 0.426682 with a standard deviation of 0.211659. Histogram of the participants' scores is available in Figure 3.2.

3.1.1 Dataset

EyePACS is the largest, publicly available dataset [34]. It was composed by EyePACS, a free platform for retinopathy screening as part of the Kaggle competition. A clinician rated the severity of DR in each of the images on a scale from 0 to 4. The dataset consists of 35 126 training images (35.3GB of size) and 53 576 testing images (53.7GB of size). The images are labelled using the 5-level classification (No DR, Mild DR, Moderate DR, Severe DR and Proliferate DR). Both training and testing labels have been released. Sample images can be seen in the Figure 3.1.

Even though the dataset is highly imbalanced, the distribution of the severity classes is similar for the train and test dataset (see Table 3.1).

Fundus photographs have been taken in different hospitals using different cameras. Therefore the photos have different resolutions, from 400x289px to 5184x3456px (see Figure 3.3).

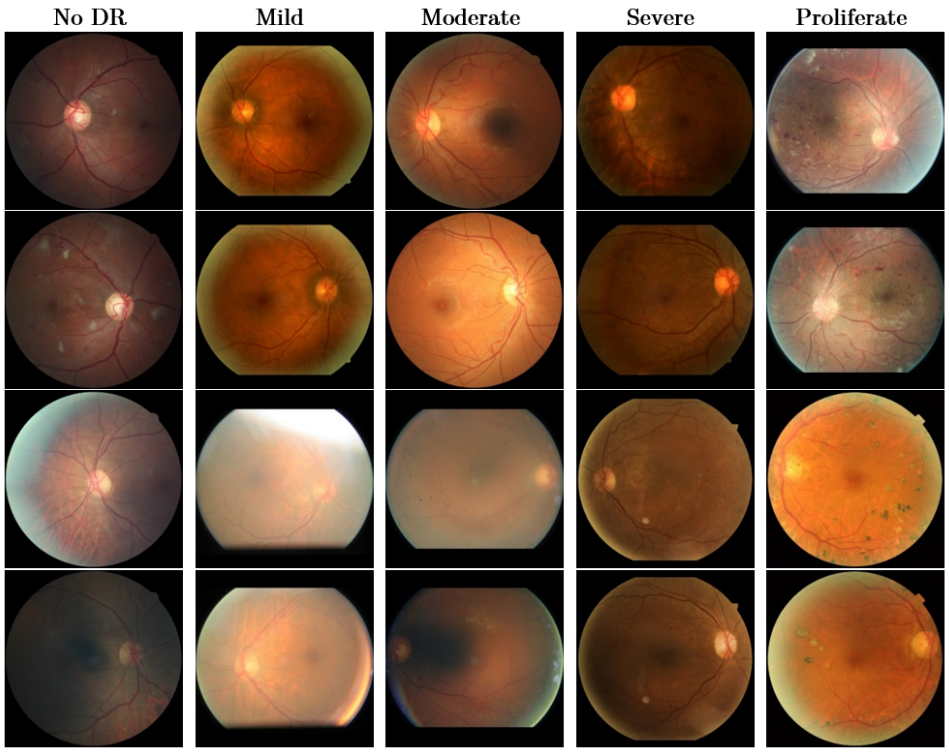


Figure 3.1: Sample images from the data. Upper two rows and bottom two rows contain data from the same patient (left and right eye).

As reported by Islam et al.[35] and Gao et al.[36], the images indeed vary in quality and contain artefacts.

To explore the relationship between the severity of two eyes, heatmap has been created (see Figure 3.4).

3.1.2 Summary

After closing of the competitions, the users were encouraged to publish their methods on the competition website [32]. The following is a summary of their posts.

1st place

Competition winners rescaled the pictures to the size of 300x300px, subtracted the local average colour and clipped the retina to 90% size to remove the boundary effect. The augmentations used were random scale, random rotation and random skew. The architecture consisted of three convolutional networks with fractional max-pooling. After that, class probabilities and metadata

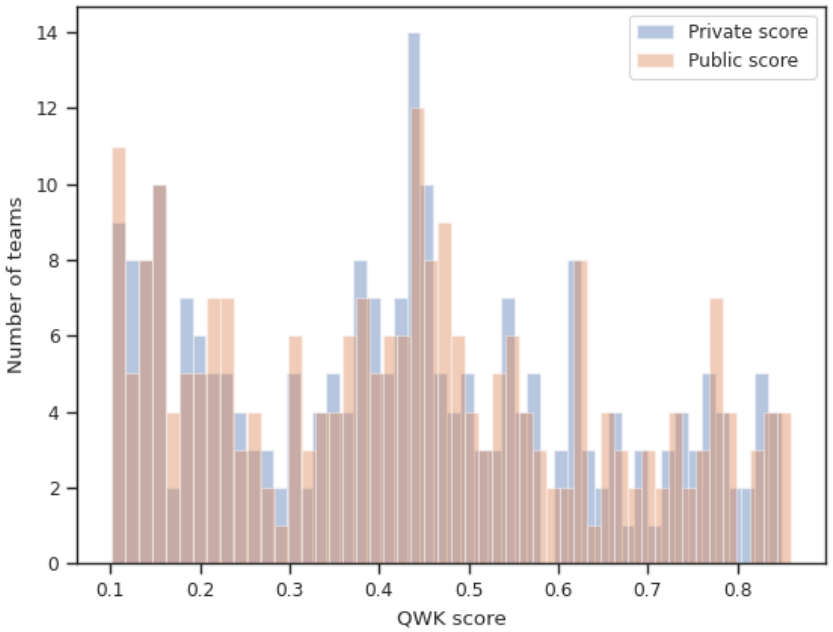


Figure 3.2: Histogram of the results of the participants of the Diabetic Retinopathy Detection 2015.

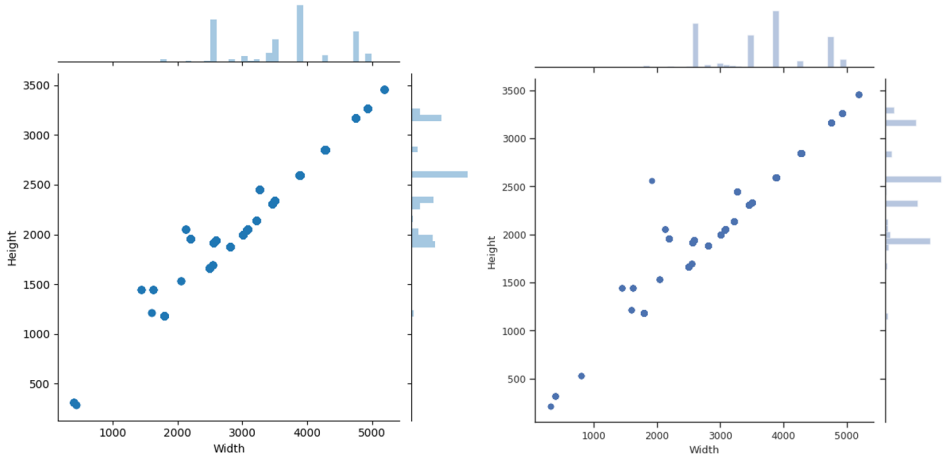


Figure 3.3: Distribution of photos resolutions in the training dataset (left) and test dataset (right) of the EyePACS dataset. Bar lines around the plot represent the count of particular width/height.

(probability of the person's other eye severity, size of the original images, a variance of the original images, the variance of the preprocessed images) were used as the input for random forest. The final prediction was obtained by averaging the probability distributions of multiple tests and thresholding.

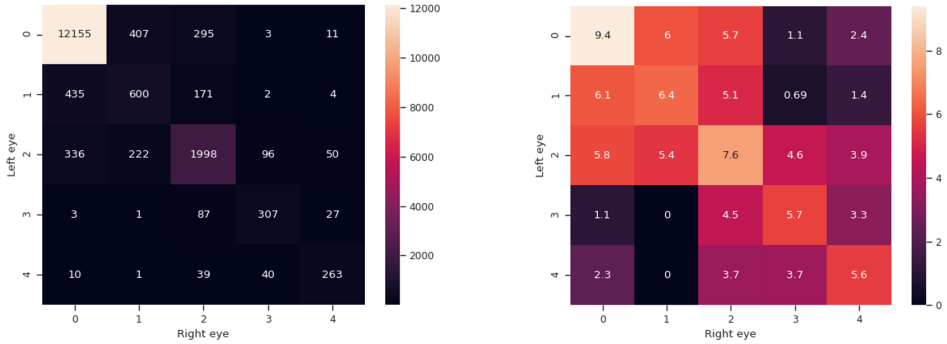


Figure 3.4: Heatmap of a patient DR severity of both eyes in linear (left) and logarithmic (right) scale.

2nd place

The team on the 2nd place resized the images to sizes 512, 256, 128px and applied rotation, translation, scaling, stretching and Krizhevsky colour augmentation. The architecture consisted of convolutional and dense layers followed by leaky rectifier units. L2 weight decay and L2 loss function were applied. Training started with resampling such that all classes were present in equal fractions, then gradually decreased the balancing after each epoch to arrive at final resampling weights of 1 for class 0 and 2 for the other classes. The optimiser was SGD with Nesterov momentum (0.9) with a fixed learning rate schedule over 250 epochs. Batch size 24 to 48 for large networks and 128 for the smaller ones was used. Per patient blend was also used – the authors extracted mean and standard deviation of RMSPool layer for 50 pseudorandom augmentations for three sets of weights (best validation score, best kappa, final weights) for net A and B; then standardised all features to have zero mean and unit variance and used them to train a small network with L1 regularisation in the first layer and L2 regularisation everywhere else. The blending network used was Adam optimiser, L2 loss and the model was trained for 100 epochs. The output values were then thresholded.

3rd place

The team on the 3rd place used images from size 384 to 1024 and applied random affine transformations, cropped, flipped, rotated and scaled the images to the desired model input size. The model used was an ensemble of 9 convolutional networks. Pre-Lu weight initialisation helped trained models with large numbers of layers. The participant reported that the key to success was the use of larger images and blending information of both eyes.

5th place

Participant on the 5th place used images of size 512 and applied a variety of transformations - cropping, colour balance adjustment, brightness, contrast, flipping, rotating and zooming. Some classes were oversampled to get a more uniform distribution of classes in batches; however, somewhere in the middle, oversampling stopped, and images were sampled from the true training set distribution. The models used SGD and Nesterov momentum for almost 100k iterations. The loss function was a combination of continuous kappa loss together with the cross-entropy (log) loss. The author ensembled a few models using the mean of their log probabilities for each class, converting these to normal probabilities in (0, 1) again and using weighted probabilities to apply the ranking procedure to assign labels. The participant proposed to split the image into four (or sixteen) non-overlapping (or only slightly overlapping) squares parallel and then combining these representations. However, this did not seem to work. In the solution, the author also tried spatial transformation layers with the intention to use some coarse input to direct the attention to some parts of the image in higher resolution, but training this total architecture end-to-end was incredibly difficult. This participant also reported, that camera artefacts seem to be fairly common. In the end, the pseudolabelling approach was adopted, which helped to push the score of a single model. The author also reported that most errors came from predicting class 0 when it really was class 2.

3.2 APTOS 2019 Blindness Detection

This competition ran from 27. 6. 2019 to 5. 9. 2019 and the task was the same as with the 2015 competition. The rule was to compete in a maximum team size of 5. The competition was sponsored by Aravind Eye Hospital & PG Institute of Ophthalmology.

In total, 2929 teams have participated, and score leaderboard has been created. Public leaderboard, that evaluated performance on 20% of the test data, was available during the competition. Private leaderboard, that evaluated performance on 80% of the test data, was available only after the submission deadline. The final positions were based on the private leaderboard.

In the private leaderboard, the winner obtained QWK score of 0.936129. The total average score of all the participants was 0.787760. As some of the submissions reached very low QWK, a filter has been added to account only the participants with QWK above 0.10. This criterion was fulfilled by 2715 teams and the average score of these teams was 0.849887 with a standard deviation of 0.126608. Many participants reported, that the results from the public leaderboard were very different to the results from the private

Severity	Label	Number of training images
No DR	0	1805 (49.3%)
Mild	1	370 (10.1%)
Moderate	2	999 (27.3%)
Severe	3	193 (5.3%)
Proliferate	4	295 (8.1%)

Table 3.2: Distribution of the APTOS dataset

leaderboard. Histogram is provided in Figure 3.5.

3.2.1 Dataset

Aravind Eye Hospital in India released a dataset as a part of the APTOS 2019 Blindness Detection competition [37]. It composes of 3662 training images (12.0GB) and 1928 testing (2.4GB) images. Only the training labels have been released. Multiple attempts have been made to obtain the test labels but with no success. Neither the organisers nor the top participants were willing to share the labels.

The imbalance of this dataset is not as significant as in the case with the EyePACS (see Table 3.2), the resolution of the images is similar. There is one image per eye, two images per patient (left and right eye).

3.2.2 Summary

The participants were also encouraged to publish their methods on the website [38]. The following is a summary of their methods.

1st place

Winning team place applied no preprocessing, just resizing. Augmentations consisted only of horizontal flip to reduce running time. The class imbalance was not treated in any way. The model used was an ensemble of 8 models (2x InceptionResnetV2 image size 512px, 2x InceptionV4 image size 512, 2x Seresnext50 image size 512, 2x SeresNeXt101 image size 384). For the last pooling layer, the author found the generalised mean pooling (used the default p=3 as initial value) better than the original average pooling. The first stage consisted of training on the training dataset, while in the second stage, models were trained using the pseudo-labelled (soft) test dataset. Loss used was SmoothL1 loss optimised with the Adam optimiser. The outputs of

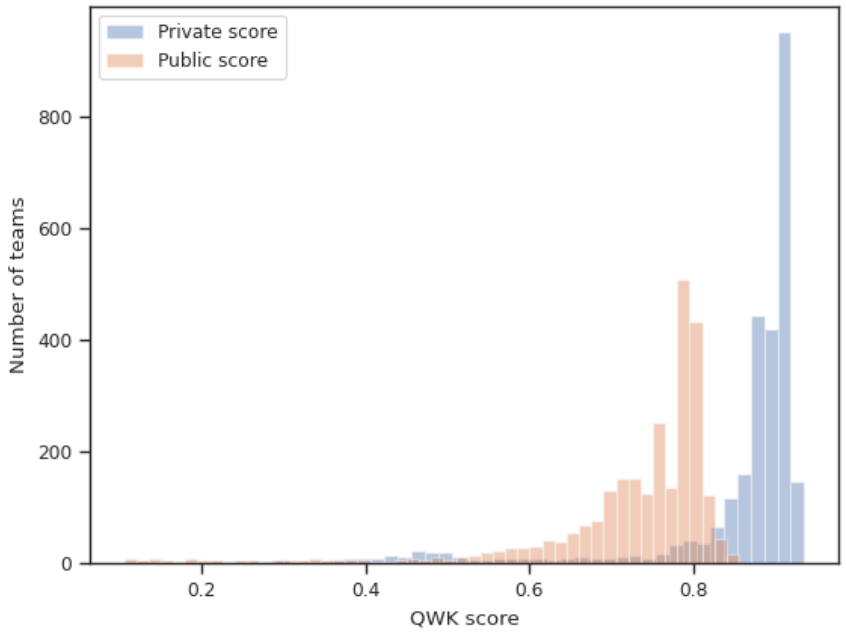


Figure 3.5: Histogram of the results of the participants of the APTOS 2019 Blindness Detection.

the ensemble were averaged and thresholded with optimised levels (0.7, 1.5, 2.5, 3.5) to get the final score.

2nd place

Competitor scoring on the 2nd place used image sizes 300, 460 and 456px and applied a variety of transformations from the Albumentations library [39] – blur, flip, random brightness and contrast, shift, scale, rotate, elastic transform, transpose, grid distortion, hue saturation value, CLAHE and coarse dropout. Model ensembled multiple Efficient-nets (B3 image size 300, B4 image size 460, B5 image size 456). The blending was done by a simple mean. The author also adopted pseudo labelling approach of the test data.

4th place

Fourth place was won by a researcher, that cropped the images to sizes 224, 240, 256, 320, 352 and 376px and applied the following transformations: dihedral, random crop, rotation, contrast, brightness, cutout, perspective transform and CLAHE. The author reported that large models on high-resolution were much more likely to be overfitting. He, therefore, used small

Severity	Label	Number of training images	Number of validation images
No DR	0	532 (44.3%)	174 (43.5%)
Mild	1	139 (11.6%)	46 (11.5%)
Moderate	2	232 (19.3%)	92 (23.0%)
Severe	3	(17.8%)	68 (17.0%)
Proliferate	4	72 (6.0%)	20 (5.0%)

Table 3.3: Distribution of the ISBI dataset

model for high resolution and large model for low resolution, namely Efficientnets: B7 (image size 224), B6 (image size 240), B5 (image size 256), B4 (image size 320), B3 (image size 352), B2 (image size 376). The final submission was trained on the full 2015 dataset for 25 epochs.

3.3 ISBI - The 2nd Diabetic Retinopathy – Grading and Image Quality Estimation Challenge - 2020

This competition ran from 25. 10. 2019 to 25. 3. 2020. It contained three sub-challenges, from which the Sub challenge 1 was selected as the task was similar to the 2015 and 2019 competitions. The difference was that the retinal fundus images were wide-range, and two photos per eye were provided, each from a different angle.

There have been 26 participants with a top QWK of 0.930.

3.3.1 Dataset

This dataset was gathered between 2014 and 2017 in different locations across China and was released as part of the ISBI - The 2nd Diabetic Retinopathy – Grading and Image Quality Estimation Challenge [40]. The authors have extracted 2000 regular fundus images from 500 patients (2 images per eye, 4 images per patient - one with the optic disc at the centre and another one with the fovea at the centre).

The images have been divided as follows:

The attempt has been made to obtain the test labels, but the organisers argued the test photos might be used in another competition in the near future and therefore cannot be given to the public.

The images have a resolution of 1956x1934 pixels and are stored as jpg images.

3.3.2 Summary

After the competition deadline, an online session was held to allow participants to present their methods publicly. The following is a summary of that session.

1st place

The team from Shangai Jiao Tong University (China) used image sizes from 288px to 674px (mostly 384), no special preprocessing was done. The transformation list comprised of the only flip, rotate and brightness adjustments. The team adopted a regression approach on EfficientsNets B5, B3 and B1 with GeM Pooling. Models were pretrained on ImageNet and Kaggle DR dataset, finetune on DeepDR, only use feature extractor and reset FC layer when finetune. The training was done with SmoothL1 loss AdaMod optimiser and Cosine LR scheduling with warm restarts. The evaluation used 5-fold Cross-Validation. As the dataset consisted of two photos per eye, the blending of the eyes was adopted by averaging the score of two images and applying thresholds (0.55, 1.55, 2.4, 3.18).

3rd place

The team from VUNO company, South Korea used images with input size 1024x1024px and assembled all the available datasets. When labels were not present in the 5 categories, they used pseudo labels. The images were normalised into (0,1) range and underwent a battery of transformations: random flip, affine transformation (rescale, shear, translation, rotation), image perturbation (colour contrast, brightness, sharpness, RGB shift, Gamma), noise (blur, ISO noise, Gaussian Noise, JPEG compression and sun flare), elastic transformation, grid transformation and downsampling. The architecture used was EfficientNet B4 with optimised with SGD with Nesterov momentum and fixed learning rate. The loss function consisted of L1 loss. The final model was an ensemble of 10 models with optimised thresholds and max(img1, img2). Source code is available publicly [41].

5th place

5th place was achieved by a team from Beihang University (China). They used input sizes from 512px to 800px, removed the black side of the image

Severity	Label	Number of images
No DR	0	0 (0.0%)
Mild	1	1 (20.8%)
Moderate	2	21 (39.6%)
Severe	3	0 (0.0%)
Proliferate	4	21 (39.6%)

Table 3.4: Distribution of the VFN dataset

and randomly applied from three to twelve of the following transformations: random contrast, histogram equalisation, invert, rotate, posterise, solarise, SolarizeAdd, random colour, random contrast, random brightness, sharpness shear, ShearY, CutOutAbs, TranslateXabs, TranslateYAbs. The architecture used was EfficientNetB7 with a Generalised Mean pooling layer. The loss function was a cross-entropy loss, optimiser SGD with warmup. The team adopted a blending network for the two images of the same eye using a FC layer with a dropout (0.5).

7th place

7th place won the joint team of the University of Bournemouth and ETS Canada. They adopted a classification approach. Architectures used were Resnet50 and Resnetx50. Loss function was a combination of a base loss (did not specify) and a cost-sensitive cross-entropy loss. The training started on the Kaggle EyePacs dataset while heavy oversampling the minority classes and then finished on the 2020 dataset.

3.4 VFN

In cooperation with the Eye Clinic at General University Hospital in Prague (VFN), namely prof. Heissigerova (Head of Clinic) and Kovacova MD, an anonymised dataset was obtained for the purposes of this thesis.

It contains 53 images of the retina (see Table 3.4). This is the only dataset that has images obtained with dilated pupils. The images have been photographed using either standard 20-50 degree-view ophthalmoscope or a CLARUS ultra-widefield retinal camera and the images are not paired to a patient.

The No DR class lacks images as the laboratory is a nationwide centre for Diabetic Retinopathy and concentrates patients already diagnosed elsewhere. The Severe class lacks because the methodology adopted by VFN requires

fluorescein angiography and patient medical history to confirm this stage of the disease.

3.5 Other

There are also other DR related datasets, such as Messidor, IDRiD or ODIR. However, many of these have different DR severity levels (binary, 4-level), are unlabeled or appear in small portion and therefore were not put into consideration.

Chapter 4

Related work

This chapter summarizes the past attempts to solve the automatic detection problem. Section 4.1 describes the progress made in the computer vision research, listing multiple authors chronologically. Where available, the results and techniques of the authors were described. After that, Section 4.2 reports the commercially available systems for the detection of diabetic retinopathy.

4.1 Computer vision research

One of the first to propose the neural networks for the DR detection was Gardner et al. (1996) [42], who used neural networks to classify small patches of the original image. An ophthalmologist was then required to classify the patches for features, and the output was fed into the SVM.

In 2000, Ege et al. [43] proposed the use of Bayesian classifiers and k-nearest neighbours for DR detection.

Acharya et al. (2008) [44] and Adarsh et al. (2013) [45] adopted five-class classification using SVM and calculated the size of areas of different DR signs such as haemorrhages, micro-aneurysms, exudate and blood vessel. Roychowdhury et al. (2014) [46] used a two-level model with a Gaussian mixture model (GMM), kNN, and support vector machine (SVM). However, these types of approaches had the disadvantage of utilising a limited number of features.

After the major success of Alex Krizhevsky et al., who won the ImageNet Large Scale Visual Recognition Challenge in 2012 [47], the attention has shifted towards the Convolutional Neural Networks (CNNs).

Prat et al. (2016) [48] used 5,000 validation images, resized them to 512x512px, applied colour normalisation and several transformations (random rotation 0-90 degrees, random horizontal and vertical flips and random

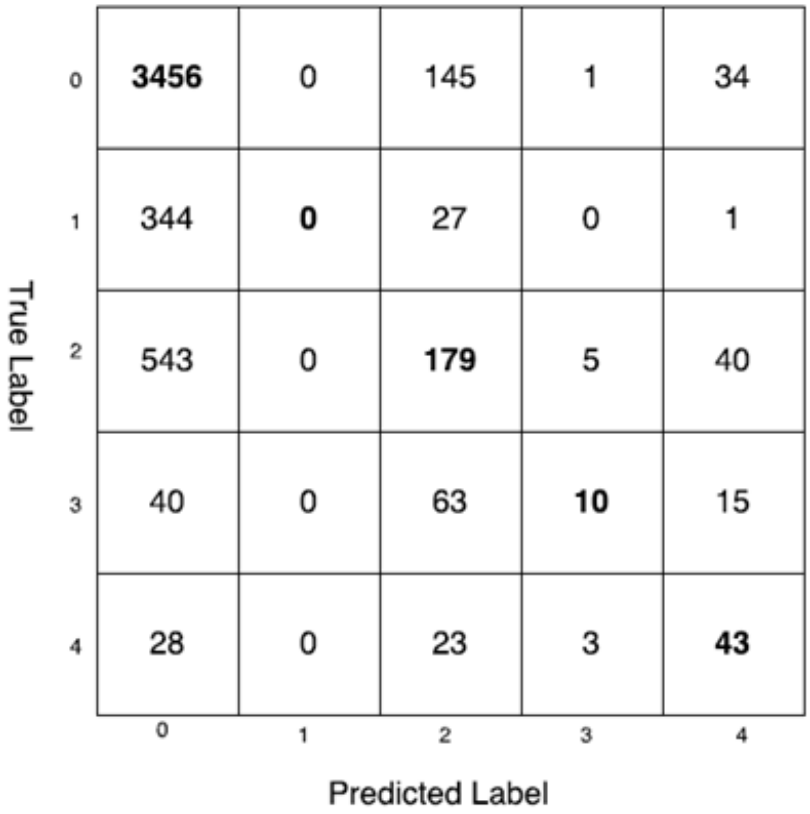


Figure 4.1: Confusion marix summarizing score of Prat et al., source [48].

horizontal and vertical shifts). They used a convolutional neural network. The network was initially pretrained on a small piece of the dataset and then finetuned it on the full dataset — used Stochastic gradient descent with Nesterov momentum. The results achieved are summarized in the Figure 4.1. They reported low sensitivity, which they accounted for model failing to detect particular features from the mild and moderate classes. Also identified the issue of images ungradability, as more than 10% of the data did not comply with the UK standards.

Islam et al. (2018) [35] used image sizes 512x512px and 448 x 448px. They rescaled the images, mapped the local average colour to 50%, clipped the images to 90% size to remove the boundary effects and applied various transformations (rotation, shearing, flip, zoom, crop). They also normalised each channel to have zero mean and unit variance over the dataset. Proposed architecture consisted of a base network consisting of 8.9 million parameters, optimised by SGD Nesterov momentum. L2 regularisation was applied. The batch size was 16. Islam et al. also took the features from the last max-pooling layer as input features for a blending network. This network blended the features for both eyes leading to a significant improvement. The network used

Figure 4.2: Various quality of images from the EyePACS dataset, source [35].

a regression approach, and the outputs were thresholded (0.5; 1.5; 2.5; 3.5). The dataset used was EyePACS. Authors mention „The dataset also contains artefacts, out of focus, too bright, and too dark images.“ (see Figure 4.2. Their methodology was to perform two binary classification for early-stage detection experiment: sick (grades 1, 2, 3, 4) vs healthy (grade 0), and low (grades 0, 1) vs high (grades 2, 3, 4). They obtained better results using the low-high DR classification approach.

Gulshan et al. (2018) [49] used Binary classification (detection of moderate or worse stages of DR). The evaluation was performed on a local dataset from India and the reported performance was equal to or exceeded manual grading.

Voets et al. (2018) [50] tried to replicate the main method from the article *Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs*, published in JAMA [51] and used Messidor-2 [52] to evaluate. Unfortunately, he not come close to the reported AUC of 0.99. However, their experiments show that training by normalising the images to a [-1, 1] range gave the best results for this replication. The input size was 299x299px the architecture used was InceptionV3 with a RMSProp optimizer [53]. They used ensemble learning by training 10 networks on the same data set and using the mean of the predictions of the ensemble to compute the final grade. They found out that 19.9% of the EyePacs dataset images are ungradable, and assumed possible, that the algorithm could learn from features for ungradable images. On the other hand, they also tried training by excluding non-gradable images, but the final performance has not increased,

Gao et al. (2019) [36] reported that the available datasets differ significantly in their annotation and quality. His approach classified the images into 4 levels of severity and used data from the local hospitals in Sichuan Province (China).

Li et al. (2019) [54] proposed to replace the max-pooling layers with fractional max-pooling in order to derive more discriminative features for classification. The output vector was combined with the image's metadata and used as input for a support vector machine (SVM) classifier. The primary dataset was EyePACS. Li et al. classified the images into into five categories. Furthermore, they designed a handy app called „Deep Retina“, which could automatically také a picture thru the ophthalmoscope and evaluate the DR

severity.

Sarki et al. (2019) [55] used the data from the Messidor and Kaggle; the transformations used included cropping, resizing, rotating and mirroring. Sarki used in total 13 Convolutional Neural Networks architectures, pre-trained on large-scale ImageNet, with a mini-batch size of 32. Binary classification (no DR/mild DR) was performed with ResNet50 and RMS prop optimiser.

Sahlsten et al. (2019) [56] proposed a novel approach, adopting only a small fraction of images, but in much higher resolutions. They used data from a Finnish provider DigiFundus. For the highest resolution (2095x2095 pixels), the architecture used was Inception-v3 with mini-batches of size 15. The learning took 40 days of consecutive model training. Their results confirm the good performance of high-resolution classification, on the other hand, the high resolution only did binary classification, and the time consumed was enormous.

Gonzales et al. (2020) [57] also used a custom made dataset called DR-AMD, which was collected in three different European medical centres (Sweden, Denmark, Spain). Their final approach consisted ensemble of two state-of-the-art CNN architectures.

Hemath et al. (2020) [58] used the Messidor dataset [59]. Proposed the use of employing both HE and CLAHE [60] techniques within image processing for each channel. The image size was reduced to 150 x 225 pixels. The model had eight layers.

Tymchenko et al. (2020) [61] found out that labelling schemes are inconsistent between datasets, and therefore decided to use the largest dataset to pretrain the CNNs. The augmentations of the images included horizontal flip, vertical flip, transpose, rotate and zoom). The method proposed use of several approaches on how to deal with DR detection – as a classification problem, as a regression problem, and as an ordinal regression problem Main training consisted of the use of Focal Loss [62] for classification head, binary Focal Loss [62] for ordinal regression head and mean-squared error for regression head. The optimiser was Rectified Adam optimiser [**?**], with cosine annealing learning rate schedule [63], weight decay [64] and dropout. Also, they penalised overconfident predictions by using label smoothing [65] The final prediction was obtained by ensambling 3 encoder architectures working at different resolutions: EfficientNet-B4 (380x380px), EfficientNet-B5 (456x456px) [66], SE-ResNeXt50 (380x380px and 512x512px) [67].

They scored 54 of 2943 competing methods (QWK score of 0.925) in the APTOS 2019 Challenge.

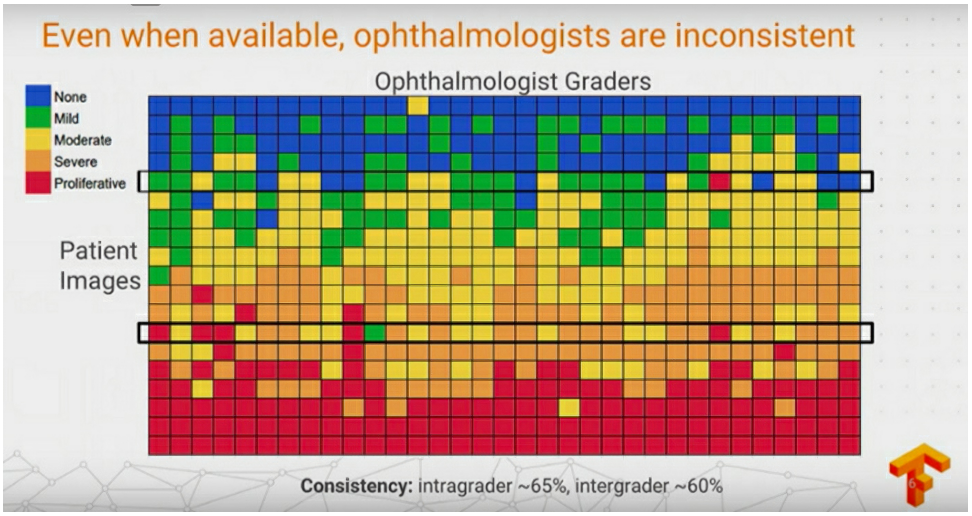


Figure 4.3: Google project presentation showing graders inconsistency, source [**?**].

4.2 Commercial systems

First commercial systems have begun to emerge in recent years. Shan et al. [68] evaluated the performance of IDx-DR, an American system, developed in collaboration with the IBM Watson, that primarily detects stages worse than mild.

Google Brain AI research team approached the problem by the following strategy: Images of size 779x779 were used as an input for the InceptionV4 model that rated them into 5-class DR severity ratings. The dataset was consisting of more than 1.6 million images. Models were then finetuned on a smaller dataset (2000 images labelled by 3 retina specialists) [69]. They also trained a secondary model (InceptionV4 [70]) to evaluate the image gradability. As part of their presentation, Google also pointed at both intragrader and intergrader inconsistency (see Figure 4.3).

This finding is also backed by Gulshan et al. [71], who evaluated the scores of different raters (see Figure 4.4).

Table 3. Quadratic Weighted κ Scores for 5-Point Diabetic Retinopathy Grading[a]

Hospital	Quadratic Weighted κ (95% CI)
Aravind	
Retina specialist (C.O.[b])	0.74 (0.71-0.76)
Trained grader (L.V.[b])	0.75 (0.72-0.79)
New model	0.85 (0.83-0.87)
Sankara	
Retina specialist (R.R.[c])	0.82 (0.80-0.84)
Trained grader (S.S.[b])	0.88 (0.86-0.89)
New model	0.91 (0.90-0.93)

[a] Only the κ score from the new model is reported as the original model was not trained to predict 5 class grades.

[b] Nonauthor technician.

[c] Dr Raman.

Figure 4.4: Grader inconsistency, source [71].

Chapter 5

Theoretical background

This chapter provides the theory for understanding the concepts and techniques used in the solution proposed by this thesis.

Section 5.1 describes the metrics used for evaluation of the performance, namely the confusion matrix, recall, precision, F-1 score, quadratic weighted kappa score and cross-validation.

Section 5.2 describes several state-of-art architectures, including the InceptionV3, DenseNet, the EfficientNet family, InceptionResnetV2 and SEResNet50. It also describes the related technique called Generalised mean poling.

Section 5.3 explains the different loss functions for classification (Cross-entropy loss, Focal loss and an advanced loss, QWK loss) and regression (L1 loss, SmoothL1 loss and L2 loss). It also outlines the related techniques, such as Mixup and cost sensitive regularisation.

Section 5.4 lists a few optimisers (SGD + momentum, Adam, AdamW) with a short description, Section 5.5 then describes the learning rate schedulers (Reduce on plateau and Cosine Annealing).

Further sections focus on the techniques used as improvements - Section 5.6 discusses the elements related to the segmentation, namely the architecture and the loss function; Section 5.7 describes the basics of random forest finally, Section 5.9 introduces a method for model explainability.

5.1 Metrics

5.1.1 Confusion matrix

Useful and simple metric is the confusion matrix. This tool provides a summary of the model predictions against the ground-truth labels. In case of

multiclass classification with N classes available, the matrix is of shape NxN, in case of binary classification, the matrix is of shape 2x2, and the following terms can be defined [72]:

- positive – indicates a prediction of 1
- negative – indicates a prediction of 0
- true – correct prediction
- false – incorrect prediction

- True Negative (TN) — model predicts negative outcome & groundtruth label is negative.
- True Positive (TP) — model predicts positive outcome & groundtruth label is positive.
- False Negative (FN) — model predicts negative outcome & groundtruth label is positive.
- False Positive (FP) — model predicts positive outcome & groundtruth label is negative.

5.1.2 Recall

Recall indicates the proportion of the true positive cases, that were correctly classified.

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{5.1}$$

5.1.3 Precision

Precision indicates the proportion of true positive predictions in all the positive predictions by the network.

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{5.2}$$

5.1.4 F1 score

F1 score is a harmonic mean of recall and precision.

$$\text{F1} = \frac{2 \cdot \text{Recall} \cdot \text{Precision}}{\text{Recall} + \text{Precision}} \tag{5.3}$$

5.1.5 Quadratic weighted kappa

All the related competitions use the Quadratic weighted kappa (or Cohen's kappa) coefficient for evaluation of the performance. Vanbelle et al. [73] provides an insight into the development of this coefficient: Cohen's kappa coefficient was first introduced in 1960 [74] to measure agreement on nominal scales. It was evaluated by Kraemer et al. [75] who proved its reliability. To emphasise more dissimilar agreement, weighted kappa coefficient has been introduced. Various weights have been tested, but the most popular became the quadratic weighted kappa introduced by [76].

The coefficient itself is calculated with the following steps:

1. Confusion matrix (denoted as Observed matrix O) is created from the predictions and groundtruths.
2. Histogram of all the predictions is created and Histogram of all the groundtruths is created. Outer product of the histograms is computed, creating Expected Value matrix E.
3. Both Observed matrix and Expected Value matrix are normalised
4. Weight matrix W with shape kxk (where k is the number of classes) is calculated using the equation:

$$w_{ij} = 1 - \frac{(i-j)^2}{(k-1)^2} \tag{5.4}$$

5. Kappa coefficient is computed using the formula:

$$\kappa = 1 - \frac{\sum_{i=1}^{k}\sum_{j=1}^{k} w_{ij}o_{ij}}{\sum_{i=1}^{k}\sum_{j=1}^{k} w_{ij}e_{ij}} \tag{5.5}$$

   where k is the number of classes, o_{ij}, and e_{ij} are elements in the observed, and expected matrices respectively; w_{ij} is calculated as follows:

The range of values of quadratic weighted kappa (QWK) coefficient is from -1 (complete disagreement) to 1 (complete agreement) and can be interpreted using the scale developer by Ashby et al. (see Table 5.1).

5.1.6 Cross-validation

The idea of cross-validation is to split the training data into K folds. For each fold $k \in (1, \ldots, K)$, model is then trained on all the folds but the k-th, and evaluated on the k-th. After evaluating on all K folds, the error is averaged over all the folds [78]. Illustration can be seen on the Figure 5.1.

Value of QWK	Strength of agreement
< 0.20	Poor
0.21 - 0.40	Fair
0.41 - 0.60	Moderate
0.61 - 0.80	Good
0.81 - 1.00	Very good

Table 5.1: Ashby's interpretation of QWK, source [77].

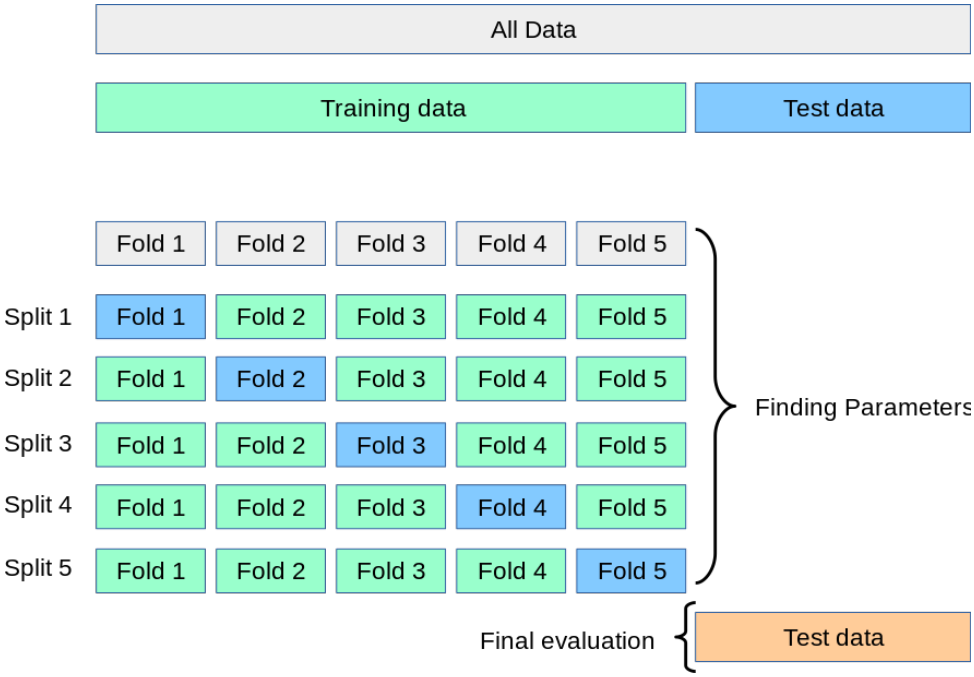


Figure 5.1: Summary of 5-fold cross-validation, source [79].

5.2 Neural networks

As with other supervised mathematical models, the task of neural networks is to find weights for the network (approximation function) that minimise the loss function, given a set of labelled training data.

Following the current trend in computer vision, this thesis is built with the use of Convolutional Neural Networks (CNNs). This type of neural networks has shown great success in image processing as it is invariant to translation [80]. The convolution is defined as

$$g(x,y) = \omega * f(x,y) = \sum_{dx=-a}^{a} \sum_{dy=-b}^{b} \omega(dx,dy) f(x+dx, y+dy) \tag{5.6}$$

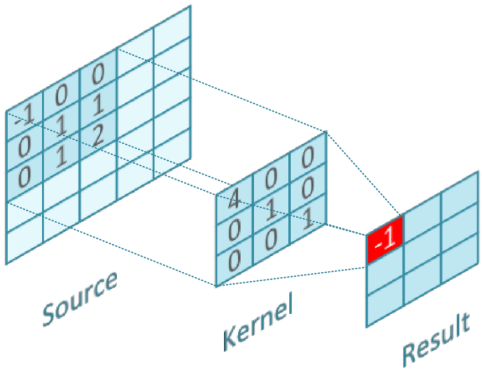


Figure 5.2: Simplified schema of convolution. The original image is denoted as Source; the filtered image is denoted as result., source [82].

where $g(x, y)$ is the filtered image, $f(x, y)$ is the original image, ω is the filter kernel [81]. Illustration is provided in Figure 5.2.

5.2.1 InceptionV3

InceptionV3 is a 42-layer convolutional neural network architecture introduced in 2015 [65]. The main idea behind it lies in factorising convolutions. Those mean a) factorising into smaller convolutions by replacing one bigger, b) factorising into asymmetric convolutions where two asymmetric convolutions replace one bigger. The convolutions are assembled into so-called „modules“ and several times repeated. At the end of the network there is a fully connected layer (FCN) and softmax layer [80]. The network also provides an auxiliary classifier with its own softmax output layer. Scheme is provided in Figure 5.3.

5.2.2 DenseNet

For the second architecture, DenseNet [83] was selected. The architecture was proposed as the next step towards more deep convolutional networks. It can be understood as a ResNet-related architecture family because both try to combine features. However, where ResNets combines features through summation before they are passed into a layer, DenseNet instead, combine features by concatenating them. Hence, each dense layer has inputs consisting of the feature-maps of all preceding convolutional blocks, and the layer also passes its outputs to all subsequent dense layers [83]. Scheme is available in Figure 5.4.

5.2.3 EfficientNet family

Proposed by Tan et al.[66], EfficientNet introduces „a new scaling method that uniformly scales all dimensions of depth/width/resolution using a simple yet highly effective compound coefficient“. The authors have developer a baseline network and scaled it up to obtain a family of models, called EfficientNets.

They have demonstrated that „EfficientNet model can be scaled up very effectively, surpassing state-of-the-art accuracy with an order of magnitude fewer parameters and FLOPS, on both ImageNet and five commonly used transfer learning datasets.“ [66]. Scheme is available in Figure 5.5.

5.2.4 InceptionResnetV2

This is a 164 layers deep neural network proposed by Szegedy et al. [70]. It is „a costlier hybrid Inception version with significantly improved recognition performance.“ [70]. The main idea behind InceptionResnetV2 lies in the Residual Inception Blocks. Scheme can be seen in Figure 5.6.

5.2.5 SEResNet50

Introduced by Hu et al. [67], SeResNet50 is a type of hybrid network utilising the Squeeze-and-Excitation blocks (SE blocks) (see Figure 5.7). These blockswhich enable dynamic channel-wise feature recalibration [67]. It is built upon the ResNet architecture [84] and uses the SE blocks along with the ResNet backbone.

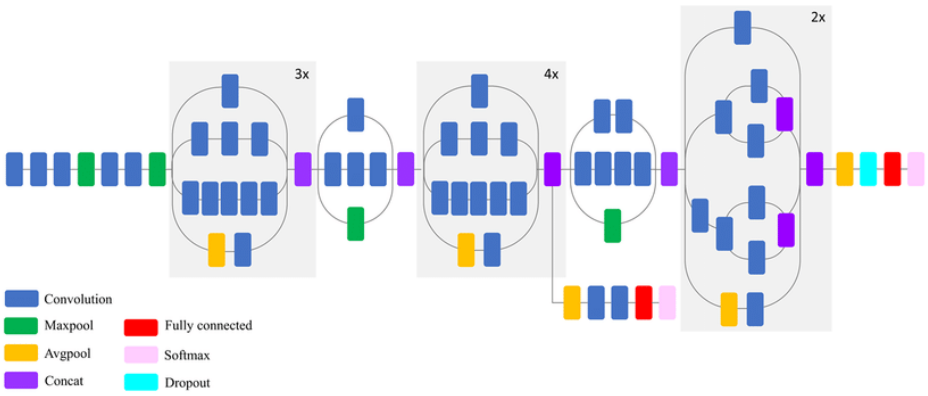


Figure 5.3: Scheme of the InceptionV3 architecture, source [85].

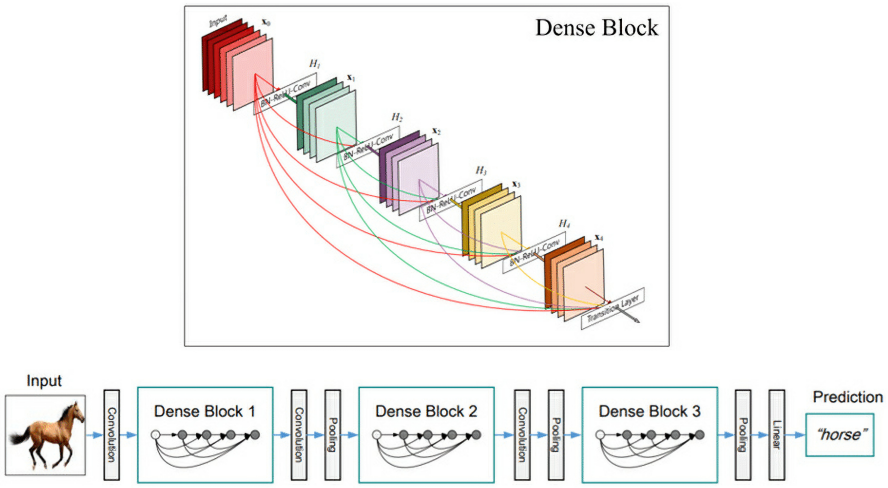


Figure 5.4: Scheme of the DenseNet architecture, source [86].

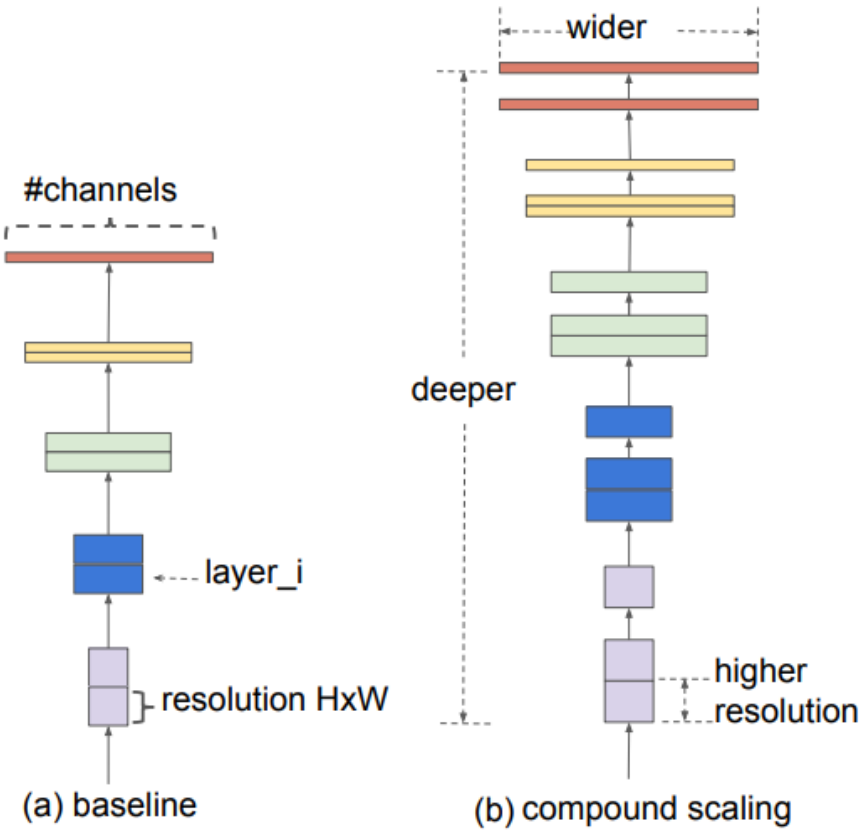


Figure 5.5: Scheme of the generic (a) and EfficientNet (b) architecturec, source [66] (modified by author)

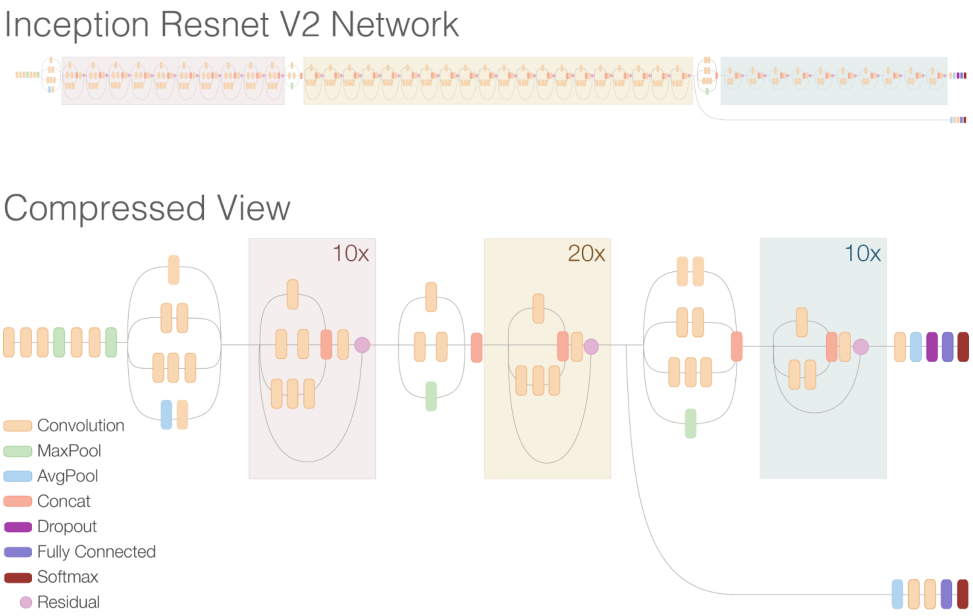


Figure 5.6: Scheme of the InceptionResnetV2 architecture, source [87].

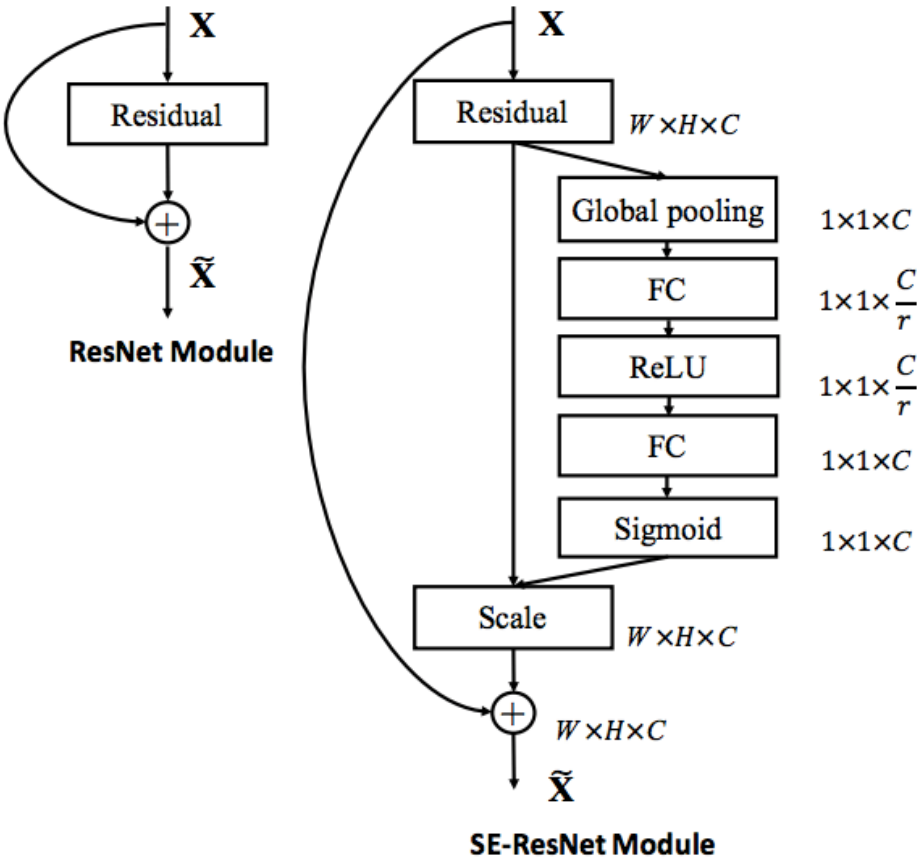


Figure 5.7: Scheme of the SE residual blocks, source [88].

5.2.6 Generalised mean pooling

Several participants of the relevant challenges, including the winner of APTOS 2019 challenge, assessed, that replacing the original average pooling layers in CNNs with generalised mean pooling [89] might positively impact the performance of the model.

Generalised mean pooling layer is based on a generalised-mean that has learnable parameters p_k, either one global or one per output dimension. The formula is as follows:

$$\mathbf{f}^{(g)} = [f_1^{(g)} \dots f_k^{(g)} \dots f_K^{(g)}]^\top, f_k^{(g)} = (\frac{1}{|X_k|} \sum_{x \in X_k} x^{p_k})^{\frac{1}{p_k}} \tag{5.7}$$

where X is the input and vector $\mathbf{f}$ is the output produce of the pooling process.

Using this formula, max-pooling can be obtained when $p_k \to \infty$ and average pooling when $p_k = 1$.

5.3 Loss function

Training the CNN is accomplished by minimising the loss function.

5.3.1 Classification

Cross-entropy loss

One way to measure the similarity of two probability distributions p, q is to use the Kullback-Leibler (KL) divergence (or relative entropy), which is defined as:

$$\mathbb{KL}(p||q) \triangleq \sum_{k=1}^{K} p_k \log \frac{p_k}{q_k} \tag{5.8}$$

Replacing the sum by an integral for probability density functions (PDFs), the formula can be rewritten as:

$$\mathbb{KL}(p||q) = \sum_{k=1}^{K} p_k \log p_k - \sum_{k=1}^{K} p_k \log q_k = -\mathbb{H}(p) + \mathbb{H}(p, q) \tag{5.9}$$

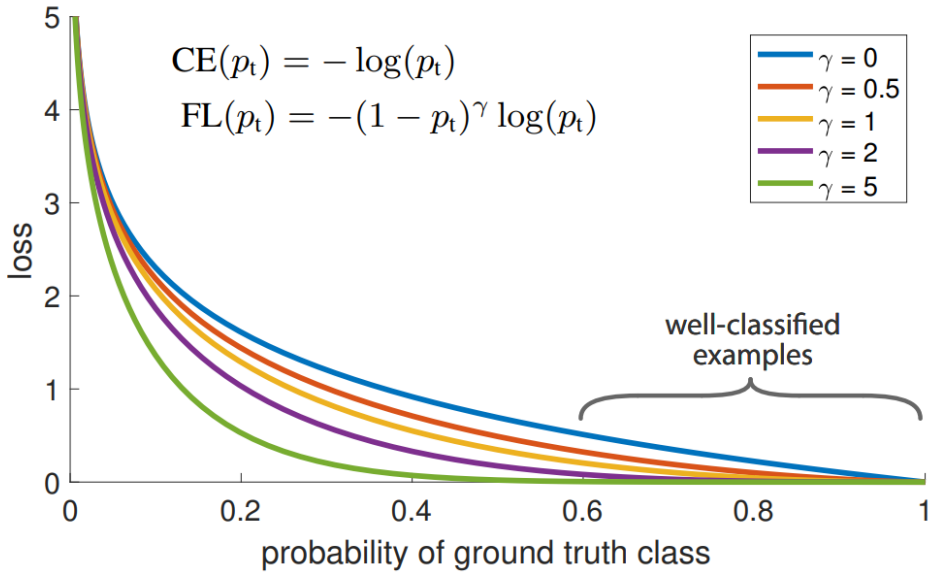


Figure 5.8: Graph of focal loss values with different value of γ, source[91].

where H(p,q) is called the cross-entropy.

$$\mathbb{H}(p, q) \triangleq -\sum_{k=1}^{K} p_k \log q_k \tag{5.10}$$

Cross entropy can also be understood as the average number of bits needed to encode data from a distribution p with model q as the definition. [90] [78]

Focal loss

The alternative way to measure the similarity of predictions is to use Focal loss [91]. It adds a factor $(1 - p_t)^\gamma$ to the standard cross-entropy criterion. Lin discovered that this loss is beneficial in improving the accuracy in cases where an extreme foreground-background class imbalance is present [91]. The plot of Focal loss with different parameter choice is in the Figure 5.8.

Advanced loss

Arazo et al.[92] and Tanaka et al.[93] have proposed the use of a robust loss function with regularisation to overcome problems arising from overfitting to noisy labels.

Their approach adopts categorical cross-entropy

$$l^\star(\theta) = -\sum_{i=1}^{N} y_i \log\left(h_\theta(x_i)\right) \tag{5.11}$$

where $h_\theta(x)$ are the probabilities predicted by the model.

Two regularisation terms are added to improve convergence. First term discourages the classifying of all the samples into one class (which is especially common in the early stages of training). It does that using the following term:

$$R_A = \sum_{c=1}^{C} p_c \log\left(\frac{p_c}{\overline{h}_c}\right) \tag{5.12}$$

where p_c is the prior probability distribution for class c and $\overline{h}_c$ denotes the mean softmax probability of the model for class c across all samples in the dataset. $\overline{h}_c$ is approximated by using mini-batches.

The second term tries to prevent the state of weak guidance and forces model to concentrate the prediction on one class using the following term:

$$R_H = -\frac{1}{N} \sum_{i=1}^{N} \sum_{c=1}^{C} h_c^c(x_i) \log(h_\theta^c(x_i)) \tag{5.13}$$

where $h_\theta^c(x_i)$ denotes the c class value of the softmax output $h_\theta(x_i)$

The complete formula for the semi-supervised loss is then

$$l = l^\star + \lambda_A R_A + \lambda_H R_H \tag{5.14}$$

where λ_A and λ_H control the contribution of each regularisation term [92].

Mixup

Mixup is a technique where the model is trained on convex combinations of sample pairs and the corresponding labels.

$$\begin{aligned} x &= \delta x_p + (1-\delta) x_q \\ y &= \delta y_p + (1-\delta) y_q \end{aligned} \tag{5.15}$$

where δ is randomly sampled from a beta distribution. In theory, mixup should lead to reduced prediction confidence and improved model calibration.

Cost sensitive regularisation

To mimic the ordinary classification, a semi-supervised loss can be extended with a cost-sensitive term. The formula is defined as:

$$L_{CS}(\theta) = \sum_{(x_i, y_i) \in D} \sum_{y_j \neq y_i} C(y_i, y_j; x_i) \log\left(P(y_j | x_i; \theta)\right) \tag{5.16}$$

where $C(y_i, y_j; x_i)$ is a positive cost of mislabeling an instance x with real label y_i into label y_j.

QWK loss

The most optimal approach would be to adopt the loss directly optimising the evaluation metric QWK. On the other hand, none of the state of art methods have adopted this method as the QWK function itself is non-differentiable. There have been attempts to approximate function and create its soft version, but the results show that the Hessian is computationally expensive and loss is susceptible to hyperparameter selection [94]

5.3.2 Regression

Successful solutions from related competitions propose to use regression models. In particular, they recommend the use of the following loss functions.

L1

L1, also called Least absolute deviations (LAD) or Mean absolute error (MAE), measures the mean of absolute values of the residuals between each true and predicted label.

$$\mathrm{L1} = \frac{1}{N}\sum_{i=1}^{N}|y_i - y_i'| \tag{5.17}$$

where y_i is the groundtruth value, y_i' is the predicted value and N is the number of samples.

L2

L2 loss, also called Mean squared error (MSE) or Mean squared deviation (MSD), measures the mean of squared residuals between each true and predicted label.

$$\mathrm{L2} = \frac{1}{N}\sum_{i=1}^{N}(y_i - y_i')^2 \tag{5.18}$$

where y_i is the groundtruth value, y_i' is the predicted value and N is the number of samples.

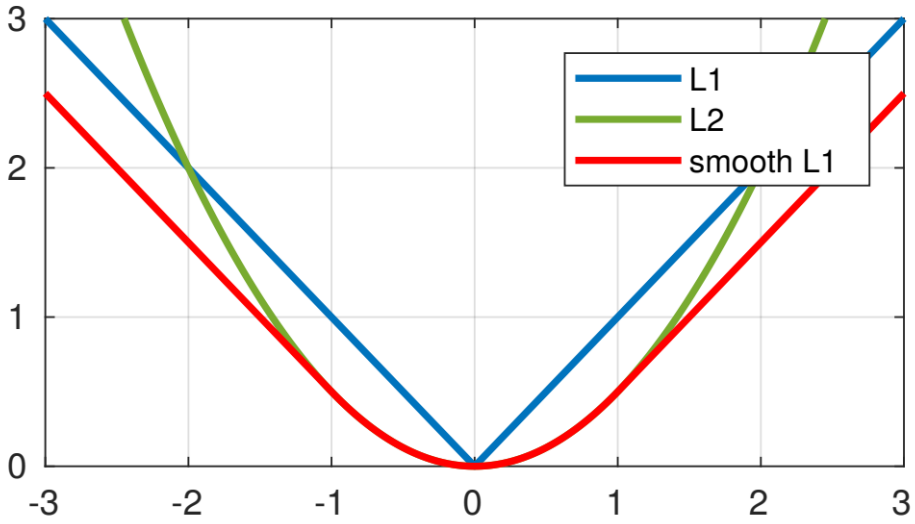


Figure 5.9: Summary of regression loss values on interval [-3,3], source [95].

SmoothL1 loss

SmoothL1 loss can be understood as a combination of L1 and L2 losses. It is defined as follows [95]:

$$\begin{aligned} \text{SmoothL1} &= \frac{1}{N}\sum_{i=1}^{N} z_i \\ z_i &= \begin{cases} \frac{1}{2}(y_i - y_i')^2, & \text{if } |(y_i - y_i')| < 1 \\ |(y_i - y_i')| - \frac{1}{2}, & \text{otherwise} \end{cases} \end{aligned} \tag{5.19}$$

where y_i is the groundtruth value, y_i' is the predicted value and N is the number of samples.

The plots of the L1 loss, L1Smooth loss and L2 loss are depicted in the Figure 5.9.

5.4 Optimiser

Most cases in deep learning include optimising a loss function that cannot be actually evaluated for computational reasons. In these cases, iterative numerical optimisation based on gradient approximation has been most widely adopted [80].

Algorithm 1 Stochastic Gradient Descent

Require: Learning rate η.
Require: Initial parameter θ.
while Stopping criterion not met **do**
  Sample a minibatch of m examples from the training set $\{x^{(1)}, \ldots, x^{(m)}\}$.
  Set $g = 0$
  for $i = 1$ to m **do**
    Compute gradient estimate:

$$g \leftarrow g + \nabla_\theta L(f(x^{(i)}; \theta)), y^{(i)}; \theta).$$

  end for
  Apply update: $\theta \leftarrow \theta - \eta g$
end while

Figure 5.10: Algorithm of SGD, source [96].

5.4.1 Stochastic gradient descent with momentum

Stochastic gradient descent (SGD) is an iterative method based on the gradient descent with stochasticity originating at randomly sampling the gradient over a single data point instead of the whole set [80]. The assumption is that data samples are independent and identically distributed random variables (i.i.d). It is often better to compute the gradient of a mini-batch of m data cases instead of from a single data point. [78]

The algorithm of SGD is described in the Figure 5.10.

Momentum presents an extra adjustment to the SGD that accumulates the gradients of the past steps to improve the convergence and limit unwanted oscilations.

5.4.2 Adam

Adam optimiser [97] presents a solution designed to combine the approaches introduced by momentum and RMSprop [98]. Plainly speaking, while momentum accelerates the search towards the minimum, RMSprop minimises oscillations on the way [99].

5.4.3 AdamW

In 2017, some researches began advising not to use adaptive gradient methods such as Adam [100], because of its poor performance and regularisation. Therefore, Loshchilov et al. [101] proposed the improvement to Adam by

Algorithm 2 Adam with L_2 regularization and Adam with decoupled weight decay (AdamW)

given $\alpha = 0.001, \beta_1 = 0.9, \beta_2 = 0.999, \epsilon = 10^{-8}, \lambda \in \mathbb{R}$
initialize time step $t \leftarrow 0$, parameter vector $\boldsymbol{\theta}_{t=0} \in \mathbb{R}^n$, first moment vector $\boldsymbol{m}_{t=0} \leftarrow \boldsymbol{0}$, second moment vector $\boldsymbol{v}_{t=0} \leftarrow \boldsymbol{0}$, schedule multiplier $\eta_{t=0} \in \mathbb{R}$
repeat
  $t \leftarrow t + 1$
  $\nabla f_t(\boldsymbol{\theta}_{t-1}) \leftarrow \text{SelectBatch}(\boldsymbol{\theta}_{t-1})$ ▷ select batch and return the corresponding gradient
  $\boldsymbol{g}_t \leftarrow \nabla f_t(\boldsymbol{\theta}_{t-1}) + \lambda \boldsymbol{\theta}_{t-1}$
  $\boldsymbol{m}_t \leftarrow \beta_1 \boldsymbol{m}_{t-1} + (1 - \beta_1)\boldsymbol{g}_t$ ▷ here and below all operations are element-wise
  $\boldsymbol{v}_t \leftarrow \beta_2 \boldsymbol{v}_{t-1} + (1 - \beta_2)\boldsymbol{g}_t^2$
  $\hat{\boldsymbol{m}}_t \leftarrow \boldsymbol{m}_t / (1 - \beta_1^t)$ ▷ β_1 is taken to the power of t
  $\hat{\boldsymbol{v}}_t \leftarrow \boldsymbol{v}_t / (1 - \beta_2^t)$ ▷ β_2 is taken to the power of t
  $\eta_t \leftarrow \text{SetScheduleMultiplier}(t)$ ▷ can be fixed, decay, or also be used for warm restarts
  $\boldsymbol{\theta}_t \leftarrow \boldsymbol{\theta}_{t-1} - \eta_t \left(\alpha \hat{\boldsymbol{m}}_t / (\sqrt{\hat{\boldsymbol{v}}_t} + \epsilon) + \lambda \boldsymbol{\theta}_{t-1} \right)$
until *stopping criterion is met*
return optimized parameters $\boldsymbol{\theta}_t$

Figure 5.11: Comparison of Adam and AdamW, source [101] .

replacing the L2 regularisation with decoupled weight decay and called this AdamW.

Training with AdamW should be more robust, providing faster convergence and similar results to training with SGD with momentum [101]. The algorithm is in the Figure 5.11.

5.5 Learning rate scheduler

While training a neural network, it is often beneficial to change the learning rate as the training progresses.

5.5.1 Reduce on plateau

This method is based on binding the learning rate to some evaluation metrics, most commonly validation loss. The idea is to start with a large learning rate to quickly approach the optimal solution and then reduce the learning rate to further improve the result [78]. A sample record of training loss with the Reduce on plateau scheduler is depicted in Figure 5.12.

5.5.2 Cosine Annealing

Loshchilov et al. [103] stated two difficulties in training a neural network. Firstly, they stated that the local minima should not be sharp in order to generalise well. Secondly, that it is often difficult to overcome low gradients saddle points plateaus. To adress these issues, the proposed method uses

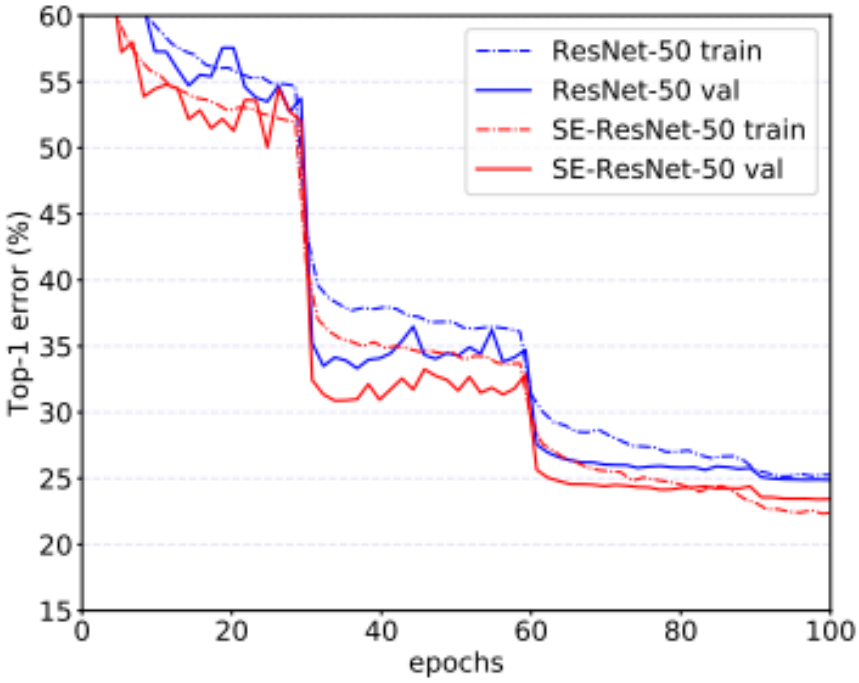


Figure 5.12: Sample use of reduce on plateau, source [102].

cyclic learning rates with warm restarts. Illustration is provided in the Figure 5.13.

5.6 Segmentation

5.6.1 U-Net

U-Net is a neural network architecture that was specifically designed for the semantic segmentation in biomedical application and contains two paths (see Figure 5.14). The first path, called the encoder, is contracting and is used to extract the context. The second path, called the decoder, is expanding and enables the segmentation.

In recent years, there has been a trend to use the U-Net idea of encoder/decoder, but to replace these elements with renowned CNN architectures [108].

5.6.2 Loss

Dice loss was proposed for the optimisation of the models. This loss is based on the Sørensen–Dice coefficient (also called F1 score or simply Dice's coefficient)

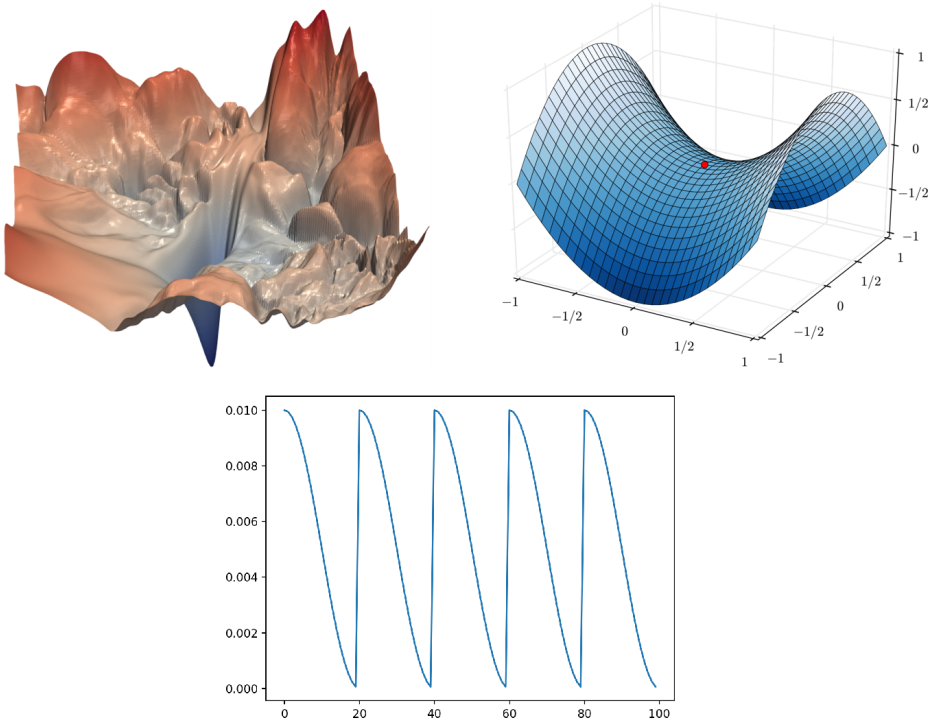


Figure 5.13: Illustration of points mentioned by Loshchilov et al.: On the left, the surface of a loss function (ResNet-56 without skip connections). On the right, illustration of saddle point plateau. On the bottom, the curve of the learning rate produced by Cosine annealing with warm restarts. Sources [104], [105], [106].

[109]. It is a statistical coefficient designed to evaluate the similarity of two samples. When defined on discrete sets, the formula is as follows:

$$\text{DICE} = \frac{2|X \cap Y|}{|X| + |Y|} \tag{5.20}$$

where X and Y are the two sets

When using with binary predictions, the formula can be transformed into the following form:

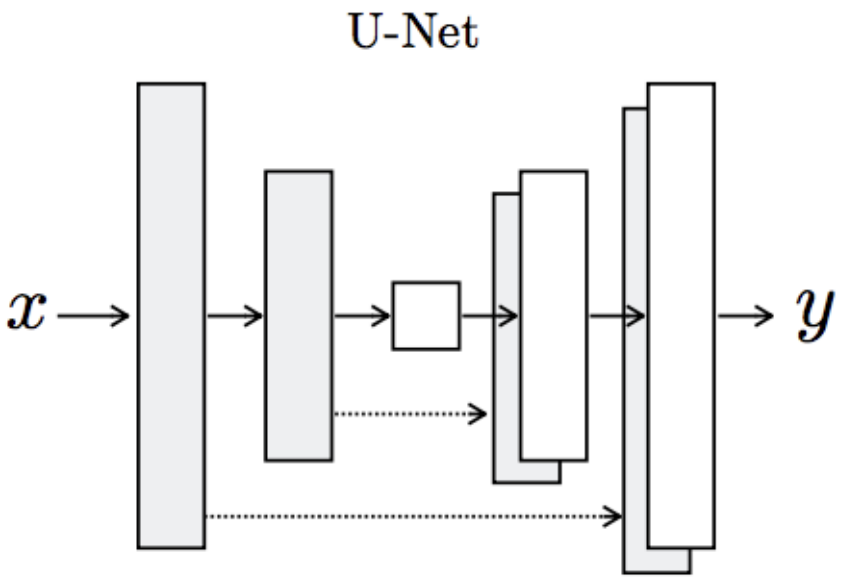
$$\text{DICE} = \frac{2\text{TP}}{2\text{TP} + \text{FP} + \text{FN}} \tag{5.21}$$

where TP are true positives, FP are false positives and FN are false negatives.

The value ranges between 0 and 1.

5.7 Random forest

Random forest is an ensemble learning method that uses a group of Decision Trees. Each tree is trained on a different random subset of the data. Within



.

Figure 5.14: U-Net network schema, source [107].

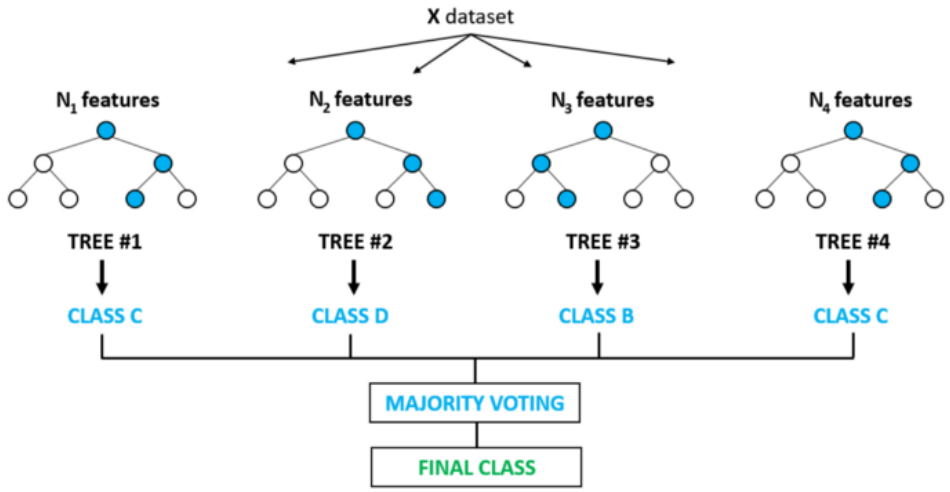
each tree, a random subset of features is used for splitting at each node, and the result is a collection of models, where no two trees are the same. The ensemble then outputs a consensus prediction for each input [110] [111]. The scheme is depicted in Figure 5.15.

5.8 Pseudolabelling

When there are not enough labelled images with consistent labels, Arazo et al. [92] propose to learn from unlabeled data by generating pseudo-labels using the network predictions. They have demonstrated, that naive pseudo-labelling overfits to incorrect pseudo labels and suggest the use of mixup augmentation and setting a minimum number of labelled samples per minibatch as efficient regularisation techniques.

5.9 Model explainability

Neural networks are often criticised for their non-transparent structure and low explainability. SHapley Additive exPlanations can be used to overcome that [113]. SHapley Additive exPlanations (SHAP) are based on Shapely values, that arise from game theory and were used in coalition games. Shapely



.

Figure 5.15: Schematics of random forest, source [112].

value is defined as the average marginal contribution of a feature value across all possible coalitions.

SHAP allows to compute both feature importance (how much is each predictor contributing to the total output) and single case output explanation (what leads to a particular output of the network). [114] [113].

Chapter 6

Methods

This chapter describes the proposed solution. It is divided into several blocks.

Section 6.1 describes the whole scope of the project and its structure. It outlines repository, the folders and main files along with its usage. It also shows the tracking and helper tools as well as the oversight algorithm.

Section 6.2 discusses the baseline solution. Baseline solution consists of single image classification. The section describes not only the preprocessing and data augmentation but also the parameters of the used neural network models (InceptionV3 and DenseNet) with the loss functions (Cross-entropy and Focal loss), optimisers (SGD + momentum and Adam) and learning rate scheduler (Reduce on plateau). In the end, it proposes two configurations for detection.

Section 6.3 proposes several improvements to the baseline solution, but still maintains the task of DR detection from a single image. In the beginning, the section describes the new preprocessing and augmentation. After that, it characterises the two approaches used by this thesis - the classification and the regression approach. For each approach, specific neural networks and loss functions are proposed. In addition to that, state-of-art optimisers (AdamW) and a scheduler (cosine annealing) are proposed. In the end, this section states 9 models for the detection.

Section 6.4 proposes to improve DR detection by utilising the correlations between the two eyes of the same person. To do so, it proposes two algorithms (Winner takes it all and random forest).

Section 6.5 explores the possibility of DR detection improvement using unlabeled data (pseudolabeling). It describes the training process and the methods used.

Section 6.6 proposes the improvement using the additional output (vessel segmentation masks). This was accomplished with the use of segmentation

CNNs. First, the section summarizes the process of creating a combined dataset. Secondly, it contains the information on the preprocessing of this dataset and the augmentations. After that, this section provides a brief overview of the training of the segmentation network, and in the end, it outlines the creation of the additional output for the EyePACS dataset.

Section 6.7 introduces a method designed for improving the DR detection by ensemble classification. This ensembling is done with the help of a dedicated neural network. The section describes the process of acquiring the data for the ensemble and the training of the neural network. It also describes the application of SHAP for feature importance extraction.

Section 6.8 then describes the process of evaluation of the ensamble in the related competitions and datasets.

6.1 Framework

Author decided to use the PyTorch library (version 1.5.1) [115] and Python programming language (version 3.7.4) [116].

The code is attached to this thesis and is also available as Gitlab repository. It is structured as follows:

- Folder `A_preprocessing` contains preprocessing and file division.
- Folder `B_classification` contains the classification framework.
- Folder `C_regression` contains the regression framework.
- Folder `D_segmentation` contains the framework for segmenting vessels.
- Folder `E_postprocessing` contains methods for eye blending and ensambling.
- Folder `F_miscellaneous` contains several data analysis codes.
- Folder `G_plots` contains visualisation codes and the related figures.
- Folder `H_oversight` contains the oversight code.
- Folder `Datasets` contains the EyePACS, APTOS, ISBI and VFN datasets, as well as segmented vessel masks.

A_preprocessing

Each dataset needs to be first preprocessed using the `prep_preprocess.py` file. Using flags, the user can select the input and output folder as well as

the desired output image size. After this preprocessing is done, the images have to be divided into their corresponding classes. That is performed by the code `prep_divide_files_XXX.py`, where `XXX` denotes the name of the dataset. Folder `examples` that contains the sample usage of the code.

B_classification

This folder contains two files. The first one, `diabCNN_classification.py`, is the backend and contains the basic elements such as the train loop, validation loop, all the function definitions and value checks. The second one, `config.py`, is the frontend. It enables model hyperparameters settings. This folder also contains a folder with examples.

C_regresion

The structure of this folder is similar to the classification folder. It also contains the examples as well as the backend file `diabCNN_regression.py` and frontend configuration file `config.py`.

D_segmentation

In this folder, there are several codes related to the segmentation task. File `create_segmentation_dataset.py` inputs the images for the segmentation, augments them and creates the training patches. File `diabCNN_segmentation.py` is the backend for the segmentation network training. The `config.py` is once again used as frontend enabling parameter selection. After the training is done, file `segment_vessels_folder.py` takes all the images in a particular folder, splits them into tiles, which are then evaluated by the segmentation networks, and in the end merged back, creating the whole image segmentation mask.

E_postprocessing

This folder contains the implementation of blending methods. The two files, `post_blending_random_forest.py` and `post_blending_winner.py`, provide the blending of the two eyes. File `post_ensamble_preparation.py` transforms the output predictions from multiple networks (in CSV file) into one large NumPy array, that is then used by the `post_ensamble_neural_net.py` to train the ensemble network. All the codes include functions for export of the predictions, that can be then used for submission.

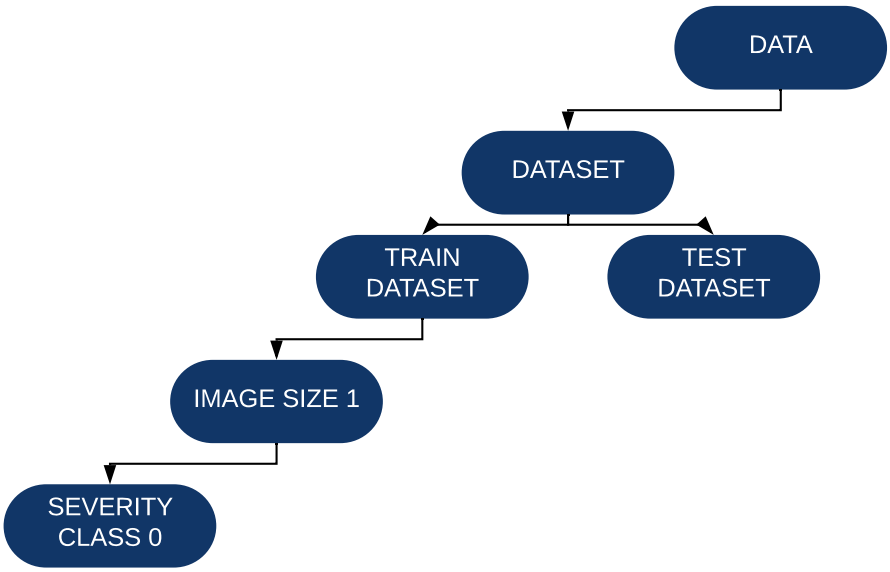


Figure 6.1: Structure of datasets. Only one branch depicted, others were omitted.

H_oversight

This folder contains a single file `parse_slurm_outs_folder.py`, that is used for the oversight. A folder containing reports from the slurm batch jobs needs to be specified. All the reports inside of this folder are processed, visualisations are created, and a Pandas DataFrame summarising the results is printed out.

Datasets

The datasets (EyePACS, APTOS, ISBI, VFN) are stored in the structure described on Figure 6.1. The data is divided into datasets, which are then split to train/test parts. In each of these, there are folders with different image sizes, that contain folders with severity classes. Inside these folders are the images stored in the jpeg format.

Apart from these folders, the framework also lists all the used libraries (`requirements.txt` file) and includes some tools that are worth mentioning:

LR finder

A utility called LR finder has been used to properly select the initial learning rate of the models. The idea is to train for a single epoch using increasing

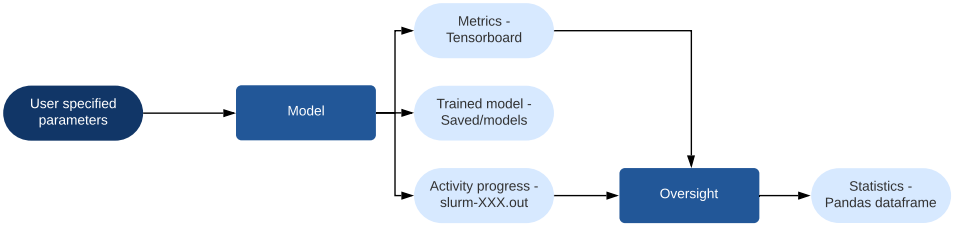


Figure 6.2: Schematics of the tracking system.

learning rate and then estimate boundary learning rates [117]. The tool is available in both the classification and regression backends.

Batch size finder

This thesis proposes a utility to improve training quality. The idea is to utilise the GPU card memory as much as possible. The maximum available batch size of neural network models is accessed using the bisection method. The implementation of this tool is available in both the classification and regression backend.

Tracking

A method of tracking the progress was created (see Figure 6.2. After each epoch, the metrics (loss, accuracy, QWK and confusion matrix) were saved into the Tensorboard environment; the model parameters were saved in a `Saved/Models` folder and the activity progress with hyperparameters was logged in the slurm file. Data collected from all the epochs of the training was packed together.

6.2 Baseline solution

A baseline solution consisting of a simple classification pipeline was established to evaluate the proof of concept. It was inspired by the work of Gulshan et al. [51]. The solution ran on the CMP computational grid [118]. The used server was named boruvka. Boruvka is a machine with 32 cores, 256GB RAM and 8x CUDA GPUs. However, the training was performed on a single GeForce GTX 1080 Ti, 250W card, that was used along with a 11178MiB of memory.

Figure 6.3: Sample images after the preprocessing

6.2.1 Preprocessing

In the baseline solution, this thesis adopted the preprocessing of raw images published by Voets et al. [119]. The idea of the algorithm was to locate the centre and radius of each fundus and resize each image to the size of 299px, with the fundus at its center. Firstly, the image was converted to grayscale, then the contours have been found using the algorithm developed by Suzuki et al. [120]. The largest contour was selected and the minimum enclosing circle [121] was found around this contour. The centre of the retina circle and its radius was then obtained using moments [122] and the original image was cropped adjustingly. In the end, the image was resized to the size of 299x299px and saved as a jpeg image.

6.2.2 Augmentation

To augment the dataset, prevent overfitting and improve generalisation, augmentation of the images was done using the Albumentations library [123]. Namely, training images were randomly flipped along the vertical and horizontal axes, and their contrast was randomly adjusted. The normalisation of the images was done on each channel to the range [0,1]. Testing images were only normalised.

6.2.3 Neural networks

Inspired by the work of Voets et al. [50], Sahlsten et al. [56] and Gao et al. [36], InceptionV3 [65] has been selected as the architectures for the classification of DR.

InceptionV3

InceptionV3 model used weights pretrained on the ImageNet [124]. The output layer has been adjusted to consist of 5 output neurons (classification

approach), whose weights have been initialised using Xavier initialisation [125]. The auxiliary classifier was not used. The implementation of InceptionV3 has been taken from the PyTorch framework.

DenseNet

Pretrained weights from ImageNet database were also used in training. The output layer has been adjusted to consist of 5 output neurons (classification approach), whose weights have been initialised using Xavier initialisation [125]. The implementation of DenseNet has been taken from the PyTorch framework.

6.2.4 Loss

Cross entropy and focal loss were used in the baseline solution. The focal loss parameter γ was set to $\gamma = 2$.

6.2.5 Optimisers

SGD with momentum and the Adam optimiser were used. PyTorch implementation of SGD with momentum used the value 0.9 as the coefficient of momentum, implementation of Adam used the values 0.9, 0.999 as β_1, β_2 respectively.

6.2.6 Learning rate

First attempts were made with a fixed learning rate, but soon the Reduce on plateau scheduler has been adopted.

6.2.7 Summary

The whole pipeline is depicted in Figure 6.4. Several model-loss-optimiser combinations were tested. The models were trained on the EyePACS train dataset in a learning loop for several epochs. After each epoch, instead of validation on a small number of images, the model was evaluated on the complete EyePACS test dataset and the metrics (loss, accuracy, QWK score, confusion matrix) were computed.

Two final models consisting of one InceptionV3 model (see Table 6.1 for parameters) and one DenseNet model (see Table 6.2 for parameters) have been selected.

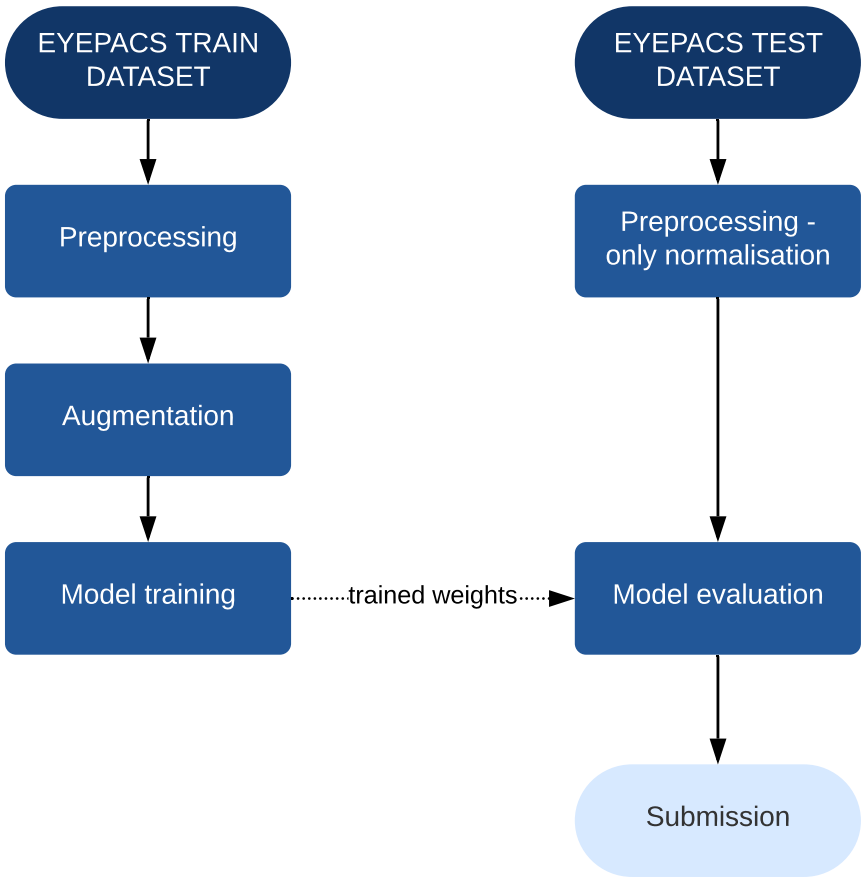


Figure 6.4: Schematics of the baseline solution flow.

Parameter	**Value**
Model name	InceptionV3
Image size	299x299px
Loss function	Focal loss
Optimizer	Adam
Scheduler	None
Learning rate	0.001 (first 9 epochs), 0.0001 (next 10 epochs)
Batch size	64

Table 6.1: Parameters of the InceptionV3 model.

Parameter	Value
Model name	DenseNet
Image size	299x299px
Loss function	Focal loss
Optimizer	Adam
Scheduler	Reduce on plateau
Learning rate	0.0001 (initial)
Batch size	32

Table 6.2: Parameters of the DenseNet model.

6.3 General improvements

To increase the accuracy of the baseline solution, several improvements have been made.

6.3.1 Computing power

The RCI cluster [126] was used for training. The cluster consists of CPU and GPU based compute nodes. Training on the GPU was done on a single NVIDIA Tesla V100, 300W card with a 32480MiB of graphic memory.

6.3.2 Preprocessing

As the images of the retina were in taken with different resolutions, different quality of exposure and with a number of optical artefacts, simple thresholding was not able to generalise well. Therefore, custom preprocessing has been created. The aim of it was to locate the eye, pad the edges if necessary and crop a square-like shape around the retina circle. The NumPy [127] and OpenCV [128] libraries were used for the transformations.

The input image has been padded from top and bottom (by $\frac{1}{3}$ of its height) and from left and right (by $\frac{1}{3}$ of its width). After that, a grayscale copy of the original image was created. This copy was used in two branches. The first branch applied edge detection using the Canny algorithm [129], followed by the finding of contours using the Suzuki algorithm [120] and finding of a convex hull around these contours using Sklansky algorithm [130]. The second branch applied CLAHE - adaptive histogram equalisation. Following CLAHE, the image was thresholded using the Otsu algorithm [131] and the binary output was subjected to morphological closing [132]. Images created by these two branches were then merged using pixel-wise maximum. Following this, the resulted image was blurred by convolving the image with a normalised

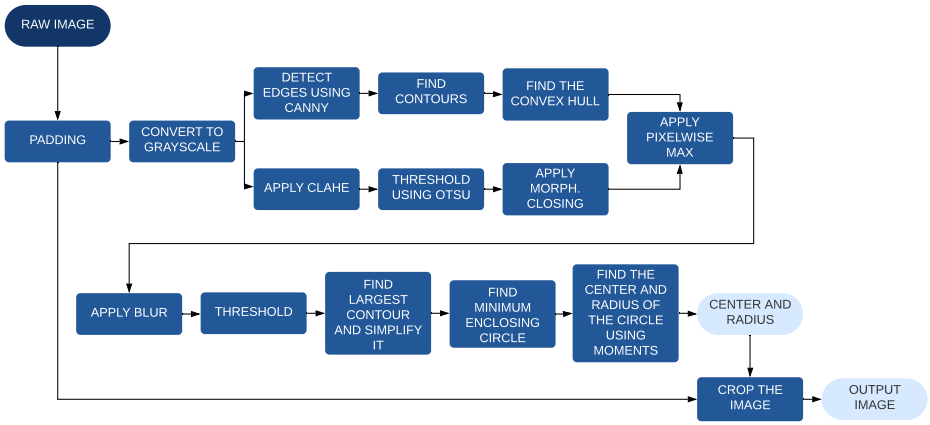


Figure 6.5: Scheme of the proposed preprocessing

box filter [133] and thresholded once again. In the binary output, the largest contour was found and simplified using Ramer–Douglas–Peucker algorithm [134]. A minimum enclosing circle was located on the largest contour [121], and its centre and radius were obtained using moments [122]. Once the coordinates of the center of the circle and its diameter were known, the padded image was cropped so that a square-like box was achieved. In cases where the retina was not of circular shape, the square box was tight on the sides with black space on the top and bottom. The cropped image was scaled to different sizes (from 299x299 up to 1042x1024px). If downsampling of the original image was needed, it was done by resampling using pixel area relation. If upsampling of the image was needed, a bicubic interpolation over 4x4 pixel neighbourhood was done. Final images were saved as a jpeg image with minimum compression.

The scheme of this preprocessing can be found in Figure 6.5, the implementation in the `prep_preprocess.py` file (in the `A_preprocessing` folder). All the images from the EyePACS dataset as well as from APTOS, ISBI and VFN datasets were preprocessed.

6.3.3 Augmentations

Experiments were done with various transformations from the Albumentations library, including vertical flip, horizontal flip, random contrast, random brightness, random gamma and scaling the pixel values on the interval [0,1].

In the later stages of development, training images were augmented only using vertical flip (with a probability of 0.5), horizontal flip (with a probability of 0.5) and scaling the pixel values on the interval [0,1]. Testing images were only scaled to the interval [0,1].

6.3.4 Approach

We propose baseline solution improvement by adopting both classification and regression approaches.

Classification

As in the baseline solution, the classification approach had 5 output neurons (one for each class) and the output of i-th neuron provided probability predicted by the model, that the image belonged to the i-th severity class.

Regression

To better account for the relationship between the severity classes, the regression approach consisted of 1 output neuron with a continuous variable at its output. The variable represented the expected severity and generally should fall into the [0,4] range.

Ordinal regression/classification

In addition to pure classification and regression, there is a third option available – ordinary regression/classification. Even though some authors did adopt this approach, they reported minimum improvement [61], [135]. This thesis, therefore, did not implement this approach.

6.3.5 Neural networks

One of the aims was to try the architectures proposed in the related work and focus on improving the accuracy of a single image prediction. Several neural network architectures have been selected from related work:

- Classification approach: EfficientNetB3, EfficientNetB5, EfficientNetB7, InceptionResnetV2
- Regression approach: EfficientNetB3, EfficientNetB5, EfficientNetB7, InceptionResnetV2, SEResNet50, InceptionV4

Generalized mean pooling

To improve the accuracy of some models (InceptionV4 and EfficientNetB3), generalised mean pooling (GeM) was implemented. In this thesis, the value

of parameter p was selected $p = 3$ as proposed in the original paper [89].

6.3.6 Loss functions

Various loss functions have been tested for both classification and regression

Classification

Classification approach adopted the loss as proposed by Arazo et al.[92], which will be further denoted as the advanced loss. The loss consisted of three terms, and the λ_A, λ_H values have been set to 0.8 and 0.4 respectively. In some cases, the training images were mixed-up.

Cost sensitive loss was also tested. The values in the C matrix were selected to be linearly proportional to the distance between the real label y_i and predicted label y_j as proposed Lin et al. [136]. The cost-sensitive term was added to the advanced loss with a weighted factor of either 10, 1 or 0.1.

Regression

Regression approach adopted the L1, L1Smooth and L2 loss functions.

6.3.7 Optimisers

To improve the convergence and implement regularisation, AdamW optimiser was selected as the optimiser.

6.3.8 Learning rate

This thesis proposes to use the cosine annealing of the learning rate (with a T_{max} value of 15 epochs) to improve the results.

6.3.9 Summary

This thesis proposes to use the 5 final classification networks (see Tables 6.3 and 6.3 for parameters) and 4 final regression networks (see Tables 6.5 and 6.6 for parameters).

Model name	EfficientNet B3	EfficientNet B7	Inception ResnetV2
Image size	512x512px	299x299px	512x512px
Loss function	advanced	advanced	advanced
Mixup	No	Yes	Yes
Cost-sensitive regularization	No	No	No
Optimizer	AdamW	AdamW	AdamW
Scheduler	Cosine annealing	Cosine annealing	Cosine annealing
Learning rate (initial)	0.002	0.0002	0.0002
Batch size	34	32	37
Model tag	m1	m2	m3

Table 6.3: Summary of the final classification models, part 1/2.

Model name	EfficientNet B5	DenseNet
Image size	299x299px	299x299px
Loss function	advanced	advanced
Mixup	No	Yes
Cost-sensitive regularization	No	No
Optimizer	AdamW	AdamW
Scheduler	Cosine annealing	Cosine annealing
Learning rate (initial)	0.002	0.002
Batch size	56	141
Model tag	m4	m5

Table 6.4: Summary of the final classification models, part 2/2.

6.4 Blending of the two eyes

Another improvement proposed by this thesis is to improve the prediction quality by combining the predictions from both eyes. Such information should be relevant [6], and signs of correlations have been found in the EyePACS dataset (see Figure 3.4). Two approaches have been made, both designed to work with the outputs from the classification networks. The scheme of the two methods is described in Figure 6.6.

Model name	EfficientNet B3	SeResNetX50
Image size	512x512px	512x512px
Mixup	Yes	Yes
Loss	L2	L2
Optimizer	AdamW	AdamW
Scheduler	Cosine annealing	Cosine annealing
Learning rate (initial)	0.002	0.00002
Batch size	35	50
Model tag	m6	m7

Table 6.5: Summary of the final regression models, part 1/2.

Model name	EfficientNet B3	EfficientNet B3 + GeM
Image size	512x512px	512x512px
Mixup	Yes	Yes
Loss	L1Smooth	L1Smooth
Optimizer	AdamW	AdamW
Scheduler	Cosine annealing	Cosine annealing
Learning rate (initial)	0.002	0.002
Batch size	35	32
Model tag	m8	m9

Table 6.6: Summary of the final regression models, part 2/2.

Winner takes it all

Probabilities of all classes from both eyes were acquired for each patient in the EyePACS dataset. Class with the highest available probability prediction was then used as a label for both of the eyes. That means, the output of this method always contained the same predictions for both eyes.

Random Forrest

The second approach was built upon the random forest. Input vector to the random forest was created for each patient and consisted of probabilities of all classes for both eyes (10 dimensions in total). The output vector consisted of two continuous variables – the severity class of left and right eye, respectively. The model was trained on patients from the EyePACS training dataset, and grid search was performed to obtain the model with the best hyperparameters. The best model has been evaluated on the EyePACS test dataset.

Later in the development, these blending methods were replaced by a more complex method described in the Ensemble classification section.

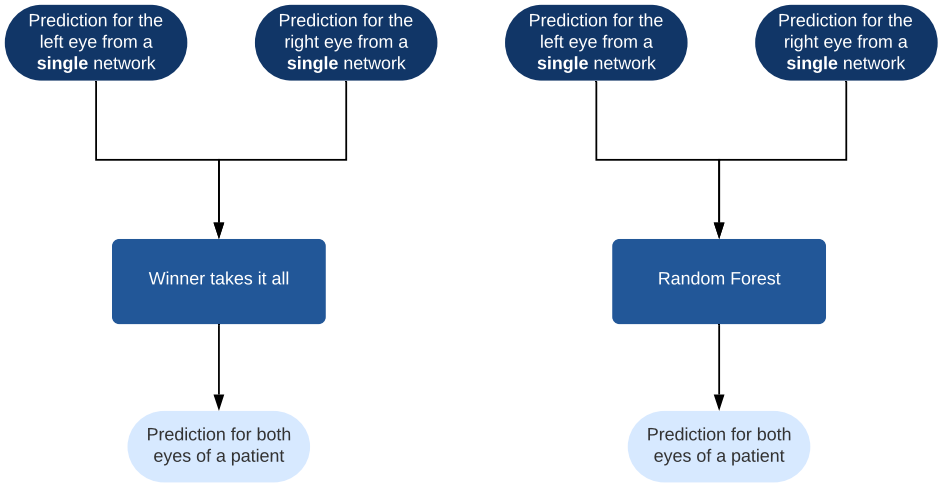


Figure 6.6: Schematics of the proposed blending algorithms. Winner takes it all (left) and random forest (right).

6.5 Using unlabeled data

It is generally known that with neural networks, more data lead to better accuracy and generalisation. Therefore, this thesis proposes to use the pseudolabeling appproach as described by Arazo et al. [92]. In detail, the training was done on a mix of training images and test images. Training images had their ground truth labels, whereas the testing images used pseudo labels based on the network predictions. None of the published ground-truth labels for the testing data has been used. Application of this technique has been discussed with Arazo, and parameters were selected accordingly.

Models specified in the Subsection 6.3.9, have been pretrained using mixup and selected for further training on pseudolabels. The pseudolabels for the training were obtained using the ensemble model (Section 6.7). During the training, the pseudolabels of each batch were updated after each epoch.

Model name	EfficientNet B3	EfficientNet B7	Inception ResNetV2	EfficientNet B5	DenseNet
Loss function	advanced	advanced	advanced	advanced	advanced
Image size	512x512px	299x299px	512x512px	299x299px	299x299px
Original QWK	0.796	0.795	0.781	0.772	0.693

Table 6.7: Impact of pseudolabeling on classification networks.

Model name	SEResNet50	EfficientNetB3	EfficientNetB3	EfficientNetB3 + GeM
Loss function	L2 loss	L2 loss	L1 smooth	L2 loss
Image size	512x512px	512x512px	512x512px	512x512px
Original QWK	0.819	0.821	0.809	0.794

Table 6.8: Impact of pseudolabeling on regression networks.

6.6 Vessel segmentation

In order to further improve the performance of DR detection, this thesis proposes to use vessel segmentation masks as an extra input (along with the RGB image) for the classification.

Oliviera et al. [137] summarises the related work in this area and proposes the use of convolutional neural networkds for the task of semantic segmentation.

6.6.1 Datasets

As the number of images in the datasets available for vessel segmentation is low, several of the datasets were combined. Namely:

- The DRIVE dataset [138] has been created in order to establish a standard in comparative studies on blood vessel segmentation. It consists of 20 training images with a pixel-wise masks annotating blood vessels and 20 testing images where the annotated masks were not available at the time of thesis writing.
- High-Resolution Fundus (HRF) Image Database [139] contains 15 images of healthy patients, 15 images of patients with diabetic retinopathy and 15 images of glaucomatous patients. Annotated Binary segmentation masks are available for each image.
- CHASEDB1 is a dataset made by Kingston University [140]. It contains 28 retinal images with annotated masks.

In total, 93 images with the corresponding binary segmentation masks were acquired and divided by 80:20 ratio into 75 train images and 18 test (validation) images.

```
albumentations_test_transform = albi.Compose([ # for testing dataset
    albi.RandomCrop(p=1, height=128, width=128),
])

albumentations_train_transform = albi.Compose([ # for train dataset
    albi.RandomCrop(p=1, height=128, width=128),
    albi.HorizontalFlip(p=0.5),
    albi.VerticalFlip(p=0.5),
    albi.RandomRotate90(p=0.5),
    albi.Transpose(p=0.5),
    albi.ShiftScaleRotate(shift_limit=0.01, scale_limit=0.04, rotate_limit=0, p=0.25),
    albi.RandomBrightnessContrast(p=0.5),
    albi.RandomGamma(p=0.25),
    albi.IAAEmboss(p=0.25),
    albi.Blur(p=0.01, blur_limit = 3),
    albi.OneOf([
        albi.ElasticTransform(p=0.5, alpha=120, sigma=120 * 0.05, alpha_affine=120 * 0.03),
        albi.GridDistortion(p=0.5),
        albi.OpticalDistortion(p=1, distort_limit=2, shift_limit=0.5)
    ], p=0.8),
])
```

Figure 6.7: Complete list of applied transformations.

6.6.2 Preprocessing

Related literature [141], [142], [143], [144] proposes to either extract the green channel or convert the original image to grayscale as the vessels have the most contrast in it. To enhance the image, CLAHE is applied to the grayscale image.

This preprocessing has been applied on the segmentation dataset, and all the images were saved in the resolution 512x512px. As the dataset is of small size, we propose to create patches from the augmented preprocessed images.

Albumentations library is used to heavily augment the images and extract 500 patches of size 128x128 from every image. The complete list of transformations with parameters is in Figure 6.7. Some of the parameters of the transformations were copied from the example usage notes available on the Albumentations website [145].

This operation resulted in the creation of the final segmentation dataset, that contained 37500 $(75 \cdot 500)$ training patches and 9000 $(18 \cdot 500)$ testing (validation) patches.

6.6.3 Segmentation architecture

For segmentation, U-Net [146] like models were selected. This thesis adopted three of these networks and used Python implementations available online [147]:

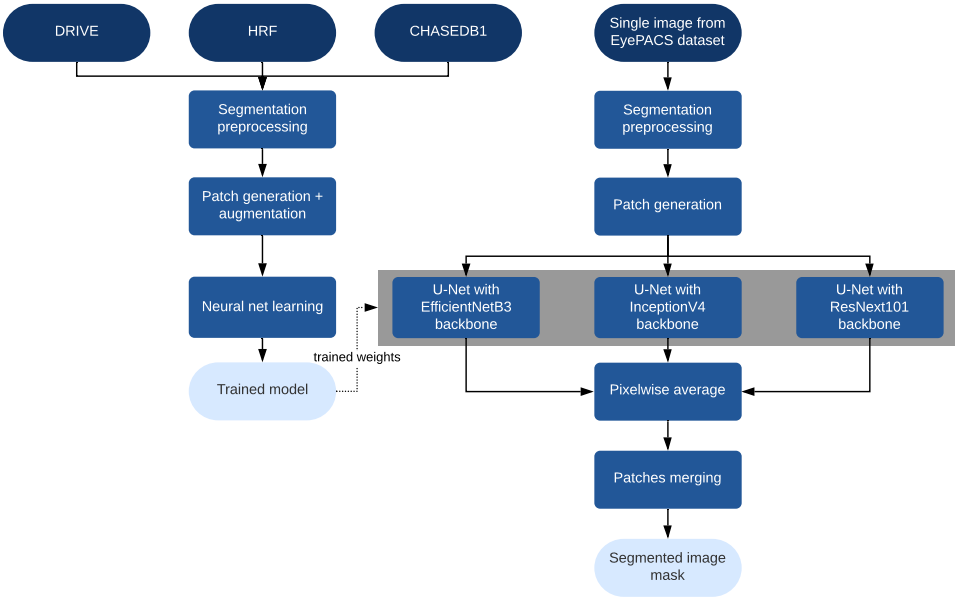


Figure 6.8: Schematics of the segmentation approach. On the left is depicted the training of the segmentation models on segmentation datasets. On the right is depicted the creation of segmented masks of the EyePACS dataset.

- UNET with EfficientNetB3 [148] backbone (10 milion parametres)
- UNET with InceptionV4 [149] backbone (41 milion parametres)
- UNET with ResNext101 [150] backbone (46 milion parametres)

Use of multiple CNN architectures was motivated by considering the trade-off between bias and variance [78]. Different types of CNNs were used to ensure the diversity of the predictions. The predictions were in the form of segmented masks with a probability of blood vessel (in the range [0,1]) for each pixel.

As a loss function, Dice loss was selected. The implementation of Dice loss from Liu [151] has been used in this thesis.

Creation of the segmentation masks

After training and evaluation of the three segmentation models, an ensemble pipeline was created (see Figure 6.8). This pipeline took the images from the EyePACS dataset one by one, every time converting the preprocessed 512x512px version into grayscale, applying CLAHE and splitting it into 16 non-overlapping patches of size 128x128px. These patches were then inputted into the three segmentation models, and output predictions were averaged pixel-wise. Resulting masks were then multiplied by a factor of 255 to get a grayscale image of the predicted vessels.

These grayscale images were then merged with the original RGB images to obtain a 4 channel images and four neural networks architectures were modified to accept the 4-channel input. As the networks contained the pretrained weights from the ImageNet dataset, weights for the 4th channel were copied from the already present green channel.

Two scenarios were proposed. In the first one, the networks were trained from scratch on the training dataset without the use of pseudolabels. In the second one, the networks were trained with the help of pseudolabels on both the training and the testing datasets. Their summary is in the Table 6.9.

Model name	SEResNet50	EfficientNet B3	EfficientNet B3	EfficientNet B3 + GeM
Loss function	L2 loss	L2 loss	L1 smooth	L2 loss
Image size	512x512px	512x512px	512x512px	512x512px
Scheduler	Cosine annealing	Cosine annealing	Cosine annealing	Cosine annealing
Learning rate	0.00002	0.002	0.002	0.002
Batch size	45	28	28	25

Table 6.9: Summary of the networks proposed for the training with the 4 channel input.

6.7 Ensemble classification

The final improvement was suggested to arise from ensembling and blending at the same time. In total, the ensamble consisted of 9 networks (specified in the Subsection 6.3.9).

One vector was created from their outputs (5 probability classes in case of classification, 1 continuous value variable in case of regression) for both eyes. The pipeline is specified in the Figure 6.9.

6.7.1 Neural network

Specialised ensembling neural network has been created. The architecture was proposed to be shallow to prevent overfitting. Input vector had a dimension of 58, output vector had the dimension of 2 – (regression outputs for left and right eye). Details of the architectures are available in Figure 6.10. This thesis also proposes an alternative architecture, that has four hidden layers instead of two.

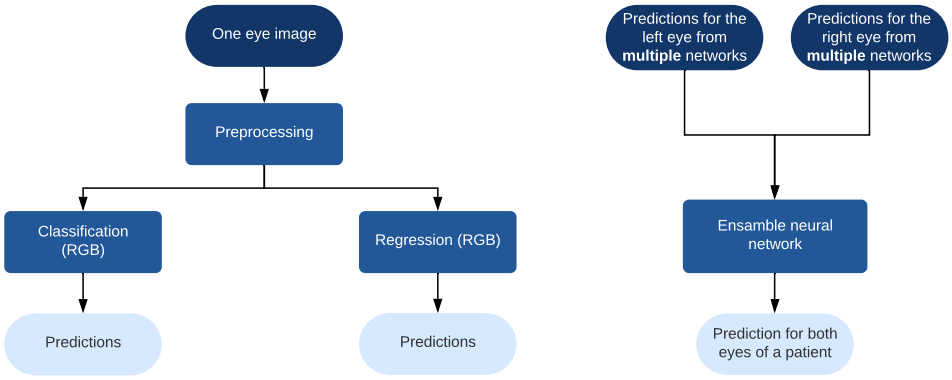


Figure 6.9: Multiple prediction on a single image (left). Schematics of the proposed ensamble method (right).

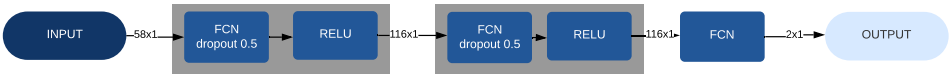


Figure 6.10: Schematics of the proposed ensembling neural network architecture with described data size.

The AdamW optimiser was used alongside cosine annealing scheduler to optimise the L2 loss function. The neural network has been trained on the EyePACS training set and evaluated on the testing dataset.

6.7.2 Model explainability

In this thesis, a straightforward form of SHAP has been implemented in the `post_ensamble_neural_network.py` file. The purpose was to create visualisation of feature importance of the ensembling neural network. It could also be used to explain the individual predictions of CNNs, but this application has been left undone. The SHAP evaluator was fitted to 10000 randomly chosen patients from the training dataset, and the feature importance visualisation depicts the mean SHAP values of the predictors over 300 randomly chosen patients.

6.8 Evaluation

The ensemble model trained on the EyePACS training dataset and then evaluated in the related competitions.

- In case of the Diabetic Retinopathy Detection 2015 competition, it was evaluated on the testing dataset.

- In case of the APTOS 2019 Blindness Detection, the ensemble was evaluated on the training part of the APTOS dataset.
- In case of the 2020 – ISBI - The 2nd Diabetic Retinopathy – Grading and Image Quality Estimation Challenge, the ensemble was finetuned by training for one epoch on the training part of the ISBI dataset and then evaluated on the validation part of the ISBI dataset.
- In case of VFN, no patient information was provided and the dataset was evaluated by a single EfficientNet B3 model (tag m6).

Chapter 7

Results

This chapter states the results of the proposed methods. Unless stated otherwise, the results are from the EyePACS testing dataset.

Section 7.1 provides a summary of the baseline solution, namely the preprocessing and the performance of the two selected models.

Section 7.2 reports the results obtained by the improved preprocessing and from the selected 5 classification and 4 regression neural networks.

Section 7.3 describes the performance of the two blending algorithms, Winner takes it all and random forest.

Following this, the Section 7.4 reports the impact of using unlabeled data on selected models.

Section 7.5 contains the results of training the vessel segmentation networks as well as the impact of using the vessel segmentation masks as additional input for the classification.

Section 7.6 reports the influence of the ensemble neural network.

Section 7.7 provides a final evaluation of the proposed solution and the results in related competitions.

7.1 Baseline solution

7.1.1 Preprocessing

Sample images from the baseline preprocessing are available in Figure 7.1.

Figure 7.1: Sample images after the preprocessing as proposed in the baseline solution.

7.1.2 Neural networks

The results of the two networks proposed in Subsection 6.2.7 are described. Namely, the results of the InceptionV3 model are described in the Table 7.1, the results of the DenseNet model are described in the Table 7.2.

Parameter	Value
Model name	InceptionV3
Best QWK	0.5335
Best epoch	17
Model size	85.355MB
Model training device	boruvka
Model training time	4.47h

Table 7.1: Distribution of the EyePACS dataset

Parameter	Value
Model name	DenseNet
Best QWK	0.5960
Best epoch	6
Model size	27.662MB
Model training device	boruvka
Model training time	2.65h

Table 7.2: Distribution of the EyePACS dataset

7.2 General improvements

7.2.1 Preprocessing

Figure 7.2 shows a single image from the EyePACS dataset during various stages of the preprocessing. The statistics of the improved preprocessing of the EyePACS dataset are available in Table 7.3.

The proposed preprocessing improved the quality of the trained models (in terms of QWK score). Figure 7.3 shows a representative example, where two identical EfficientNetB3 networks were trained on the 299x299px images using cosine annealing. One model was trained on the data preprocessed by the preprocessing from the baseline solution, the second one on the data from the improved preprocessing. The results represents a trend which was seen across multiple models.

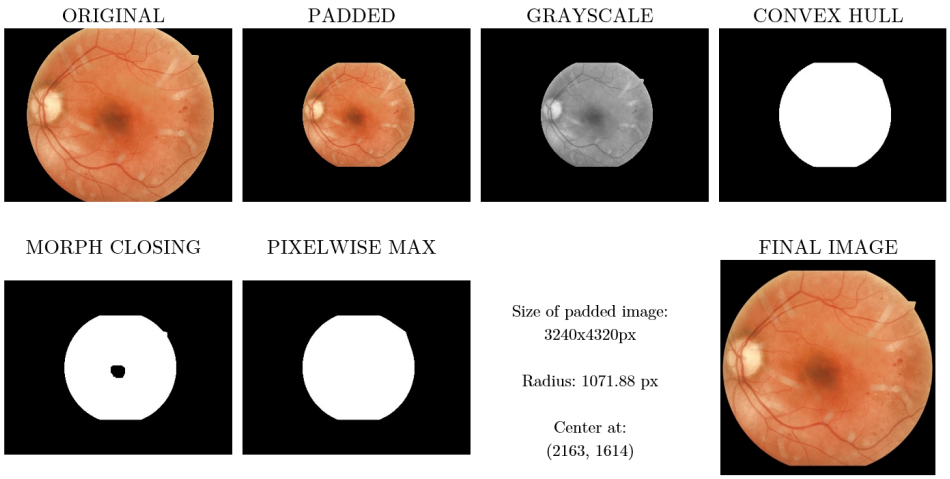


Figure 7.2: Single image during the preprocessing.

Data	Number of images	Duration of preprocessing	Time per image
Training data	35126	38610.4s	0.91s
Testing data	53576	45741.8	1.17s

Table 7.3: Distribution of the EyePACS dataset

7.2.2 Neural networks

The final results of the classification networks are summarized in the Tables 7.4 and 7.5. The final results of the regression networks are summarized in the Tables 7.6 and 7.7.

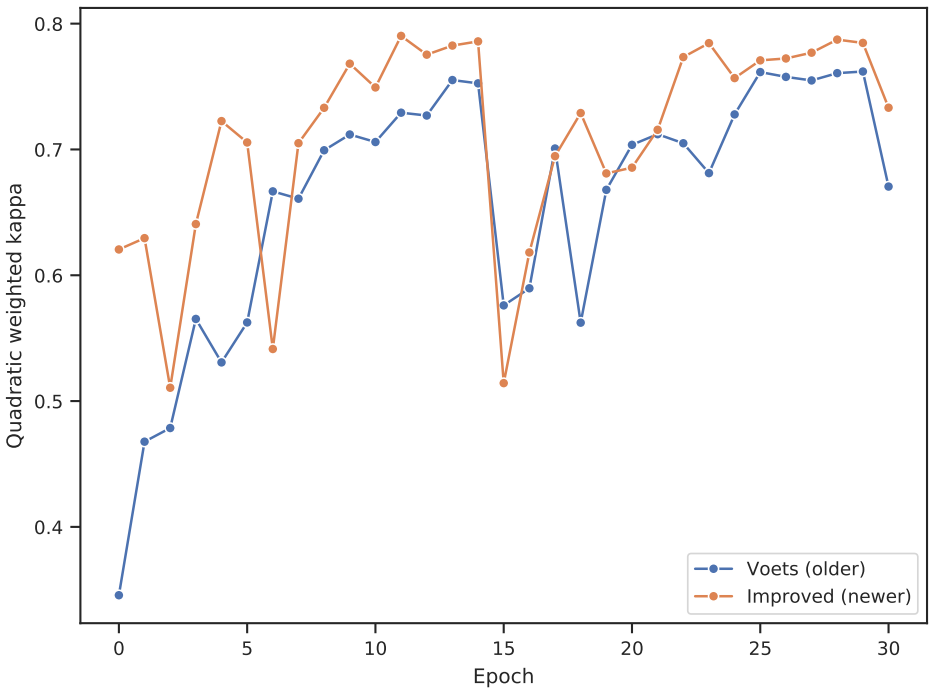


Figure 7.3: Comparison of preprocessing.

Model name	EfficientNet B3	EfficientNet B7	Inception ResnetV2
Tag	m1	m2	m3
Image size	512x512px	299x299px	512x512px
Mixup	No	Yes	Yes
Best QWK	0.796	0.795	0.781
Best epoch	9	5	11
Model size	42.26MB	250.65MB	212.65MB
Model training device	RCI	RCI	RCI
Model training time	11.025h	2.3h	19.25h

Table 7.4: Summary of the best classification models, part 1/2.

7.3 Blending of the two eyes

The results of the proposed Winner takes it all algorithm are in Table 7.8, results of the random forest classifier are summarized in Table 7.9.

Model name	EfficientNet B5	DenseNet
Tag	m4	m5
Image size	299x299px	299x299px
Mixup	No	Yes
Best QWK	0.772	0.693
Best epoch	10	38
Model size	111.582MB	27.67MB
Model training device	RCI	RCI
Model training time	3.167h	17.1h

Table 7.5: Summary of the best classification models, part 2/2.

Model name	EfficientNet B3	SeResNetX50
Tag	m6	m7
Image size	512x512px	512x512px
Mixup	Yes	Yes
Loss	L2	L2
Best QWK	0.821	0.819
Best epoch	14	9
Model size	42.237MB	102.803MB
Model training device	RCI	RCI
Model training time	15.63h	12.9h

Table 7.6: Summary of the best regression models, part 1/2.

Model name	EfficientNet B3	EfficientNet B3 + GeM
Tag	m8	m9
Image size	512x512px	512x512px
Mixup	Yes	Yes
Loss	L1Smooth	L1Smooth
Best QWK	0.808	0.793
Best epoch	14	13
Model size	42.237MB	42.237MB
Model training device	RCI	RCI
Model training time	17.23h	16.25h

Table 7.7: Summary of the best regression models, part 2/2.

7.4 Using unlabeled data

Impact of the pseudolabelling approach on the networks specified in Section 6.5 is described. The classification networks' performance is in the Table

Model name	DenseNet	InceptionV3	EfficientNet B3
Original QWK	0.67	0.62	0.796
Improved QWK	0.71	0.66	0.801

Table 7.8: Performance of the winner takes it all algorithm.

Parameter	Value
Pretrained Model	EfficientNet B3
Number of trees	40
Max depth	50
Original QWK	0.796
Improved QWK	0.818

Table 7.9: Random forrest summary.

7.10, whereas the impact on the regression networks is summarized in the Table 7.11. The training time of the networks with pseudolabelling was not recorded, but generally prolonged the time of the original training by half.

Model name	EfficientNet B3	EfficientNet B7	Inception ResNetV2	EfficientNet B5	DenseNet
Loss function	advanced	advanced	advanced	advanced	advanced
Image size	512x512px	299x299px	512x512px	299x299px	299x299px
Original QWK	0.796	0.795	0.781	0.772	0.693
Improved QWK	0.801	0.807	0.795	0.775	0.749
Best epoch	5	9	6	11	8

Table 7.10: Impact of pseudolabeling on classification networks.

7.5 Vessel segmentation

The results of the three segmentation networks are described in the Table 7.12. Figure 7.4 shows samples of the predicted patches and their corresponding groundtruths.

Results of the creation of the segmentation masks for the EyePACS dataset is represented in Figure 7.5.

Model name	SEResNet50	EfficientNetB3	EfficientNetB3	EfficientNetB3 + GeM
Loss function	L2 loss	L2 loss	L1 smooth	L2 loss
Image size	512x512px	512x512px	512x512px	512x512px
Original QWK	0.819	0.821	0.809	0.794
Improved QWK	0.831	0.829	0.820	0.815
Best epoch	6	7	9	7

Table 7.11: Impact of pseudolabeling on regression networks.

The impact of 4 channel DR detection was evaluated on the selected regression networks in two scenarios. The first was done without the use of pseudolabeling (Table 7.13) and the second one with the use of pseudolabeling (Table 7.14).

Parameter	**EfficientNetB3 backbone**	**InceptionV4 backbone**	**ResNext101 backbone**
Best epoch	39	94	9
IOU background	0.927	0.927	0.926
IOU vessel	0.673	0.671	0.672
Precision background	0.977	0.977	0.974
Precision vessel	0.746	0.745	0.753
Recall background	0.948	0.948	0.949
Recall vessel	0.874	0.872	0.862

Table 7.12: Results of the segmentation network.

Model name	SEResNet50	EfficientNet B3	EfficientNet B3	EfficientNet B3 + GeM
Loss function	L2 loss	L2 loss	L1 smooth	L2 loss
Image size	512x512px	512x512px	512x512px	512x512px
Best QWK	0.709	0.788	0.787	0.788
Best epoch	6	7	7	7

Table 7.13: Performance of regression networks that use 4 channel input (no pseudolabelling).

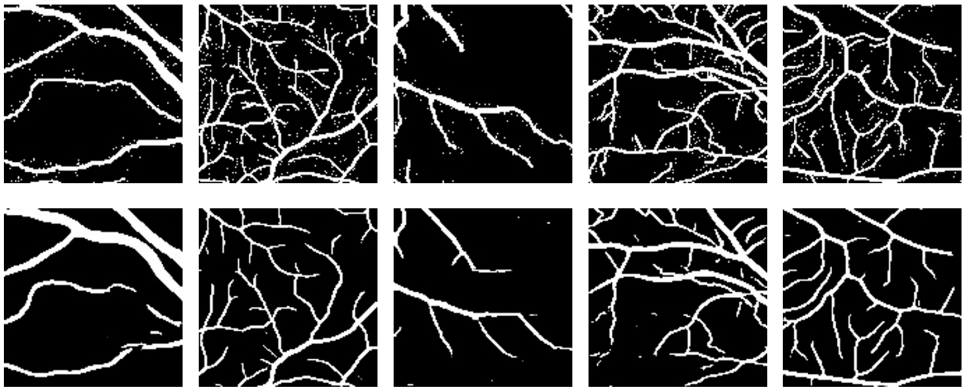


Figure 7.4: Comparison of the ground-truth masks (first row) with the corresponding masks predicted by the trained UNet InceptionV4 model (second row).

Model name	SEResNet50	EfficientNet B3	EfficientNet B3	EfficientNet B3 + GeM
Loss function	L2 loss	L2 loss	L1 smooth	L2 loss
Image size	512x512px	512x512px	512x512px	512x512px
Best QWK	0.825	0.837	0.838	0.837
Best epoch	8	7	7	7

Table 7.14: Performance of regression networks that use 4 channel input (with pseudolabelling).

7.6 Ensemble classification

The results of the two proposed architectures are outlined in Table 7.15. Results are discussed in detail in the Subsection 7.7.1.

The SHAP results describing the importance of individual predictors (networks' outputs) are described in Figure 7.6.

Number of hidden layers	2	4
Best QWK	0.850	0.807
Best epoch	13	10
Model size	82KB	980KB
Model training device	RCI	RCI
Model training time	2.31min	11.12min

Table 7.15: Summary of the ensemble neural network architectures

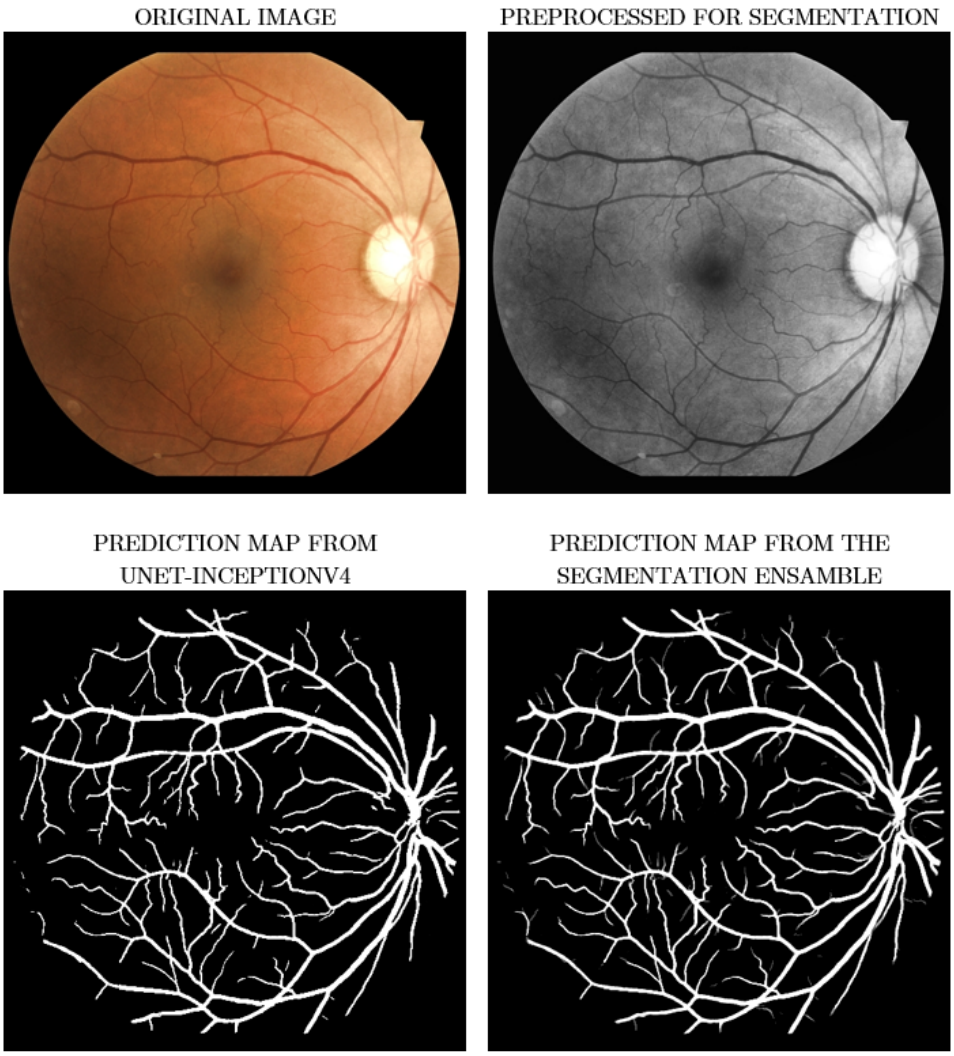


Figure 7.5: Sample image during the creation of the segmentation masks.

7.7 Evaluation

7.7.1 Diabetic Retinopathy Challenge - 2015

The final ensemble model was submitted into the competition and achieved QWK score of 0.85097. The detailed results are in the Figure 7.9. The proposed solution achieved first place in the competition. The results are discussed in more detail in the Table 7.16 and in the confusion matrices (see Figure 7.7 and Figure 7.8)

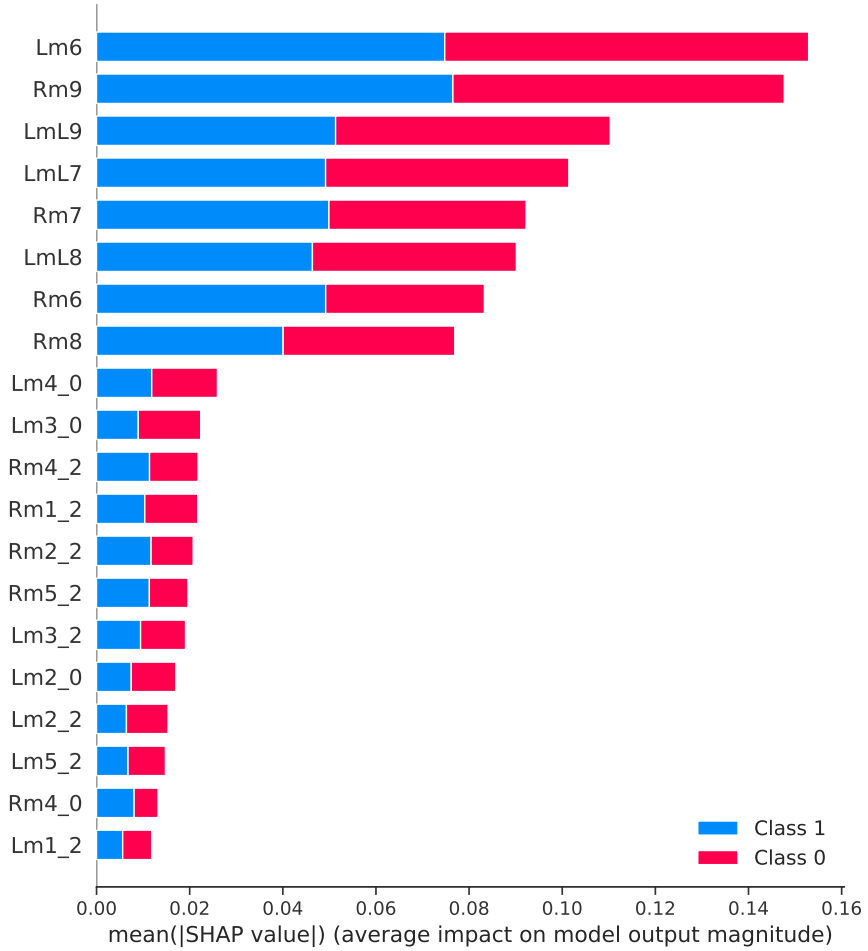


Figure 7.6: Feature importance of the ensamble network. The first letter in the label (L/R) denotes the eye for which the prediction was originally made. The following two chars (e.g. m6) represent the model tag, as described in Subsection 6.3.9. In case of classification networks (m1-m5), there is an extra number behind an underline, that marks the output of the particular DR severity class. Class 0 describes the prediction of the ensemble for the left eye; class 1 describes the prediction for the right eye.

7.7.2 APTOS 2019 Blindness Detection

The final achieved QWK score on the training data is 0.876. The more detailed results are in the Table 7.17 and in the confusion matrices (Figure 7.10 and Figure 7.11).

Label	Class	Precision	Recall	F1 score	Support
0	No DR	0.94	0.93	0.93	39533
1	Mild DR	0.27	0.42	0.33	3762
2	Moderate DR	0.76	0.42	0.54	7861
3	Severe DR	0.26	0.72	0.38	1214
4	Proliferate DR	0.75	0.40	0.52	1206

Table 7.16: Summary of the final model performance on the EyePACS dataset.

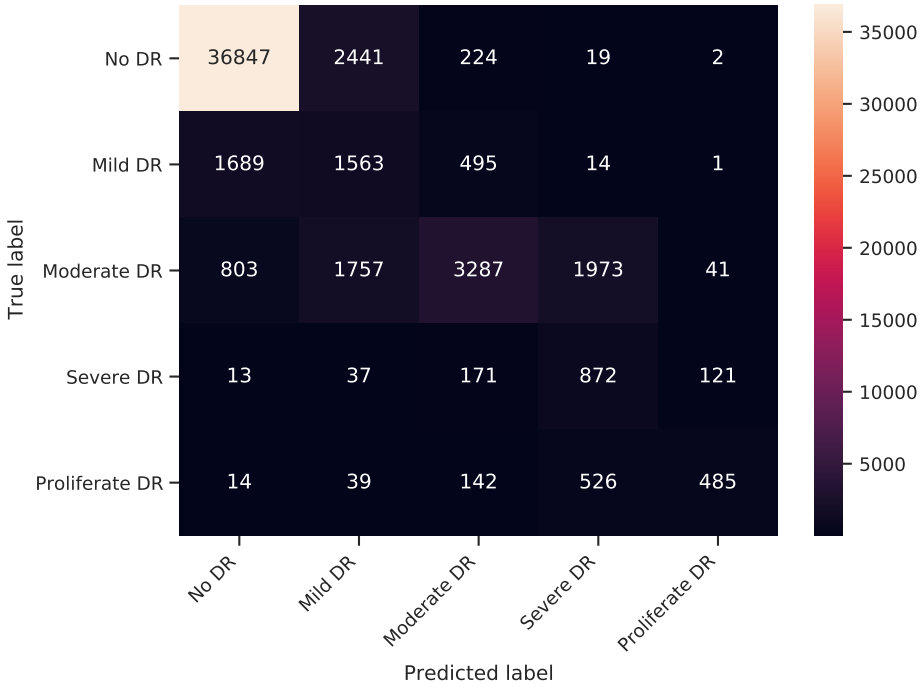


Figure 7.7: Confusion matrix of the final ensemble model on the EyePACS testing dataset.

7.7.3 ISBI - The 2nd Diabetic Retinopathy - Grading and Image Quality Estimation Challenge - 2020

The final achieved QWK score on the validation data is 0.756. The results are described in detail in Table 7.18 and in the confusion matrices (see Figure 7.12 and Figure 7.13).

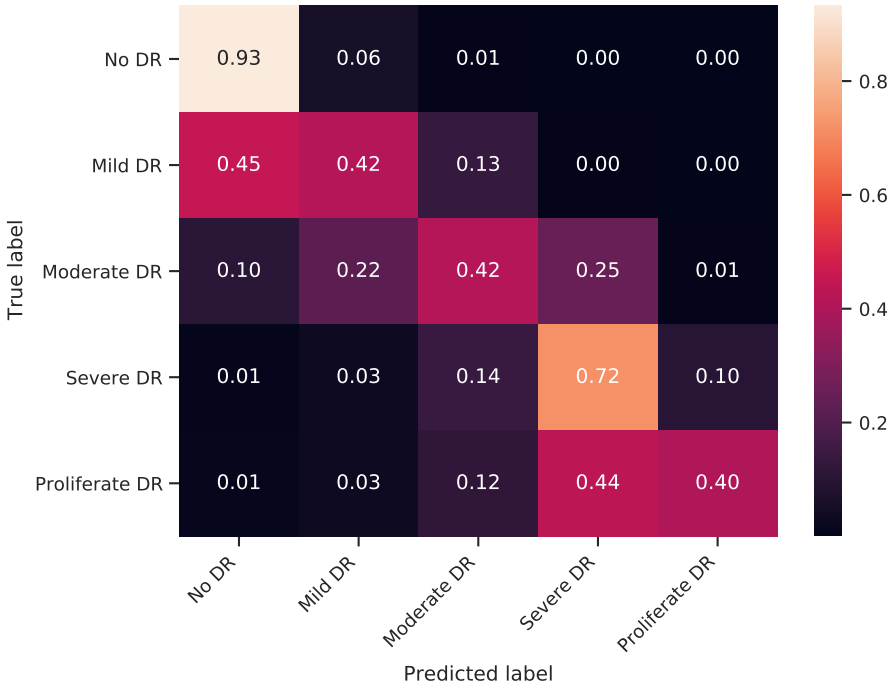


Figure 7.8: Normalised confusion matrix of the final ensemble model on the EyePACS testing dataset.

Submission and Description	Private Score	Public Score
SUBMISSION_seventhTry.csv by Vojtech Poriz EnsambleNNfrom9	0.85097	0.85467

Figure 7.9: Screenshot of the submission system of the competition.

Label	Class	Precision	Recall	F1 score	Support
0	No DR	0.99	0.94	0.96	1805
1	Mild DR	0.20	0.07	0.10	370
2	Moderate DR	0.56	0.46	0.50	999
3	Severe DR	0.16	0.67	0.26	193
4	Proliferate DR	0.65	0.43	0.52	295

Table 7.17: Summary of the final model performance on the APTOS training dataset.

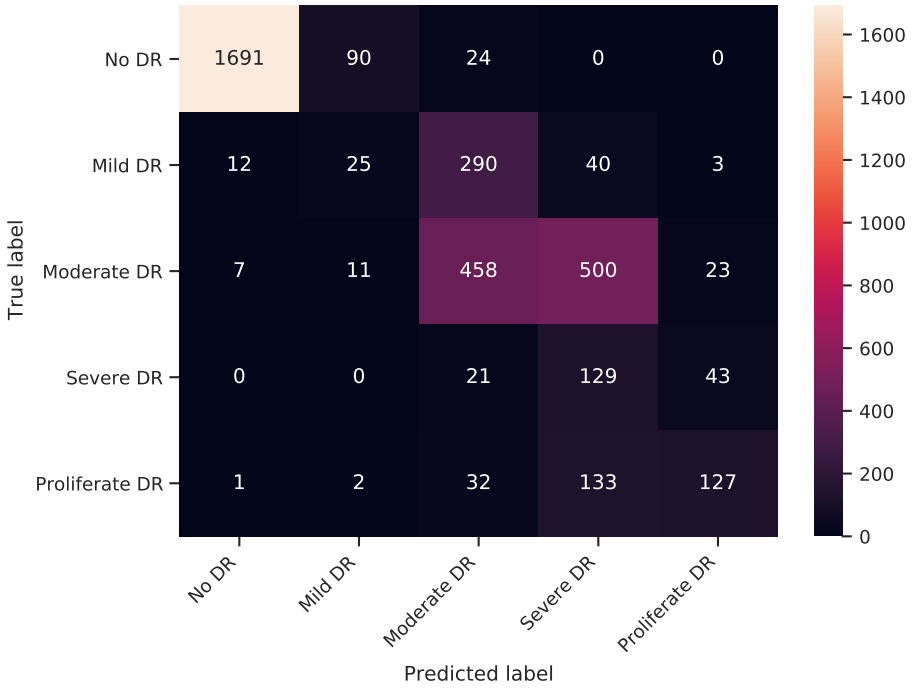


Figure 7.10: Confusion matrix of the final ensemble model on the APTOS training dataset.

Label	Class	Precision	Recall	F1 score	Support
0	No DR	0.81	0.53	0.64	174
1	Mild DR	0.21	0.48	0.29	46
2	Moderate DR	0.56	0.46	0.50	999
3	Severe DR	0.16	0.67	0.26	193
4	Proliferate DR	0.65	0.43	0.52	295

Table 7.18: Summary of the final model performance on the ISBI validation dataset.

7.7.4 VFN

The final QWK score of the EfficientNet B3 model (tag m6) on this dataset is 0.510. The result is summarized in Table 7.19 and in the confusion matrices (see Figure 7.14 and Figure 7.15).

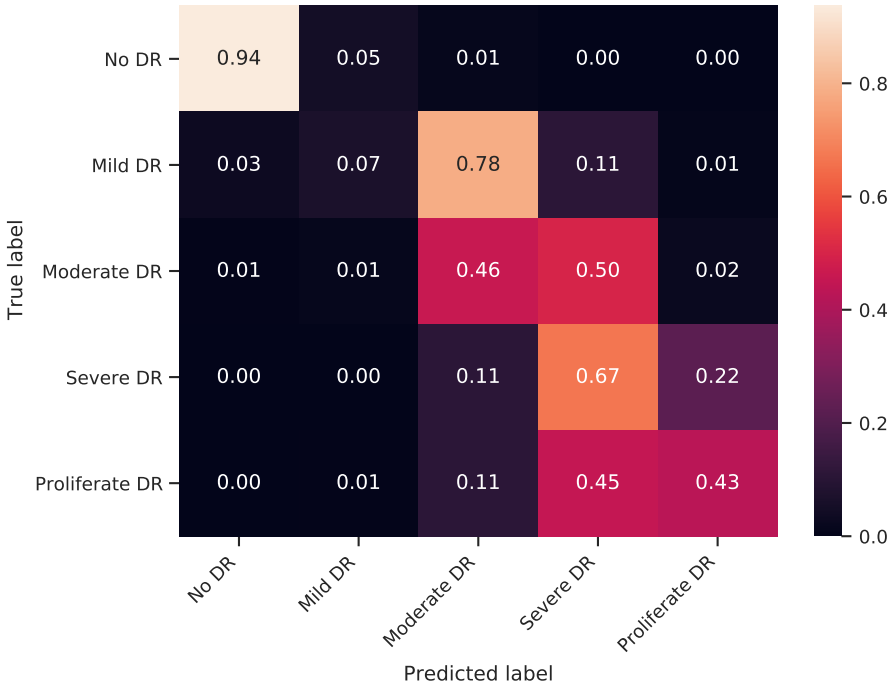


Figure 7.11: Normalised confusion matrix of the final ensemble model on the APTOS training dataset.

Label	Class	Precision	Recall	F1 score	Support
0	No DR	0.00	0.00	0.00	0
1	Mild DR	0.33	0.09	0.14	11
2	Moderate DR	0.40	0.40	0.40	20
3	Severe DR	0.00	0.00	0.00	0
4	Proliferate DR	1.00	0.20	0.33	20

Table 7.19: Summary of the final model performance on the VFN dataset.

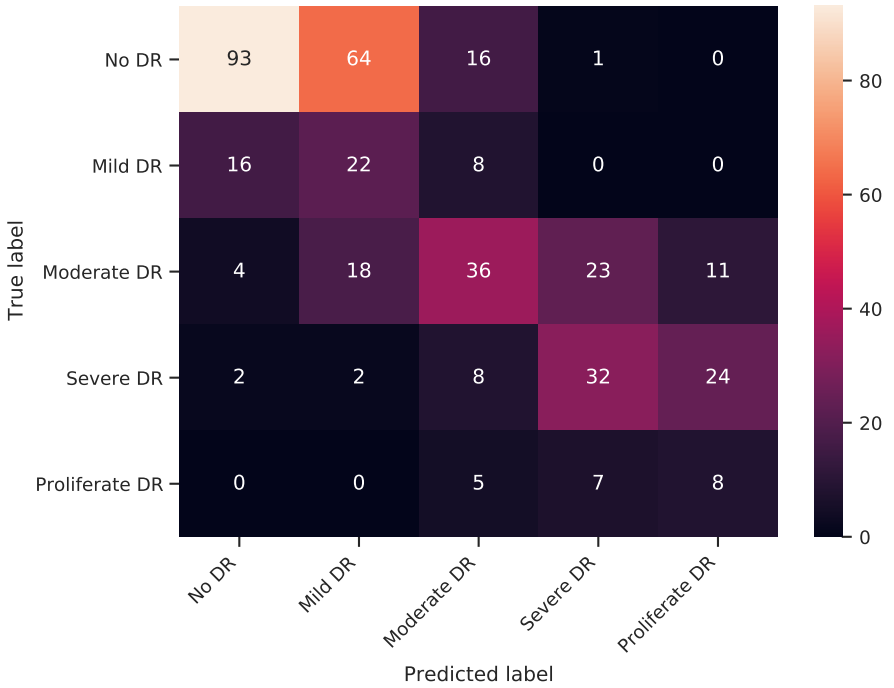


Figure 7.12: Confusion matrix of the final ensemble model on the ISBI validation dataset.

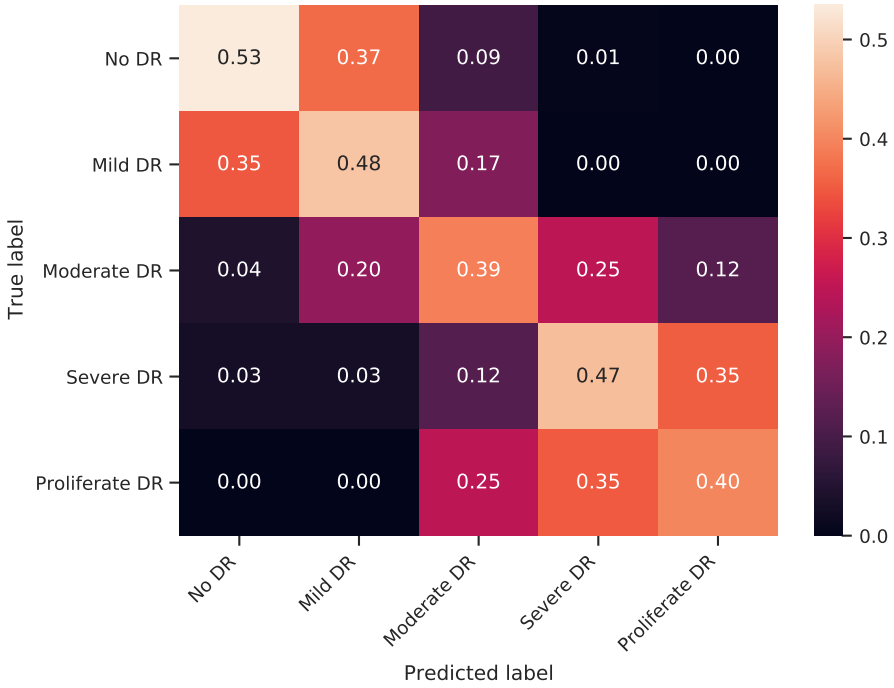


Figure 7.13: Normalised confusion matrix of the final ensemble model on the ISBI validation dataset.

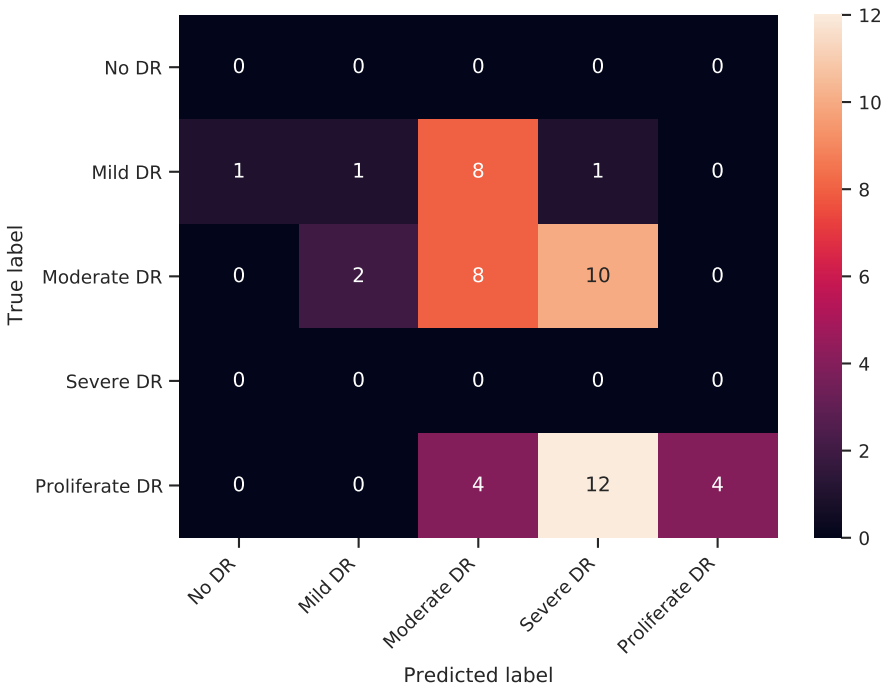


Figure 7.14: Confusion matrix of the final ensemble model on the VFN dataset.

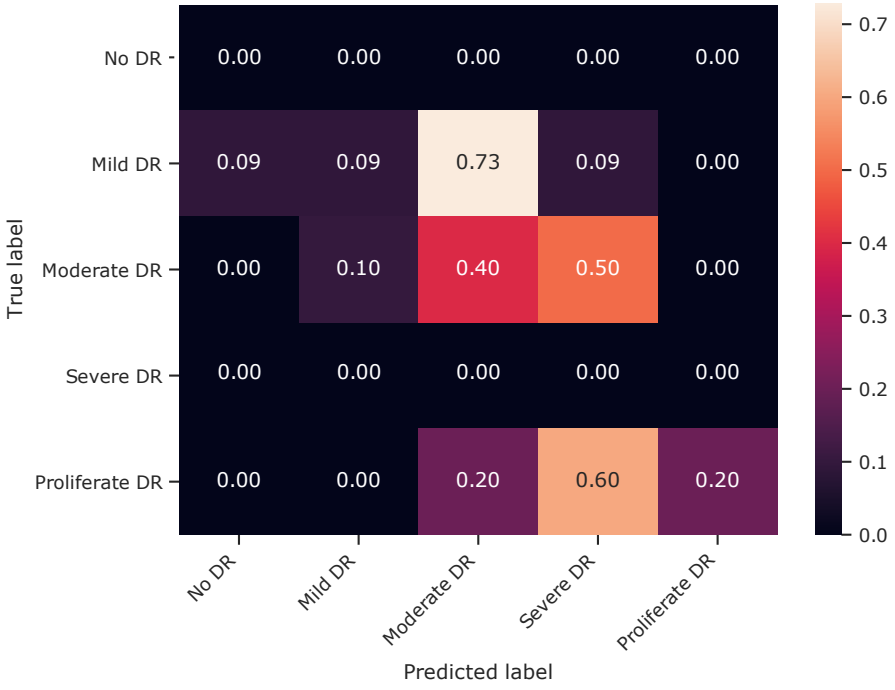


Figure 7.15: Normalised confusion matrix of the final ensemble model on the VFN dataset.

Chapter 8

Discussion

This chapter comments on the results and methods.

Section 8.1 summarizes the contribution of different baseline methods.

Section 8.2 reflects on the proposed preprocessing algorithm, discusses the influence of hyperparameters and state of the art architectures.

In Section 8.3, the author makes remarks on the algorithms, that use the blending of both eyes of the same person.

Section 8.4 comments on the use of unlabeled data.

Section 8.5 reports the observations made during the development of the vessel segmentation approach.

Section 8.6 mentions the points observed while using the ensemble approach. It also provides an explanation of the model predictions.

Section 8.7 discusses the results in the related competitions and datasets.

8.1 Baseline solution

Both networks from the baseline solution achieved an average performance in terms of the QWK score in comparison to other participants of the Diabetic Retinopathy Detection Challenge 2015.

Cross entropy loss was not performing well in training, probably because of the class imbalance. With the use of Focal loss, models began to show a significant improvement.

The results of SGD with momentum did not lag behind the Adam optimiser. However, a more cautious approach was needed to find the appropriate

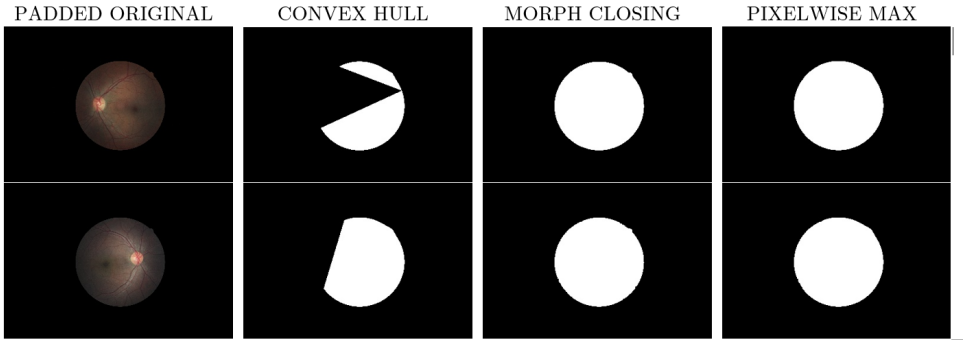


Figure 8.1: Cases, where the convex hull did not identify the retina circle correctly but morphological closing did.

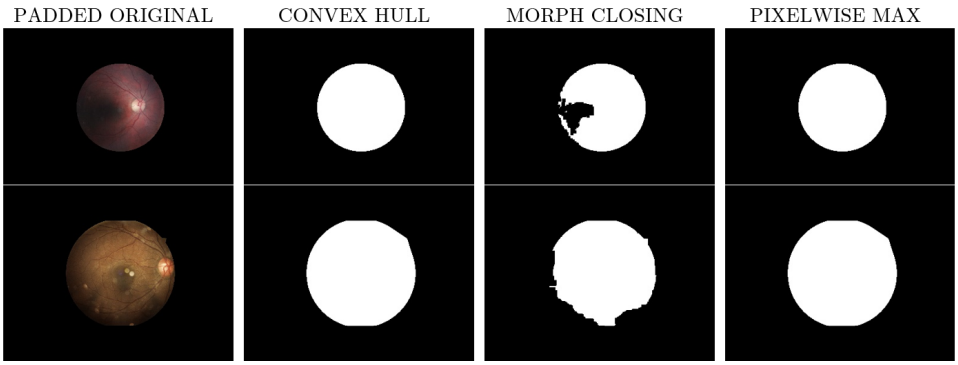


Figure 8.2: Cases, where the convex hull did identify the retina circle correctly but morphological closing did not.

learning rate and adjust it correctly during the training. Therefore, the Adam optimiser was the preferred choice.

8.2 General improvements

8.2.1 Improved preprocessing

The improved preprocessing generally focused on improving the detection of the retina circle using two separate detection branches. The Figures 8.1 and 8.2 show the situations, where the detection in one of the branches failed, however, the second branch was successful, resulting in the correct retina circle. This might be the key difference in the improved model training performance.

Figure 8.3: Result of the LR finder tool for the SeResNetX50 architecture.

8.2.2 Image sizes

Only two image sizes (299x299px and 512x512px) were among the ones used by the final models. The performance of size 299x299px was only slightly worse than in the case of the 512x512px, but the training time reduction was dramatic. Other image sizes, such as 1024x1024px were also tested, but the performance was not significantly better.

8.2.3 LR finder

This tool was essential for the successful training of the models. For example, SeResNetX50 was trained with cosine annealing with the initial learning rate of $2 \cdot 10^{-3}$ and then with the same setting but with the initial learning rate of $2 \cdot 10^{-5}$. This change, induced by the results from the LR finder (see Figure 8.3), improved the best QWK score from 0.55 to 0.819.

8.2.4 Classification and regression approach

Classification

The advantage of the classification models was the class probability output of the network because it showed more detailed insight into the network. This insight was then used in the blending and ensembling. On the other hand, the QWK score of the classification models was generally lower than the score of the regression models. This fact could be caused by the missing inter-class relationship. When this relationship was injected into the classification networks by using cost-sensitive regularisation, the results did not improve.

Regression

Among regression models, the EfficientNetB3 showed the best overall performance and was therefore selected for various experiments. The L2 and L1Smooth losses achieved the best performance in most cases, but each network was slightly different and required testing.

Generalized mean pooling

The influence of this method was inconsistent. In some cases, it led to a slight increase of performance, but in most cases, it models with GeM behaved the same way as without it.

8.2.5 Batch size

As is generally known, the batch size can influence the learning of the model. In this thesis, the influence of different batch sizes was also observed.

Among the classification models that used the advanced loss, a minimum batch size of 8 was required to enable training. That is because of the terms in the loss that estimate the distribution of the classes in a batch. This fact was limiting when training on the boruvka GPU server. With larger batch sizes, the training was possible, and the results were similar to each other (see Figure 8.5).

Among the regression networks, there was no limitation factor (apart from selecting very small batch sizes), and the batch size mostly influenced only the speed of convergence.

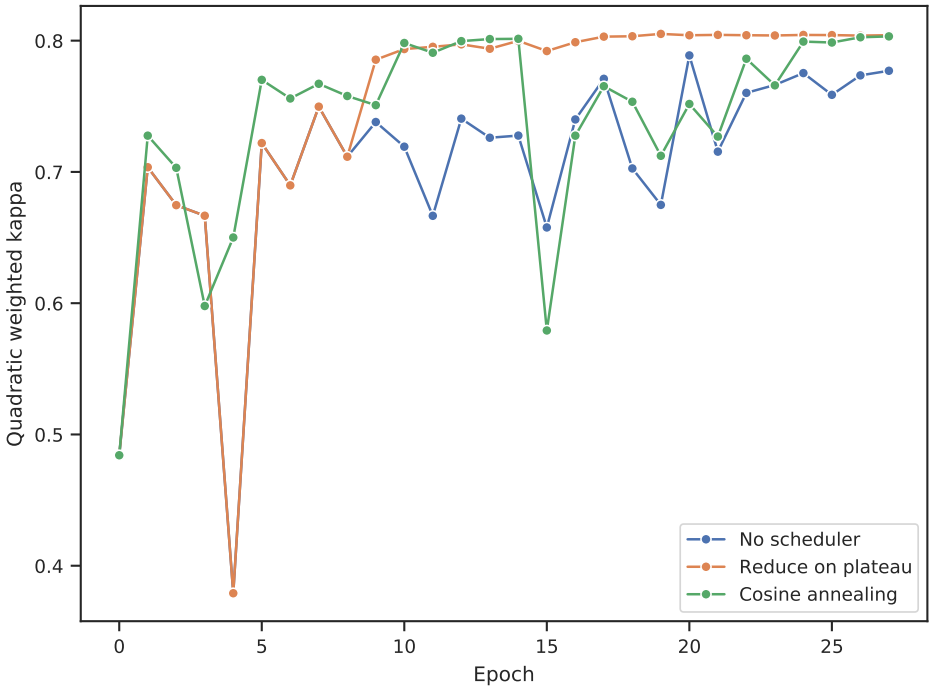


Figure 8.4: Influence of different schedulers on training a network (EfficientNet B3).

8.2.6 Learning rate schedulers

Influence of the schedulers was overall very positive. In the end, both Reduce on plateau and Cosine annealing achieved similar results (see Figure 8.4 for a representative example with the EfficientNetB3 model). However, cosine annealing induced faster convergence and in some cases led to a slightly improved performance.

8.3 Blending of the two eyes

Both algorithms improved the results to some extend. The winner takes it all was a very primitive method, and its benefits were more significant with worse-performing models (around QWK 0.60).

8.4 Using unlabeled data

The results of this method were not as convincing as expected. The cause lies in the number of parameters involved that needs to be correctly set. The

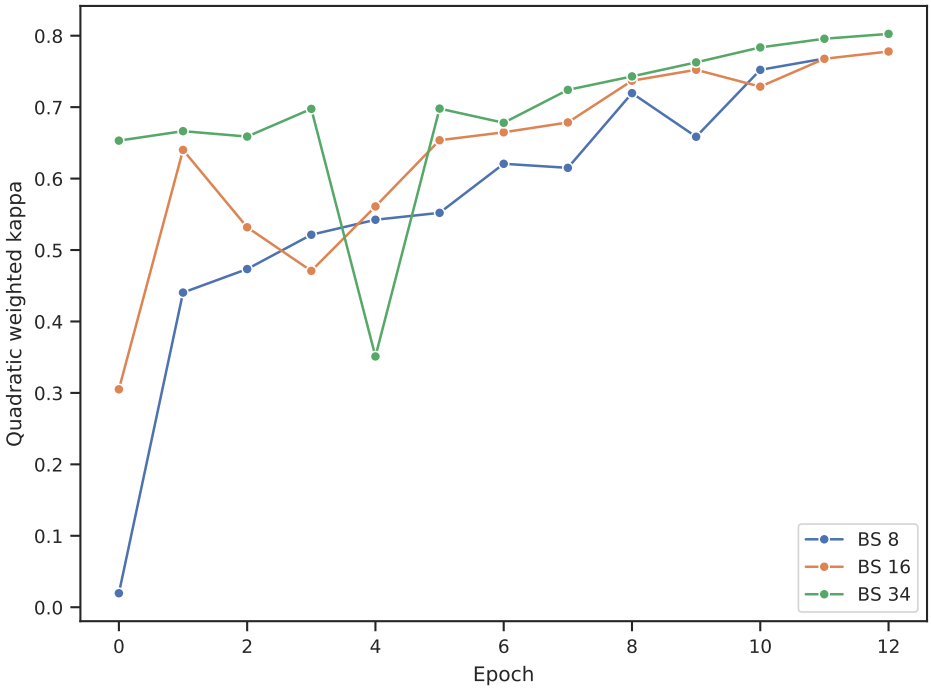


Figure 8.5: Influence of batch size on training of classification network (EfficientNet B3).

author of this thesis was not able to find suitable values. However, many participants from the APTOS 2019 Challenge reported, that if properly tuned, the performance increase should be visible.

8.5 Vessel segmentation

Vessel segmentation was a challenging task itself. The creation of the custom segmentation dataset was difficult as the images were available in different resolutions. There was a high level of noise in the segmentation ground-truth masks. However, the segmentation models were able to successfully eliminate the noise and the results were overall satisfactory.

As with the use of 4 channel classification, the results show that the networks performed better when trained on a larger (pseudolabelled) dataset. In this case, the models achieved the best performance of a single model.

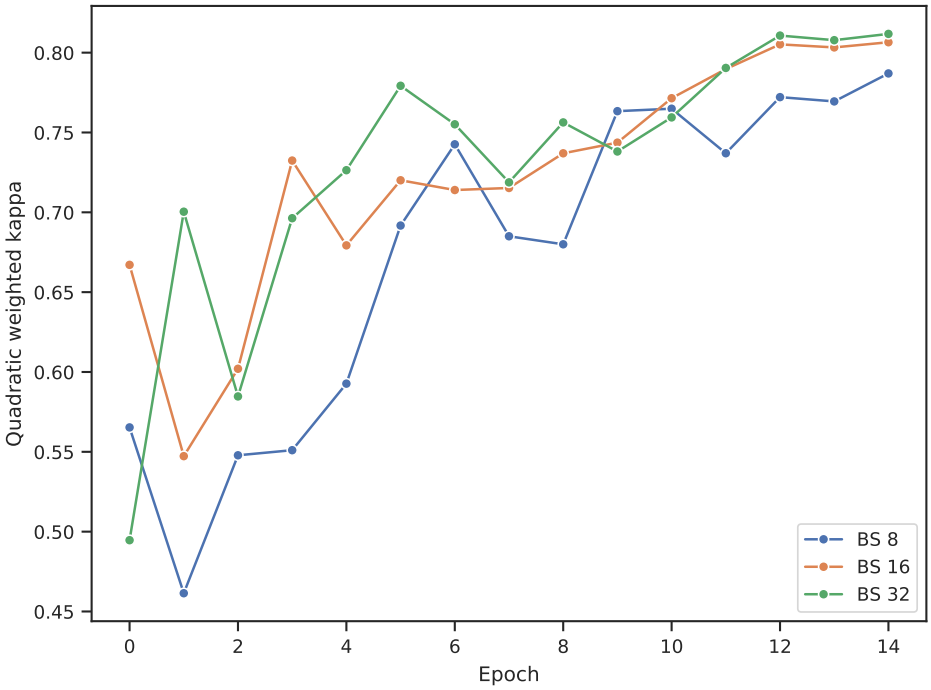


Figure 8.6: Influence of batch size on training of regression network (EfficientNet B3).

8.6 Ensemble classification

The results show that combining the blending of the two eyes and the predictions of multiple networks leads to the highest QWK scores.

As for the architecture, a very shallow network is sufficient for the ensemble task. If more layers were introduced, the performance was lower.

The results of the SHAP show that the ensemble model learnt to predict the labels using almost solely the regression networks. This fact could be used to replace the classification networks in the ensemble by other networks, possibly the 4 channel regression networks.

8.7 Analysis of the results

8.7.1 Diabetic Retinopathy Challenge - 2015

The results of the final ensemble model were analysed in detail to explore the most important misclassified examples. The medical opinion of prof.

Heissigerova is attached in the image description.

One of the most frequent groups of misclassifications was the group of images that were predicted to contain No DR, however the true labels stated, that these images presented the Mild DR stage (see Figure 8.7).

Another frequent group was the group of images predicted to belong to the Mild DR class, however according to the official labels they belong to the No DR class (see Figure 8.8).

A significant group of images belonging officially to the Severe DR were misclassified into No DR class (see Figure 8.9).

The next group is a group of images predicted to belong to the 0 (No DR class), but their true labels were (Proliferate DR). Even though this group of misclassifications is small in size, it is interesting to look at some of the examples (see Figure 8.10).

These things considered, the final proposed model has its limitations and flaws, but the quality of the EyePACS dataset is also questionable.

It is also worth mentioning that achieving the first place in such competition should not be overrated, as the pace of progress in deep learning is considerable and many of the proposed methods did not exist at the time of the competition.

8.7.2 APTOS 2019 Blindness Detection

The final result of the ensemble is promising, however, it cannot be compared directly with the other participants. Most of the errors arise from predicting the Moderate DR class when the true label is Mild DR.

8.7.3 ISBI - The 2nd Diabetic Retinopathy - Grading and Image Quality Estimation Challenge - 2020

The final result of the ensemble network could be influenced by the different format of the photos.

8.7.4 VFN

The final result is poor in comparison to other datasets. However, the dataset is very small in size, and the result might suffer from different methodology of the DR classification used in the Czech republic.

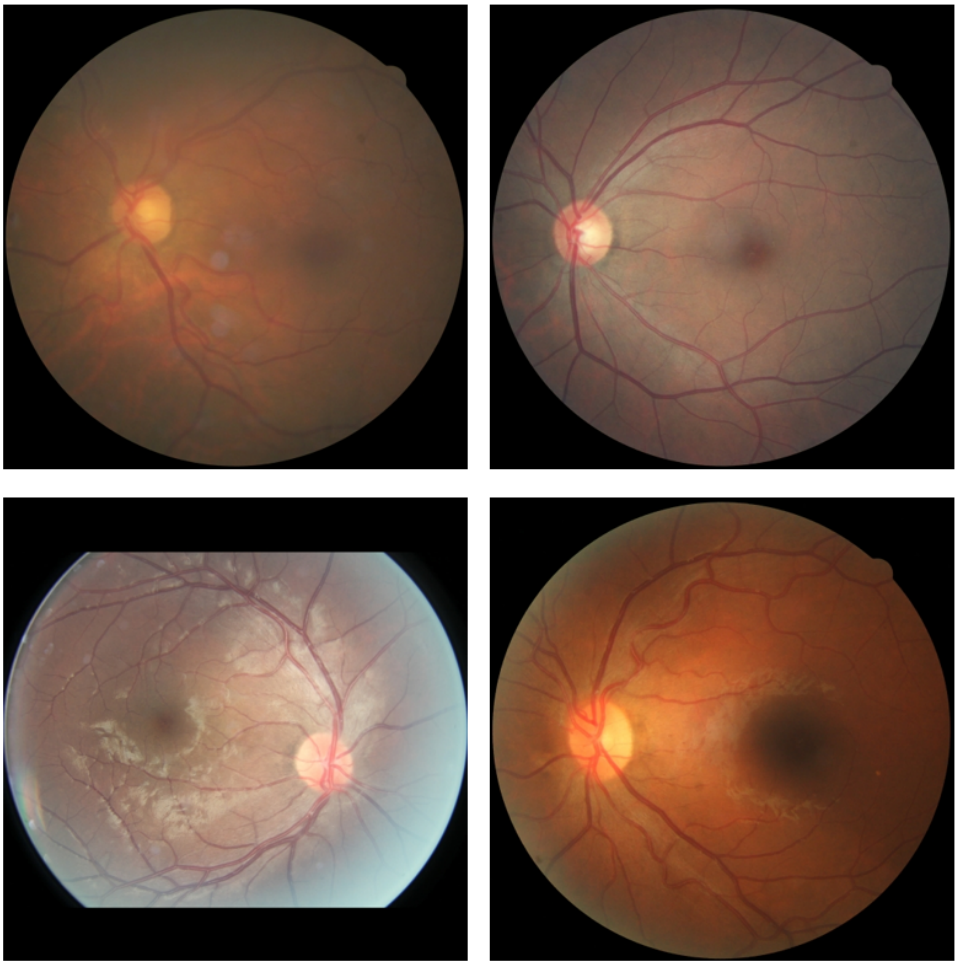


Figure 8.7: Images, where the model predicted 0 (No DR), but the true label is 1 (Mild DR). According to the prof. Heissigerova: The upper left picture contains artefact and belongs to the No DR class. The upper right image is normal fundus photography and belongs to the No DR class. The lower left image is either image of a child retina or an image of the retina after surgery. It could be considered Mild DR. The lower right image contains artefacts and belongs to the No DR class.

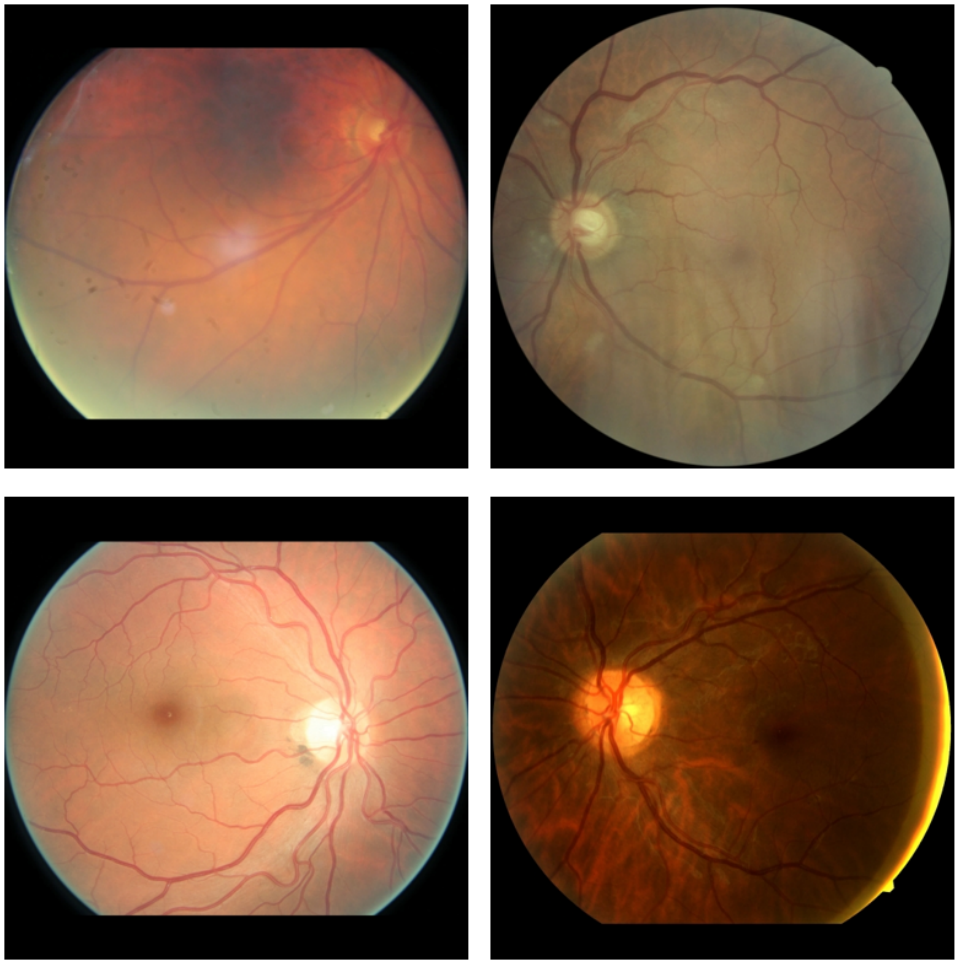


Figure 8.8: Images, where the model predicted 1 (Mild DR) but the true label is 0 (No DR). According to the prof. Heissigerova: The upper left image contains artefact (dust on the lens of the camera) and belongs to the No DR class. The upper right image belongs to the No DR. The lower two images are normal fundus images and belong to the No DR class.

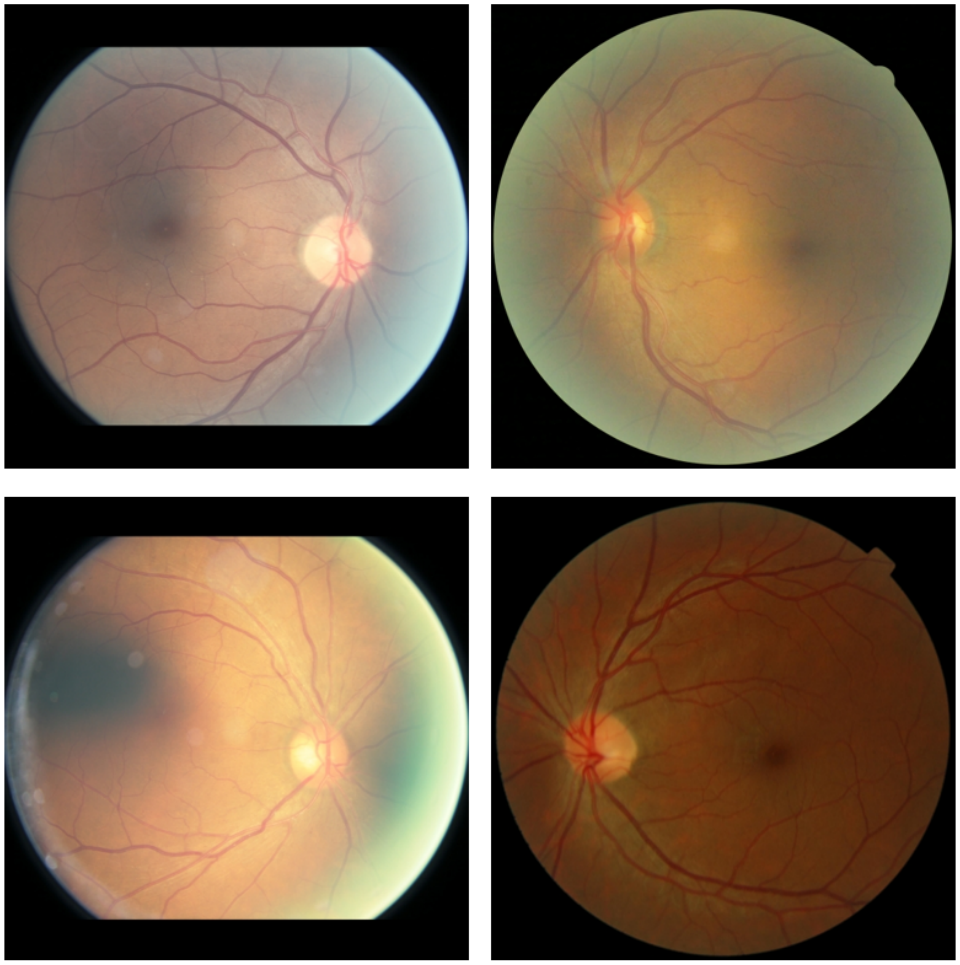


Figure 8.9: Images, where the model predicted 0 (No DR), but the true label is 2 (Severe DR). According to the prof. Heissigerova: The upper left picture belongs to the No DR class. The upper right image contains artefact but also belongs to the No DR class. The lower left image contains either an artefact or a pigment spot. The image belongs to the No DR class. The lower right image also presents no signs of DR.

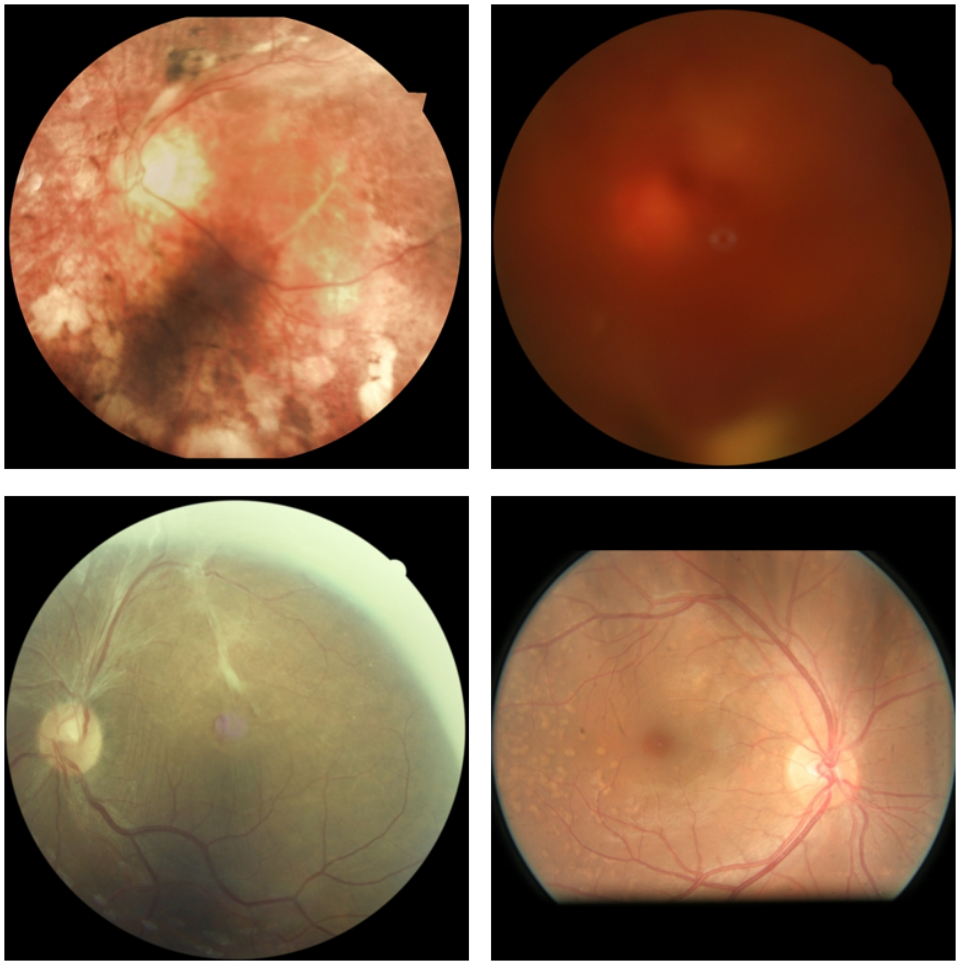


Figure 8.10: Images, where the model predicted 0 (No DR), but the true label is 4 (Proliferate DR). According to the prof. Heissigerova: The upper left image presents degenerative myopia, not DR. Therefore, it should belong to the No DR class. The upper right image presents vitreous haemorrhage and could be considered Proliferate DR. The lower left image belongs most probably to the No DR class. The lower right image presents signs of retinal degeneration, however, it is not caused by diabetic retinopathy.

Chapter 9

Conclusion

This thesis has developed a solution based on the convolutional neural networks for the purpose of diabetic retinopathy detection, as defined in the Diabetic Retinopathy Detection 2015 Competition.

The best performing baseline architecture consisted of DenseNet network and achieved the QWK score of 0.5960.

Then this thesis improved the solution by creating a custom preprocessing and adopting state of the art deep learning networks and methods. The best performing standalone model was regression-based EfficientNet B3, which achieved a QWK score of 0.821.

The results were improved by using the correlation from both eyes of the same person. Two methods were implemented, with the best improvement achieved by the random forest classifier (from QWK 0.796 to QWK 0.818).

Another proposed improvement used the training on pseudolabels. This approach was advantageous with the regression networks. The best score (QWK 0.831) achieved was using the EfficientNet B3 model.

This thesis also proposed to take advantage of the additional input, the use of vessel segmentation masks. Ensamble of three segmentation networks (InceptionV4, EfficientNet B3, ResNext101) provided the segmentation masks. When using the RGB images with the additional input and training on the training dataset, the DR detection did not improve sufficiently. However, when the additional input strategy was combined with the pseudolabels, the QWK score improved significantly. The final score (QWK 0.837) was achieved with the EfficientNet B3 network trained on the pseudolabels.

The final improvement arised from ensembling multiple classification and regression networks by a specialised neural network. The final score achieved after ensembling 9 models was QWK 0.850.

The results from the final ensemble model were submitted into the Dia-

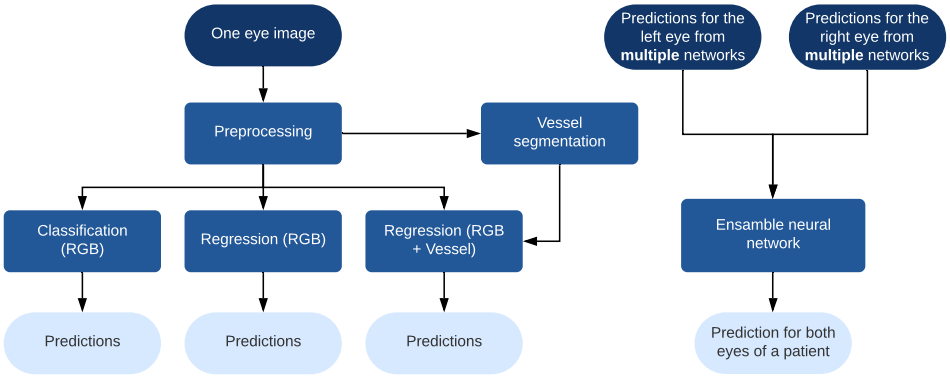


Figure 9.1: Multiple prediction on a single image (left). Schematics of the proposed ensemble method (right).

betic Retinopathy Detection 2015 Competition, achieving a state of the art performance and first place. In other related competitions, the results were compared to some extend.

The author recommends future work to focus on improving the quality of the pseudolabeling approach and pursue the combined use of additional input and pseudolabels.

The models selected in the ensemble could be replaced by better performing models, and the overall architecture would benefit from including the 4-channel input models into the ensemble (see Figure 9.1).

To make the solution serviceable in the real world, the author suggests creating a cloud-based application. Such application would provide the medical staff with a website interface and could be used in the department of diabetology, where a trained ophthalmologist is not available.

Bibliography

[1] L. Guariguata, D. Whiting, I. Hambleton, J. Beagley, U. Linnenkamp, and J. Shaw, "Global estimates of diabetes prevalence for 2013 and projections for 2035," *Diabetes Research and Clinical Practice*, vol. 103, no. 2, pp. 137–149, feb 2014. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0168822713003859

[2] E. J. Duh, J. K. Sun, and A. W. Stitt, "R E V I E W Diabetic retinopathy: current understanding, mechanisms, and treatment strategies The clinical challenge of diabetic retinopathy," *JCI Insight*, 2017. [Online]. Available: https://doi.org/10.1172/jci.insight.93751

[3] V. Bartoš and T. Pelikánová, *Praktická diabetologie*. Maxdorf, 2003.

[4] J. Rybka, *Diabetes mellitus - komplikace a přidružená onemocnění: diagnostické a léčebné postupy*. Grada, 2007.

[5] B. D. Kels, A. Grzybowski, and J. M. Grant-Kels, "Human ocular anatomy," *Clinics in Dermatology*, vol. 33, no. 2, pp. 140–146, mar 2015. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0738081X1400234X

[6] J. Heissigerová, *Oftalmologie. Pro pregraduální i postgraduální přípravu.* Maxdorf, 2018.

[7] "Eye anatomy image." [Online]. Available: https://www.genes-vision.ch/retinalearn/eye-anatomy/ (Accessed 12.8.2020).

[8] A. M. Hendrick, M. V. Gibson, and A. Kulshreshtha, "Diabetic Retinopathy," *Primary Care: Clinics in Office Practice*, vol. 42, no. 3, pp. 451–464, sep 2015. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S009545431500038X

[9] University of Michigan and Keylog Eye Center, "Anatomy of the Eye." [Online]. Available: https://www.umkelloggeye.org/conditions-treatments/anatomy-eye (Accessed 2020-08-12).

[10] P. Kuchynka, *Oční lékařství*. Grada, 2016.

[11] Georgia State University, "Rods and Cones." [Online]. Available: http://hyperphysics.phy-astr.gsu.edu/hbase/vision/rodcone.html (Accessed 2020-08-12).

[12] C. A. Curcio, K. R. Sloan, R. E. Kalina, and A. E. Hendrickson, "Human photoreceptor topography," *The Journal of Comparative Neurology*, vol. 292, no. 4, pp. 497–523, feb 1990. [Online]. Available: http://doi.wiley.com/10.1002/cne.902920402

[13] E. Rencová, "Diabetická oftalmopatie," *Vnitřní lékařství*, 2006.

[14] D. A. Antonetti, R. Klein, and T. W. Gardner, "Diabetic Retinopathy," *New England Journal of Medicine*, vol. 366, no. 13, pp. 1227–1239, mar 2012. [Online]. Available: http://www.nejm.org/doi/10.1056/NEJMra1005073

[15] L. Valešová and J. Hycl, *Diabetická retinopatie*, 2002.

[16] The Institute of Health Information and Statistics of the Czech Republic, "Diabetická retinopatie." [Online]. Available: https://nsc.uzis.cz/zdraveoci/index.php?pg=diabeticka-retinopatie (Accessed 12.8.2020).

[17] C. Kirkpatrick, "Intraretinal Microvascular Abnormality (IrMA)." [Online]. Available: https://webeye.ophth.uiowa.edu/eyeforum/atlas/pages/IRMA.htm (Accessed 2020-08-12).

[18] "Early Treatment Diabetic Retinopathy Study Design and Baseline Patient Characteristics," *Ophthalmology*, vol. 98, no. 5, pp. 741–756, may 1991. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0161642013380099

[19] V. Hoang Minh and R. Sznitman, "Medical Imaging and Applications Visual Question Answering for Diabetic Retinopathy Screening," Ph.D. dissertation, 2018.

[20] "DR vision impairment image." [Online]. Available: http://www.eyedocsottawa.com/services-special-vision/special-vision-concerns/diabetes/ (Accessed 2020-08-12).

[21] H. Murgatroyd, "Effect of mydriasis and different field strategies on digital image screening of diabetic eye disease," *British Journal of Ophthalmology*, vol. 88, no. 7, pp. 920–924, jul 2004. [Online]. Available: http://bjo.bmj.com/cgi/doi/10.1136/bjo.2003.026385

[22] P. Veselý and P. Beneš, *Vyšetřovací metody v optometrii*. Grada, 2019.

[23] I. E. Mohamed, M. A. Mohamed, M. Yousef, M. Z. Mahmoud, and B. Alonazi, "Use of ophthalmic B-scan ultrasonography in determining

the causes of low vision in patients with diabetic retinopathy," *European Journal of Radiology Open*, vol. 5, pp. 79–86, 2018. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S2352047718300091

[24] E. B. Schroeder, R. Hanratty, B. L. Beaty, E. A. Bayliss, E. P. Havranek, and J. F. Steiner, "Simultaneous Control of Diabetes Mellitus, Hypertension, and Hyperlipidemia in 2 Health Systems," *Circulation: Cardiovascular Quality and Outcomes*, vol. 5, no. 5, pp. 645–653, sep 2012. [Online]. Available: https://www.ahajournals.org/doi/10.1161/CIRCOUTCOMES.111.963553

[25] M. D. José Pedro De La Cruz, M. D. José Antonio González-Correa, M. D. Ana Guerrero, and M. D. Felipe Sánchez de la Cuesta, "Pharmacological approach to diabetic retinopathy," *Diabetes/Metabolism Research and Reviews*, vol. 20, no. 2, pp. 91–113, mar 2004. [Online]. Available: http://doi.wiley.com/10.1002/dmrr.432

[26] J. G. F. Dowler, "Laser management of diabetic retinopathy," *JRSM*, vol. 96, no. 6, pp. 277–279, jun 2003. [Online]. Available: http://jrsm.rsmjournals.com/cgi/doi/10.1258/jrsm.96.6.277

[27] G. Rao, "Physicians battle India's diabetic retinopathy crisis," 2007. [Online]. Available: https://www.healio.com/news/ophthalmology/20120325/physicians-battle-india-s-diabetic-retinopathy-crisis (Accessed 2020-08-12).

[28] W. M. D. W. Zaki, M. A. Zulkifley, A. Hussain, W. H. W. Halim, N. B. A. Mustafa, and L. S. Ting, "Diabetic retinopathy assessment: Towards an automated system," *Biomedical Signal Processing and Control*, vol. 24, pp. 72–82, feb 2016. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S1746809415001652

[29] A. D. Fleming, K. A. Goatman, S. Philip, G. J. Prescott, P. F. Sharp, and J. A. Olson, "Automated grading for diabetic retinopathy: a large-scale audit using arbitration by clinical experts," *British Journal of Ophthalmology*, vol. 94, no. 12, pp. 1606–1610, dec 2010. [Online]. Available: http://bjo.bmj.com/cgi/doi/10.1136/bjo.2009.176784

[30] V. Gulshan, R. P. Rajan, K. Widner, D. Wu, P. Wubbels, T. Rhodes, K. Whitehouse, M. Coram, G. Corrado, K. Ramasamy, R. Raman, L. Peng, and D. R. Webster, "Performance of a Deep-Learning Algorithm vs Manual Grading for Detecting Diabetic Retinopathy in India," *JAMA Ophthalmology*, vol. 137, no. 9, p. 987, sep 2019. [Online]. Available: https://jamanetwork.com/journals/jamaophthalmology/fullarticle/2734990

[31] "Kaggle." [Online]. Available: https://www.kaggle.com/c/about/host/ (Accessed 2020-08-12).

[32] "Diabetic Retinopathy Detection competition," 2015. [Online]. Available: https://www.kaggle.com/c/diabetic-retinopathy-detection/ (Accessed 2020-08-12).

[33] J. Cuadros and G. Bresnick, "EyePACS: An Adaptable Telemedicine System for Diabetic Retinopathy Screening," *Journal of Diabetes Science and Technology*, vol. 3, no. 3, pp. 509–516, 2009. [Online]. Available: https://doi.org/10.1177/193229680900300315

[34] "EyePACS Dataset," 2015. [Online]. Available: https://www.kaggle.com/c/diabetic-retinopathy-detection/data (Accessed 2020-08-12).

[35] S. M. S. Islam, M. M. Hasan, and S. Abdullah, "Deep Learning based Early Detection and Grading of Diabetic Retinopathy Using Retinal Fundus Images," dec 2018. [Online]. Available: http://arxiv.org/abs/1812.10595

[36] Z. Gao, J. Li, J. Guo, Y. Chen, Z. Yi, and J. Zhong, "Diagnosis of Diabetic Retinopathy Using Deep Neural Networks," *IEEE Access*, vol. 7, pp. 3360–3370, 2019. [Online]. Available: https://ieeexplore.ieee.org/document/8581492/

[37] "APTOS Dataset," 2019. [Online]. Available: https://www.kaggle.com/c/aptos2019-blindness-detection/data (Accessed 2020-08-12).

[38] "APTOS competition," 2019. [Online]. Available: https://www.kaggle.com/c/aptos2019-blindness-detection/overview (Accessed 2020-08-12).

[39] A. Buslaev, V. I. Iglovikov, E. Khvedchenya, A. Parinov, M. Druzhinin, and A. A. Kalinin, "Albumentations: Fast and Flexible Image Augmentations," *Information*, vol. 11, no. 2, p. 125, feb 2020. [Online]. Available: https://www.mdpi.com/2078-2489/11/2/125

[40] "ISBI Dataset," 2020. [Online]. Available: https://isbi.deepdr.org/data.html (Accessed 2020-08-12).

[41] "ISBI 3rd place solution source code." [Online]. Available: https://bitbucket.org/woalsdnd/deepdrid/src/master/ (Accessed 2020-08-12).

[42] G. G. Gardner, D. Keating, T. H. Williamson, and A. T. Elliott, "Automatic detection of diabetic retinopathy using an artificial neural network: a screening tool." *British Journal of Ophthalmology*, vol. 80, no. 11, pp. 940–944, nov 1996. [Online]. Available: http://bjo.bmj.com/cgi/doi/10.1136/bjo.80.11.940

[43] B. M. Ege, O. K. Hejlesen, O. V. Larsen, K. Møller, B. Jennings, D. Kerr, and D. A. Cavan, "Screening for diabetic retinopathy using computer based image analysis and statistical classification," *Computer Methods and Programs in Biomedicine*, vol. 62, no. 3, pp. 165–175, jul

2000. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0169260700000651

[44] R. Acharya U, C. K. Chua, E. Y. K. Ng, W. Yu, and C. Chee, "Application of Higher Order Spectra for the Identification of Diabetes Retinopathy Stages," *Journal of Medical Systems*, vol. 32, no. 6, pp. 481–488, dec 2008. [Online]. Available: http://link.springer.com/10.1007/s10916-008-9154-8

[45] P. Adarsh and D. Jeyakumari, "Multiclass SVM-based automated diagnosis of diabetic retinopathy," in *2013 International Conference on Communication and Signal Processing*. IEEE, apr 2013, pp. 206–210. [Online]. Available: http://ieeexplore.ieee.org/document/6577044/

[46] S. Roychowdhury, D. D. Koozekanani, and K. K. Parhi, "DREAM: Diabetic Retinopathy Analysis Using Machine Learning," *IEEE Journal of Biomedical and Health Informatics*, vol. 18, no. 5, pp. 1717–1728, sep 2014. [Online]. Available: http://ieeexplore.ieee.org/document/6680633/

[47] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," *Communications of the ACM*, vol. 60, no. 6, pp. 84–90, may 2017. [Online]. Available: https://dl.acm.org/doi/10.1145/3065386

[48] H. Pratt, F. Coenen, D. M. Broadbent, S. P. Harding, and Y. Zheng, "Convolutional Neural Networks for Diabetic Retinopathy," *Procedia Computer Science*, vol. 90, pp. 200–205, 2016. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S1877050916311929

[49] V. Gulshan, R. P. Rajan, K. Widner, D. Wu, P. Wubbels, T. Rhodes, K. Whitehouse, M. Coram, G. Corrado, K. Ramasamy, R. Raman, L. Peng, and D. R. Webster, "Performance of a Deep-Learning Algorithm vs Manual Grading for Detecting Diabetic Retinopathy in India," *JAMA Ophthalmology*, vol. 137, no. 9, p. 987, sep 2019. [Online]. Available: https://jamanetwork.com/journals/jamaophthalmology/fullarticle/2734990

[50] M. Voets, K. Møllersen, and L. A. Bongo, "Replication study: Development and validation of deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs," mar 2018. [Online]. Available: http://arxiv.org/abs/1803.04337http://dx.doi.org/10.1371/journal.pone.0217541

[51] V. Gulshan, L. Peng, M. Coram, M. C. Stumpe, D. Wu, A. Narayanaswamy, S. Venugopalan, K. Widner, T. Madams, J. Cuadros, R. Kim, R. Raman, P. C. Nelson, J. L. Mega, and D. R. Webster, "Development and Validation of a Deep Learning Algorithm for Detection of Diabetic Retinopathy in Retinal Fundus Photographs,"

JAMA, vol. 316, no. 22, p. 2402, dec 2016. [Online]. Available: http://jama.jamanetwork.com/article.aspx?doi=10.1001/jama.2016.17216

[52] Messidor Consortium, "Messidor2 Dataset." [Online]. Available: http://www.adcis.net/en/third-party/messidor2/ (Accessed 2020-08-12).

[53] S. Ruder, "An overview of gradient descent optimization algorithms," sep 2016. [Online]. Available: http://arxiv.org/abs/1609.04747

[54] Y.-H. Li, N.-N. Yeh, S.-J. Chen, and Y.-C. Chung, "Computer-Assisted Diagnosis for Diabetic Retinopathy Based on Fundus Images Using Deep Convolutional Neural Network," *Mobile Information Systems*, vol. 2019, pp. 1–14, jan 2019. [Online]. Available: https://www.hindawi.com/journals/misy/2019/6142839/

[55] R. Sarki, S. Michalska, K. Ahmed, H. Wang, and Y. Zhang, "Convolutional neural networks for mild diabetic retinopathy detection: an experimental study," *bioRxiv*, 2019. [Online]. Available: https://www.biorxiv.org/content/early/2019/09/09/763136

[56] J. Sahlsten, J. Jaskari, J. Kivinen, L. Turunen, E. Jaanio, K. Hietala, and K. Kaski, "Deep Learning Fundus Image Analysis for Diabetic Retinopathy and Macular Edema Grading," *Scientific Reports*, vol. 9, no. 1, p. 10750, dec 2019. [Online]. Available: http://www.nature.com/articles/s41598-019-47181-w

[57] C. González-Gonzalo, V. Sánchez-Gutiérrez, P. Hernández-Martínez, I. Contreras, Y. T. Lechanteur, A. Domanian, B. Ginneken, and C. I. Sánchez, "Evaluation of a deep learning system for the joint automated detection of diabetic retinopathy and age-related macular degeneration," *Acta Ophthalmologica*, vol. 98, no. 4, pp. 368–377, jun 2020. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1111/aos.14306

[58] D. J. Hemanth, O. Deperlioglu, and U. Kose, "An enhanced diabetic retinopathy detection and classification approach using deep convolutional neural network," *Neural Computing and Applications*, vol. 32, no. 3, pp. 707–721, feb 2018. [Online]. Available: http://link.springer.com/10.1007/s00521-018-03974-0

[59] Messidor Consortium, "Messidor Dataset." [Online]. Available: http://www.adcis.net/en/third-party/messidor/ (Accessed 2020-08-12).

[60] K. Zuiderveld, *Contrast Limited Adaptive Histogram Equalization*. USA: Academic Press Professional, Inc., 1994, pp. 474–485.

[61] B. Tymchenko, P. Marchenko, and D. Spodarets, "Deep Learning Approach to Diabetic Retinopathy Detection," mar 2020. [Online]. Available: http://arxiv.org/abs/2003.02261

[62] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal Loss for Dense Object Detection," aug 2017. [Online]. Available: http://arxiv.org/abs/1708.02002

[63] I. Loshchilov and F. Hutter, "SGDR: Stochastic Gradient Descent with Warm Restarts," aug 2016. [Online]. Available: http://arxiv.org/abs/1608.03983

[64] A. Krogh and J. A. Hertz, "A Simple Weight Decay Can Improve Generalization," in *Proceedings of the 4th International Conference on Neural Information Processing Systems*, ser. NIPS'91. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 1991, pp. 950–957.

[65] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the Inception Architecture for Computer Vision," dec 2015. [Online]. Available: http://arxiv.org/abs/1512.00567

[66] M. Tan and Q. V. Le, "EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks," may 2019. [Online]. Available: http://arxiv.org/abs/1905.11946

[67] J. Hu, L. Shen, S. Albanie, G. Sun, and E. Wu, "Squeeze-and-Excitation Networks," sep 2017. [Online]. Available: http://arxiv.org/abs/1709.01507

[68] A. Shah, W. Clarida, R. Amelon, M. C. Hernaez-Ortega, A. Navea, J. Morales-Olivas, R. Dolz-Marco, F. Verbraak, P. P. Jorda, A. A. van der Heijden, and C. Peris Martinez, "Validation of Automated Screening for Referable Diabetic Retinopathy With an Autonomous Diagnostic Artificial Intelligence System in a Spanish Population," *Journal of Diabetes Science and Technology*, p. 193229682090621, mar 2020. [Online]. Available: http://journals.sagepub.com/doi/10.1177/1932296820906212

[69] R. Sayres, A. Taly, E. Rahimy, K. Blumer, D. Coz, N. Hammel, J. Krause, A. Narayanaswamy, Z. Rastegar, D. Wu, S. Xu, S. Barb, A. Joseph, M. Shumski, J. Smith, A. B. Sood, G. S. Corrado, L. Peng, and D. R. Webster, "Using a Deep Learning Algorithm and Integrated Gradients Explanation to Assist Grading for Diabetic Retinopathy," *Ophthalmology*, vol. 126, no. 4, pp. 552–564, apr 2019. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0161642018315756

[70] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. Alemi, "Inception-v4, Inception-ResNet and the Impact of Residual Connections on Learning," feb 2016. [Online]. Available: http://arxiv.org/abs/1602.07261

[71] V. Gulshan, R. P. Rajan, K. Widner, D. Wu, P. Wubbels, T. Rhodes, K. Whitehouse, M. Coram, G. Corrado, K. Ramasamy, R. Raman, L. Peng, and D. R. Webster, "Performance of a Deep-Learning Algorithm vs Manual Grading for Detecting

Diabetic Retinopathy in India," *JAMA Ophthalmology*, vol. 137, no. 9, p. 987, sep 2019. [Online]. Available: https://jamanetwork.com/journals/jamaophthalmology/fullarticle/2734990

[72] Z. Straka, P. Švarný, and J. Kostlivá, "Classifiers," CTU FEE, Prague, Tech. Rep. [Online]. Available: https://cw.fel.cvut.cz/wiki/{_}media/courses/be5b33kui/labs/weekly/classifiers.pdf

[73] S. Vanbelle, "A New Interpretation of the Weighted Kappa Coefficients," *Psychometrika*, vol. 81, no. 2, pp. 399–410, jun 2016. [Online]. Available: http://link.springer.com/10.1007/s11336-014-9439-4

[74] J. Cohen, "A Coefficient of Agreement for Nominal Scales," *Educational and Psychological Measurement*, vol. 20, no. 1, pp. 37–46, apr 1960. [Online]. Available: http://journals.sagepub.com/doi/10.1177/001316446002000104

[75] H. C. Kraemer, "Extension of the Kappa Coefficient," *Biometrics*, vol. 36, no. 2, p. 207, jun 1980. [Online]. Available: https://www.jstor.org/stable/2529972?origin=crossref

[76] J. L. Fleiss and J. Cohen, "The Equivalence of Weighted Kappa and the Intraclass Correlation Coefficient as Measures of Reliability," *Educational and Psychological Measurement*, vol. 33, no. 3, pp. 613–619, oct 1973. [Online]. Available: http://journals.sagepub.com/doi/10.1177/001316447303300309

[77] D. Ashby, "Practical statistics for medical research. Douglas G. Altman, Chapman and Hall, London, 1991. No. of pages: 611. Price: £32.00," *Statistics in Medicine*, vol. 10, no. 10, pp. 1635–1636, oct 1991. [Online]. Available: http://doi.wiley.com/10.1002/sim.4780101015

[78] K. P. Murphy, *Machine Learning: A Probabilistic Perspective*. The MIT Press, 2012.

[79] "Cross validation image." [Online]. Available: https://scikit-learn.org/stable/modules/cross{_}validation.html (Accessed 12.8.2020).

[80] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. The MIT Press, 2016.

[81] J. Ludwig, "Image Convolution," 2020. [Online]. Available: http://web.pdx.edu/{~}jduh/courses/Archive/geog481w07/Students/Ludwig{_}ImageConvolution.pdf (Accessed 2020-08-12).

[82] "Convolution image." [Online]. Available: https://www.researchgate.net/figure/A-valid-convolution-of-a-5x5-image-with-a-3x3-kernel-The-kernel-will-be-applied-to{_}fig5 (Accessed 12.8.2020).

[83] G. Huang, Z. Liu, L. van der Maaten, and K. Q. Weinberger, "Densely Connected Convolutional Networks," aug 2016. [Online]. Available: http://arxiv.org/abs/1608.06993

[84] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," dec 2015. [Online]. Available: http://arxiv.org/abs/1512.03385

[85] "Inception scheme image." [Online]. Available: https://www.researchgate.net/figure/Schematic-diagram-of-InceptionV3-model-compressed-view{_}fig6{_}326421398 (Accessed 12.8.2020).

[86] "DenseNet scheme image." [Online]. Available: https://www.researchgate.net/figure/Original-DenseNet-architecture-from-16-The-top-figure-depcits-an-example-dense{_}fig2{_}3368902 (Accessed 12.8.2020).

[87] "InceptionResnetV2 scheme image." [Online]. Available: https://www.kaggle.com/keras/inceptionresnetv2 (Accessed 2020-08-12).

[88] "SEblock image." [Online]. Available: https://towardsdatascience.com/squeeze-and-excitation-networks-9ef5e71eacd7 (Accessed 2020-08-12).

[89] F. Radenović, G. Tolias, and O. Chum, "Fine-tuning CNN Image Retrieval with No Human Annotation," nov 2017. [Online]. Available: http://arxiv.org/abs/1711.02512

[90] T. M. Cover and J. A. Thomas, *Elements of Information Theory*. Wiley, sep 2005. [Online]. Available: https://onlinelibrary.wiley.com/doi/book/10.1002/047174882X

[91] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal Loss for Dense Object Detection," aug 2017. [Online]. Available: http://arxiv.org/abs/1708.02002

[92] E. Arazo, D. Ortego, P. Albert, N. E. O'Connor, and K. McGuinness, "Pseudo-Labeling and Confirmation Bias in Deep Semi-Supervised Learning," aug 2019. [Online]. Available: http://arxiv.org/abs/1908.02983

[93] D. Tanaka, D. Ikami, T. Yamasaki, and K. Aizawa, "Joint Optimization Framework for Learning with Noisy Labels," mar 2018. [Online]. Available: http://arxiv.org/abs/1803.11364

[94] "QWK loss." [Online]. Available: https://www.kaggle.com/c/crowdflower-search-relevance/discussion/14706 (Accessed 2020-08-12).

[95] Z.-H. Feng, J. Kittler, M. Awais, P. Huber, and X.-J. Wu, "Wing Loss for Robust Facial Landmark Localisation with Convolutional Neural Networks," nov 2017. [Online]. Available: http://arxiv.org/abs/1711.06753

[96] "SGD algorithm image." [Online]. Available: https://prinsphield.github.io/posts/2016/02/overview{_}opt{_}alg{_}deep{_}learning1/ (Accessed 12.8.2020).

[97] D. P. Kingma and J. Ba, "Adam: A Method for Stochastic Optimization," dec 2014. [Online]. Available: http://arxiv.org/abs/1412.6980

[98] G. Hinton, "RMSprop." [Online]. Available: http://www.cs.toronto.edu/{~}tijmen/csc321/slides/lecture{_}slides{_}lec6.pdf (Accessed 2020-08-12).

[99] A. Kathuria, "Adam," 2018. [Online]. Available: https://blog.paperspace.com/intro-to-optimization-momentum-rmsprop-adam/ (Accessed 2020-08-12).

[100] A. C. Wilson, R. Roelofs, M. Stern, N. Srebro, and B. Recht, "The Marginal Value of Adaptive Gradient Methods in Machine Learning," may 2017. [Online]. Available: http://arxiv.org/abs/1705.08292

[101] I. Loshchilov and F. Hutter, "Decoupled Weight Decay Regularization," nov 2017. [Online]. Available: http://arxiv.org/abs/1711.05101

[102] "Reduce on plateau image." [Online]. Available: https://datascience.stackexchange.com/questions/23213/why-does-decreasing-the-sgd-learning-rate-cause-a-massive-increase-in-accuracy (Accessed 12.8.2020).

[103] I. Loshchilov and F. Hutter, "SGDR: Stochastic Gradient Descent with Warm Restarts," aug 2016. [Online]. Available: http://arxiv.org/abs/1608.03983

[104] H. Li, Z. Xu, G. Taylor, C. Studer, and T. Goldstein, "Visualizing the Loss Landscape of Neural Nets," dec 2017. [Online]. Available: http://arxiv.org/abs/1712.09913

[105] "Saddle point image." [Online]. Available: https://en.wikipedia.org/wiki/Saddle{%}5C{_}point (Accessed 12.8.2020).

[106] "Cosine annealing image." [Online]. Available: https://towardsdatascience.com/exploring-learning-{%}0Arates-to-improve-model-performance-in-keras-e37f5e63f16c (Accessed 12.8.2020).

[107] "U-Net scheme image." [Online]. Available: https://machinelearningmastery.com/how-to-implement-pix2pix-gan-models-from-scratch-with-keras/ (Accessed 12.8.2020).

[108] B. Baheti, S. Innani, S. Gajre, and S. Talbar, "Eff-UNet: A Novel Architecture for Semantic Segmentation in Unstructured Environment," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*. IEEE, jun 2020, pp. 1473–1481. [Online]. Available: https://ieeexplore.ieee.org/document/9150621/

[109] L. R. Dice, "Measures of the Amount of Ecologic Association Between Species," *Ecology*, vol. 26, no. 3, pp. 297–302, jul 1945. [Online]. Available: http://doi.wiley.com/10.2307/1932409

[110] G. Louppe, "Understanding Random Forests: From Theory to Practice," jul 2014. [Online]. Available: http://arxiv.org/abs/1407.7502

[111] DataCamp, "Machine Learning with PySpark." [Online]. Available: https://www.datacamp.com/courses/machine-learning-with-apache-spark (Accessed 2020-08-12).

[112] "Random forest image." [Online]. Available: https://medium.com/@ar.ingenious/applying-random-forest-classification-machine-learning-algorithm-from-scratch-with-real-24ff198a1c5 (Accessed 12.8.2020).

[113] S. Lundberg and S.-I. Lee, "A Unified Approach to Interpreting Model Predictions," may 2017. [Online]. Available: http://arxiv.org/abs/1705.07874

[114] "SHAP explanation." [Online]. Available: http://bjlkeng.github.io/posts/model-explanability-with-shapley-additive-explanations-shap/ (Accessed 2020-08-12).

[115] "PyTorch." [Online]. Available: https://pytorch.org/ (Accessed 2020-08-12).

[116] "Python." [Online]. Available: https://www.python.org/ (Accessed 2020-08-12).

[117] L. N. Smith, "Cyclical Learning Rates for Training Neural Networks," jun 2015. [Online]. Available: http://arxiv.org/abs/1506.01186

[118] "CMP Info." [Online]. Available: http://cmp.felk.cvut.cz/new{_}pages/ (Accessed 2020-08-12).

[119] M. Voets, K. Møllersen, and L. A. Bongo, "Reproduction study using public data of: Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs," *PLOS ONE*, vol. 14, no. 6, p. e0217541, jun 2019. [Online]. Available: https://dx.plos.org/10.1371/journal.pone.0217541

[120] S. Suzuki and K. Be, "Topological structural analysis of digitized binary images by border following," *Computer Vision, Graphics, and Image Processing*, vol. 30, no. 1, pp. 32–46, apr 1985. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/0734189X85900167

[121] "Minimum enclosing circle." [Online]. Available: https://docs.opencv.org/2.4/modules/imgproc/doc/structural{_}analysis{_}and{_}shape{_}descriptors.html?highlight=minenclosingcircle{#}minenclosingcircle (Accessed 2020-08-12).

[122] Ming-Kuei Hu, "Visual pattern recognition by moment invariants," *IEEE Transactions on Information Theory*, vol. 8, no. 2, pp. 179–187, feb 1962. [Online]. Available: http://ieeexplore.ieee.org/document/1057692/

[123] A. Buslaev, V. I. Iglovikov, E. Khvedchenya, A. Parinov, M. Druzhinin, and A. A. Kalinin, "Albumentations: Fast and Flexible Image Augmentations," *Information*, vol. 11, no. 2, p. 125, feb 2020. [Online]. Available: https://www.mdpi.com/2078-2489/11/2/125

[124] "ImageNet," 2012. [Online]. Available: http://www.image-net.org/ (Accessed 2020-08-12).

[125] X. Glorot and Y. Bengio, "Understanding the difficulty of training deep feedforward neural networks," in *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, ser. Proceedings of Machine Learning Research, Y. W. Teh and M. Titterington, Eds., vol. 9. Chia Laguna Resort, Sardinia, Italy: PMLR, 2010, pp. 249–256. [Online]. Available: http://proceedings.mlr.press/v9/glorot10a.html

[126] "RCI Info." [Online]. Available: http://rci.cvut.cz/ (Accessed 2020-08-12).

[127] "NumPy." [Online]. Available: https://numpy.org/ (Accessed 2020-08-12).

[128] "OpenCV." [Online]. Available: https://opencv.org/ (Accessed 2020-08-12).

[129] J. Canny, "A Computational Approach to Edge Detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. PAMI-8, no. 6, pp. 679–698, nov 1986. [Online]. Available: http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=4767851

[130] J. Sklansky, "Finding the convex hull of a simple polygon," *Pattern Recognition Letters*, vol. 1, no. 2, pp. 79–83, dec 1982. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/0167865582900162

[131] N. Otsu, "A Threshold Selection Method from Gray-Level Histograms," *IEEE Transactions on Systems, Man, and Cybernetics*, vol. 9, no. 1, pp. 62–66, jan 1979. [Online]. Available: http://ieeexplore.ieee.org/document/4310076/

[132] "Morphological closing." [Online]. Available: https://docs.opencv.org/trunk/d9/d61/tutorial{_}py{_}morphological{_}ops.html (Accessed 2020-08-12).

[133] "Blur." [Online]. Available: https://opencv-python-tutroals.readthedocs.io/en/latest/py{_}tutorials/py{_}imgproc/py{_}filtering/py{_}filtering.html (Accessed 2020-08-12).

[134] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern Classification*, 2nd ed. New York: Wiley, 2001.

[135] V. Franc, "Email communication," Center for Machine Perception, Czech Technical University, Prague, Tech. Rep., 2020.

[136] H. Lin, Y. Lu, X. Han, and L. Sun, "Cost-sensitive Regularization for Label Confusion-aware Event Detection," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Stroudsburg, PA, USA: Association for Computational Linguistics, 2019, pp. 5278–5283. [Online]. Available: https://www.aclweb.org/anthology/P19-1521

[137] A. Oliveira, S. Pereira, and C. A. Silva, "Retinal vessel segmentation based on Fully Convolutional Neural Networks," *Expert Systems with Applications*, vol. 112, pp. 229–242, dec 2018. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0957417418303816

[138] "DRIVE dataset." [Online]. Available: https://drive.grand-challenge.org/ (Accessed 2020-08-12).

[139] "HRF dataset." [Online]. Available: https://www5.cs.fau.de/research/data/fundus-images/ (Accessed 2020-08-12).

[140] "CHASEDB1 dataset." [Online]. Available: https://blogs.kingston.ac.uk/retinal/chasedb1/ (Accessed 2020-08-12).

[141] S. Ravishankar, A. Jain, and A. Mittal, "Automated feature extraction for early detection of diabetic retinopathy in fundus images," in *2009 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, jun 2009, pp. 210–217. [Online]. Available: https://ieeexplore.ieee.org/document/5206763/

[142] D. Siva Sundhara Raja and S. Vasuki, "Automatic Detection of Blood Vessels in Retinal Images for Diabetic Retinopathy Diagnosis," *Computational and Mathematical Methods in Medicine*, vol. 2015, pp. 1–12, 2015. [Online]. Available: http://www.hindawi.com/journals/cmmm/2015/419279/

[143] N. Memari, A. R. Ramli, M. I. B. Saripan, S. Mashohor, and M. Moghbel, "Retinal Blood Vessel Segmentation by Using Matched Filtering and Fuzzy C-means Clustering with Integrated Level Set

Method for Diabetic Retinopathy Assessment," *Journal of Medical and Biological Engineering*, vol. 39, no. 5, pp. 713–731, oct 2019. [Online]. Available: http://link.springer.com/10.1007/s40846-018-0454-2

[144] G. Indumathi and V. Sathananthavathi, "Microaneurysms Detection for Early Diagnosis of Diabetic Retinopathy Using Shape and Steerable Gaussian Features," in *Telemedicine Technologies*. Elsevier, 2019, pp. 57–69. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/B9780128169483000052

[145] "Albumentations example." [Online]. Available: https://github.com/albumentations-team/albumentations{_}examples/blob/master/notebooks/example{_}kaggle{_}salt.ipynb (Accessed 2020-08-12).

[146] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional Networks for Biomedical Image Segmentation," may 2015. [Online]. Available: http://arxiv.org/abs/1505.04597

[147] P. Yakubovskiy, "Segmentation Models Pytorch," 2020. [Online]. Available: https://github.com/qubvel/segmentation{_}models.pytorch (Accessed 2020-08-12).

[148] M. Tan and Q. V. Le, "EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks," may 2019. [Online]. Available: http://arxiv.org/abs/1905.11946

[149] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. Alemi, "Inception-v4, Inception-ResNet and the Impact of Residual Connections on Learning," feb 2016. [Online]. Available: http://arxiv.org/abs/1602.07261

[150] S. Xie, R. Girshick, P. Dollár, Z. Tu, and K. He, "Aggregated Residual Transformations for Deep Neural Networks," nov 2016. [Online]. Available: http://arxiv.org/abs/1611.05431

[151] W. Liu, "Dice loss implementation." [Online]. Available: https://gist.github.com/weiliu620/52d140b22685cf9552da4899e2160183 (Accessed 2020-08-12).