Analysing Retinal Fundus Images with Deep Learning Models

by

Samuel Ofosu Mensah

Dissertation presented for the degree of Doctor of Philosophy in Applied Mathematics in the Faculty of Science at Stellenbosch University

Supervisor: Dr. Bubacarr Bah

Co-supervisor: Prof. Willie Brink

December 2023

Declaration

By submitting this dissertation electronically, I declare that the entirety of the work contained therein is my own, original work, that I am the sole author thereof (save to the extent explicitly otherwise stated), that reproduction and publication thereof by Stellenbosch University will not infringe any third party rights and that I have not previously in its entirety or in part submitted it for obtaining any qualification.

Date: December 2023 . . . . . .

Abstract

Convolutional neural networks (CNNs) have successfully been used to classify diabetic retinopathy but they do not provide immediate explanations for their decisions. Explainability is relevant, especially for clinicians. To make results explainable, we use a post-attention technique called gradient-weighted class activation mapping (Grad-CAM) on the penultimate layer of deep learning models to produce localisation maps on retinal fundus images after using them to classify diabetic retinopathy. Moreover, the models were initialised using pre-trained weights obtained from training models on the ImageNet dataset. The results of this are fewer training epochs and improved performance. Next, we predict cardiovascular risk factors (CVFs) using retinal fundus images. In detail, we use a multi-task learning (MTL) model since there are several CVFs. The impact of using an MTL model is the advantage of simultaneously training for and predicting several CVFs rather than doing so individually. Also, we investigate the performance of the fundus cameras used to capture the retinal fundus images. We notice a superior performance of the desktop fundus cameras to the handheld fundus camera. Finally, we propose a hybrid model that fuses convolutions and Transformer encoders. This is done to harness the benefits of convolutions and Transformer encoders. We compare the performance of the proposed model with other attention-based models and observe on-par performance.

Keywords: Diabetic retinopathy · CNN · Grad-CAM · Multi-task learning · Attention · Transformers

Acknowledgements

I want to express my sincere gratitude to my God and Father of my Lord Jesus Christ for seeing me throughout my entire doctoral journey.

I would like to extend my heartfelt appreciation to my supervisors, *Dr Bubacarr Bah* and *Prof Willie Brink*, whose unwavering dedication greatly enriched the development of my dissertation. Their invaluable guidance, insightful suggestions, and patience were pivotal in bringing my doctorate to a successful end.

I would like to acknowledge the generous support provided by the German Academic Exchange Service (DAAD), whose sponsorship not only facilitated my doctoral pursuit but also encompassed essential aspects such as health insurance and travel grants. Thanks to this support, I was able to embark on a research visit to Germany.

My sincere thanks go out to the South African Cluster for High-Performance Computing (CHPC), which kindly provided me access to supercomputing resources for conducting my experiments. Their staunch support was unwavering even in the face of difficulties.

I would like to thank the African Institute for Mathematical Sciences (AIMS) for providing me with a stimulating research atmosphere, replete with an office space conducive to scholarly pursuits.

I am thankful to *Prof Dr Philipp Berens* of BerensLab at the Hertie Institute for Artificial Intelligence and Brain Research at the University of Tübingen, as well as to *eye2you*, also located in Tübingen, for extending their warm invitation for a research visit. My scholarly horizons have been greatly widened by this opportunity.

A profound appreciation goes to my circle of friends, whose unwavering lifted my spirits throughout my doctoral journey. Your exceptional encouragement was a constant source of motivation.

I am indebted to my family, especially my parents, *Mr and Mrs Emmanuel K. Mensah*, and my siblings, *Eunice*, *Emmanuel*, and *Esther*, for their unwavering love and support throughout my academic pursuits. Your steadfast belief in me has been a cornerstone of my achievements.

Dedication

I dedicate this work to my dear parents Mr and Mrs Emmanuel K. Mensah.

Contents

List of Figures

List of Tables

List of symbols

Abreviations

APTOS	Asian Pacific Tele-Ophthalmology Society
CLAHE	Contrast Limited Adaptive Histogram Equalisation
CNN	Convolutional Neural Network
CVD	Cardiovascular Disease
CVF	Cardiovascular Risk Factor
DR	Diabetic Retinopathy
EyePACS	Eye Picture Archive Communication System
FCT	Fully Convolutional Transformer
Grad-CAM	Gradient-weighted Class Activation Mapping
HPS	Hard-Parameter Sharing
MSA	Multi-head Self-Attention
MTL	Multi-Task Learning
NPDR	Non-Proliferative Diabetic Retinopathy
PDR	Proliferative Diabetic Retinopathy
SPS	Soft-Parameter Sharing
ViT	Vision Transformer

Chapter 1

Introduction

This dissertation employs deep learning models to predict diabetic retinopathy (DR) and cardiovascular risk factors (CVFs). The occurrence of DR is significantly higher in individuals with diabetes, and is characterised by the development of lesions on the retina. This condition may result in vision loss and blindness as lesions can be formed on the macula region of the retina. Also, CVFs serve as biomarkers for cardiovascular diseases (CVD), which can potentially damage the blood supply of the retina and give rise to complications such as blood clot formation in the macula region of the retina. Timely diagnosis and intervention of these diseases can play a crucial role to slow their progression. In this context, we leverage the advantages of deep learning to predict and localise retinal lesions, with a focus on salient regions of interest found by attention mechanisms.

1.1 Background

The retina is an important part of the eye. It consists of photo-sensitive tissues and sends electric or neural signals obtained by converting photon energy from light to the brain, for interpretation through the optic disc [Abràmoff *et al.*, 2010; Gahir and Shah, 2020; Gramatikov, 2014; Rodriguez *et al.*, 2022]. For this reason, there is a continual need

to better understand the retina. This has become possible by observing and capturing images of the retina in a non-invasive manner. A retinal fundus camera can capture the blood vessels, the optic disc, the macula region and the fovea (Figure 1.1).

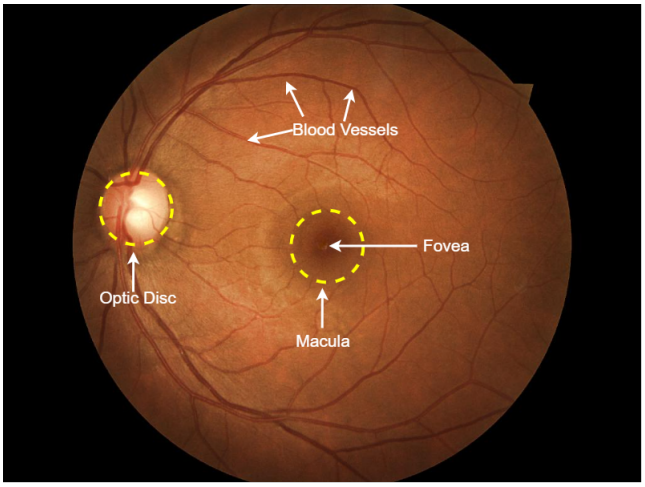


Figure 1.1: An example of a retinal fundus image showing the blood vessels, the optic disc, the macula region and the fovea.

The blood vessels (the veins and arteries) extend outward from the optic disc and they are responsible for the oxygenation and metabolism of the retina [Gramatikov, 2014; Kowluru and Chan, 2008; Purves *et al.*, 2001]. Due to different levels of oxygen in the blood vessels, the veins show a deeper red than the arteries [Gahir and Shah, 2020]. The optic disc is approximately 1.5 mm in diameter, and connects the eye to the brain's visual processing centre through the optic nerves. All blood vessels leave and enter the retina through the optic disc [Zhu *et al.*, 2012]. While seen as a bright spot on the retina, the optic disc creates a blind spot as it contains no photoreceptors [Zhu *et al.*, 2012]. Covering about 3 mm [Khan *et al.*, 2021], the macula region can be found near the centre of the retina. In the middle of the macula is the fovea, and it is the primary target as light enters the eye. The fovea is responsible for producing sharp, clear and detailed visuals, making it the most sensitive part of the retina [Zhu *et al.*, 2012]. To optimally produce high-quality images, the fovea contains a high concentration of tightly packed

visual cells [Gahir and Shah, 2020; Mookiah *et al.*, 2013] in a tiny spot, with a mean visual cells per area of $161900/\text{mm}^2$ [Kolb, 2011].

In order to prevent interruptions of light from entering the fovea, the macula region does not have any blood vessels [Mookiah *et al.*, 2013]. This is possible because a restrictive barrier, called the blood-retina barrier, isolates the macula region and fovea from circulation and regulates their flow of nutrients [Bhagat *et al.*, 2009]. Unfortunately, the accumulation of sugar in the blood could lead to the breaking of the tiny blood vessels on the retina [Mittal and Rajam, 2020]. This may result in vascular leakage where fluids such as blood, protein and lipids leak into the fovea and macula region, leading to loss of central and possibly peripheral vision [Nayak *et al.*, 2013]. This incidence is common in individuals with diabetes (Figure 1.2).

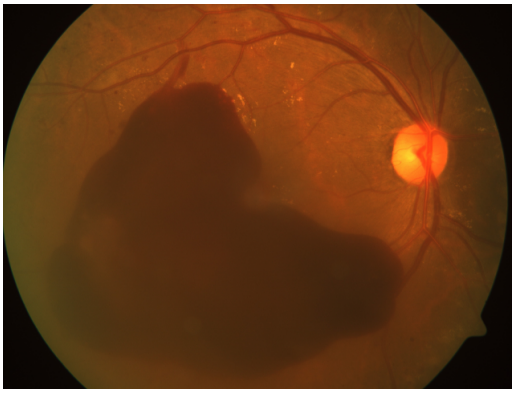
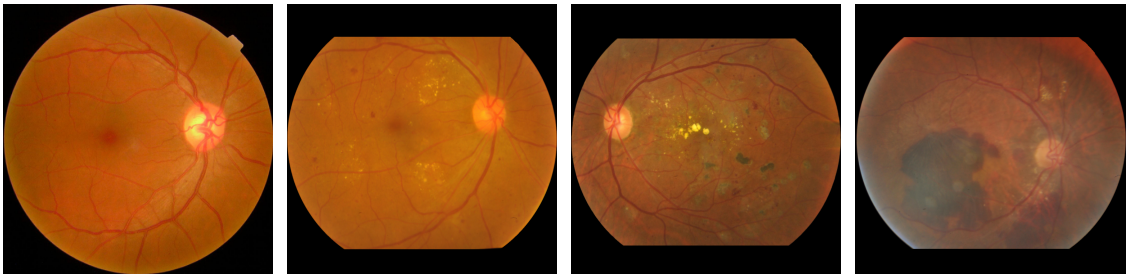


Figure 1.2: A retina with leaked blood on the fovea and macula region. This image is a random example selected from the data used for the study.

The World Health Organisation [WHO, 2022] defines diabetes as a chronic disease that occurs either when the pancreas does not produce enough insulin or when the body cannot effectively use the insulin it produces. Insulin is a hormone that regulates blood sugar levels [Ormazabal *et al.*, 2018]. Hyperglycemia is a term used to describe raised blood sugar in the body, which is common in diabetic patients. It is known to cause significant damage to blood vessels in many parts of the body, including the retina. This

usually results in restricted blood flow, a condition called ischemia. The term given to diabetic complications on the retina is diabetic retinopathy (DR) [Abràmoff *et al.*, 2010].

DR is one of the leading causes of preventable blindness among the working adult population [Teo *et al.*, 2021], that is between the ages of 20 and 74 years [Lee *et al.*, 2015]. DR is progressive [Nayak *et al.*, 2013] and can be categorised according to its level of severity [Antonetti *et al.*, 2021]. There are two main categories: non-proliferative DR (NPDR) and proliferative DR (PDR) [Antonetti *et al.*, 2021; Jadhav and Patil, 2015; Mittal and Rajam, 2020]. NPDR can be subdivided into three stages in order of severity, namely mild, moderate and severe. Usually, there is only one micro-aneurysm at the end of the blood vessels in a mild DR. The burst of the micro-aneurysm (known as haemorrhage) on the retina results in a moderate DR. In severe DR, there are several haemorrhages on the retina. PDR results in neovascularisation which is defined as the natural formation of new blood vessels. Neovascularisation mostly occurs under the surface of the retina, resulting in tractional retinal detachment [Antonetti *et al.*, 2021; Qummar *et al.*, 2019] (Figure 1.3).



(a) Mild **(b)** Moderate **(c)** Severe **(d)** PDR

Figure 1.3: Examples of diabetic retinopathy in different levels of severity. These images are random examples selected from the data used for the study.

Another common disorder that affects retinal vasculature is cardiovascular disease (CVD). It is the leading cause of death globally, and killed approximately 17.9 million people in 2019 [WHO, 2021]. CVD is defined as a group of disorders involving the heart and blood circulation system such as hypertension, stroke, coronary heart disease and pe-

ripheral vascular disease [Salkind, 2005]. Early signs of CVD can be observed on the retina [Nguyen and Wong, 2009]. Furthermore, factors including age, sex, blood pressure, smoking status, diabetes and hypertension are significant contributors to CVD. These factors are known as cardiovascular risk factors (CVFs) [Poplin *et al.*, 2018].

These kinds of observations are possible because the retina can be observed in a non-invasive manner through the pupil using a retinal fundus camera. A retinal fundus camera is a special camera that is used to capture high-resolution images of the retina. It is set up with a low-power microscope to produce a 2D and a 2.5 magnified view of the retina [Dodo, 2020; Mookiah *et al.*, 2013]. The average field of view of a retinal fundus camera is only from 30° to 50° due to the size of the pupil [Panwar *et al.*, 2016]. For this reason, a retinal fundus camera is unable to capture a full view of a retina. To address this, ophthalmologists mainly capture the central and peripherals of the retina [Al-Bander, 2018]. Top manufacturers of retinal fundus cameras include Canon, Centrevue, Crystalvue, Optovue, Topcon and Zeiss [Al-Bander, 2018; Dodo, 2020]. The use of retina images is relevant because they provide ophthalmologists with a precise in-vivo observation of the retina, enabling them to adequately monitor and document diseases that affect the retina and also help prevent blindness through early detection [Gulshan *et al.*, 2016; Nguyen and Wong, 2009].

1.2 Problem Statement

A retinal fundus image can provide helpful information to ophthalmologists [Rim *et al.*, 2020] such as the level of severity of diabetic retinopathy and biomarkers of hypertension, assisting them to precisely characterise changes on the retina [Zhang *et al.*, 2020]. However, there is a dearth of ophthalmologists for the increased demand and number of retinal fundus images created [Raman *et al.*, 2019; Rogers *et al.*, 2021]. In addition, there remains the issue of subjective diagnosis by trained specialists [Gargeya and Leng, 2017]. Hence, there is a need to develop solutions to help resolve the unmatched propor-

tion between ophthalmologists and patients [Raman *et al.*, 2019], and to help alleviate workload [Gargeya and Leng, 2017].

In this study, we consider the problem of classifying DR and CVD using deep learning on retinal fundus images. Deep learning is a sub-domain of machine learning approaches characterised by multiple neural network layers for the finding of predictive patterns based on examples [Marcus, 2018; Poplin *et al.*, 2018]. Specifically, we look at convolutional neural networks (CNNs). Over the years, CNNs have been studied extensively [Gu *et al.*, 2018], have demonstrated good performance and have dominated the domain of computer vision [He *et al.*, 2016; Simonyan and Zisserman, 2015; Szegedy *et al.*, 2015]. CNNs are known for their shift-invariance and local connectivity properties [Chu *et al.*, 2021]. Despite their success, they lack the ability to learn long-range dependencies due to poor scaling properties with respect to large receptive fields [Ramachandran *et al.*, 2019].

As other domains have adapted CNNs, recent studies in computer vision have also adapted models from different domains. Such an example is the introduction of the vision Transformer (ViT) model, which has achieved impressive results on various computer vision tasks [Dosovitskiy *et al.*, 2021]. ViT models address the scalability [Dosovitskiy *et al.*, 2021] and long-range dependency [Touvron *et al.*, 2022] issues faced by CNNs. They have dynamic attention properties [Wu *et al.*, 2021], which may be necessary for attending to lesions in DR fundus images. Also, they generalise better [Touvron *et al.*, 2022], which is important when working with real-world data. Unfortunately, the number of operations in the ViT grows quadratically with the number of pixels of an input image making it computationally expensive [Chu *et al.*, 2021].

The core component of a CNN is convolutional layers (Convs) and that of a ViT is multi-head self-attention (MSA). Park and Kim [2022] observed that Convs and MSA complement each other. In detail, they reported that (1) Convs diversify feature maps, but MSAs aggregate them, and (2) Convs are high-pass filters and MSAs are low-pass filters. In the end, Park and Kim [2022] noted that harmonising Convs with MSA yields

better results with improved robustness, compared to training a model with either Convs or MSA independently. Hence, leveraging the advantages of the two models can be beneficial.

1.3 Aims and Objectives

The aims of the study are to build and evaluate deep learning models on retinal fundus images for predicting diabetic retinopathy and cardiovascular diseases. We will develop and evaluate the performance of state-of-the-art deep learning models trained on retina images. To that end, we focus on the following objectives:

1. To obtain optimal pre-processing of retinal fundus images for subsequent tasks such as classification. In particular, we will estimate optimal hyperparameters for contrast limited adaptive histogram equalisation (CLAHE) on each image.

2. To demonstrate the impact of using transfer learning instead of randomly initialising the model weights, for classifying the severity of diabetic retinopathy in retinal images.

3. To generate localisation maps for discriminative regions on a retinal fundus image for the task of classifying diabetic retinopathy. These localisation maps can be generated to aid in providing explanations to clinicians for a model's discriminative process.

4. To demonstrate the advantages of using multi-task learning over a single-task learning model in predicting cardiovascular risk factors from retinal fundus images.

5. To present the current advantages of using a desktop fundus camera and possible future benefits of using a handheld fundus camera.

6. To propose and demonstrate the benefits of using hybrid models that combine convolutional layers and multi-head self-attention.

1.4 Contribution

Our contributions focus on building and evaluating deep learning models on retinal fundus images. Thus, the dissertation makes the following contributions.

Chapter 2. To start with, we use contrast limited adaptive histogram equalisation (CLAHE) to resolve the enhancement issues with retinal fundus images. In particular, we estimate the hyperparameters for CLAHE individually for each retinal fundus image. Next, we employ transfer learning for classification purposes and show its benefits. Finally, we use gradient-weighted class activation mapping (Grad-CAM) to generate discriminative regions on the retinal fundus images. The results were published in the following paper:

> **Paper**: Mensah, S. O., Bah, B., & Brink, W. (2021). Towards the Localisation of Lesions in Diabetic Retinopathy. In Intelligent Computing (pp. 100-107). Springer, Cham.

Chapter 3. We build a multi-task learning (MTL) model to predict cardiovascular risk factors (CVFs) from the retina images. The CVFs considered in the study are age, sex, and hypertension. In addition, we assess the performance of the MTL model on patients' ethnicity, investigate the contribution that the optic disc and the macula region have on the prediction of sex, investigate the influence that the quality of a retinal fundus have on the performance, and evaluate camera performance for the study. These results were in the following pre-print:

Due for journal submission: Mensah, S. O., Koch, L., Lies, P., Wallraven, C., Bah, B., Browatzki, B., & Berens, P. (2023). Evaluation of Multi-Task Learning for Predicting Cardiovascular Risk Factors from Retinal Fundus Images.

Chapter 4. We propose a hybrid model that fuses convolutions and Transformers, consequently benefiting from their advantages and eliminating undesirable properties from the two paradigms. In particular, we feed intermediate feature maps from a ResNet-50 model to a fully convolutional Transformer to predict diabetic retinopathy. We evaluate the proposed model using images of the Canadian Institute for Advanced Research image dataset with ten classes (CIFAR-10). We then also evaluate the performance of the proposed model along with convolution-based attention models, variants of vision Transformer models, and other hybrid models on the task of classifying severity of diabetic retinopathy. These results were published in the following paper:

Paper: Mensah, S. O., Bah, B., & Brink, W. (2022). Learning to Pay Multiple Attention with Fully Convolutional Transformers. Southern African Conference for Artificial Intelligence Research (SACAIR, 2022), pp. 67 – 77.

1.5 Dissertation Overview

In this section, we provide an overview of the dissertation and briefly describe the content of each chapter.

Chapter 2 - Diabetic Retinopathy Classification and Lesion Localisation with CNNs

CNNs have demonstrated good performance on medical images, including retinal fundus images. However, they fail to provide explanations for their discriminating pro-

cesses, making the results potentially usable for ophthalmologists but without an understanding of the decisions made by the model. Thus, this chapter focuses on providing the steps involved in generating localised maps for classifying diabetic retinopathy.

Chapter 3 - Prediction of Cardiovascular Risk Factors

It is possible to predict cardiovascular risk factors (CVFs) from retinal fundus images. However, predicting each CVF with a separate model can be time-consuming. This chapter focuses on building a multi-task learning model to simultaneously predict several CVFs, consequently saving time and reducing the number of parameters to learn.

Chapter 4 - Combining CNNs and Vision Transformers

The introduction of Transformers in the computer vision domain has resulted in improved results, compared to CNNs. One of the main components of the Transformer model is attention. CNNs and Transformers have desired underlying properties for a computer vision task. In this chapter, we propose a hybrid model that fuses convolutions and Transformers, for the task of predicting the severity of diabetic retinopathy.

Chapter 5 - Conclusion

In this chapter, we summarise the findings that were observed during the experiments conducted throughout previous chapters. We provide remarks on how the aims and objectives presented in the dissertation were met, and end with suggestions for future work.

1.6 Ethical Statement

An institutional Review Board approval was not required for the dissertation, because we utilised a database of retinal fundus images collected by APTOS and EyePACS which

are publicly available datasets containing no patient-identifiable information. Our research focuses on generating valuable insights while ensuring that the utilisation of the datasets contributes significantly to the broader scientific community.

Chapter 2

Diabetic Retinopathy Classification and Lesion Localisation with CNNs

2.1 Overview

The aim in this chapter is to localise lesions on the retina using a post-attention technique called gradient-weighted class activation mapping (Grad-CAM) [Selvaraju *et al.*, 2017]. First, we present an enhancing technique to resolve potential contrast issues in retinal fundus images. Next, we use state-of-the-art convolutional neural networks (CNNs) to extract feature maps from the data. We feed the extracted feature maps to Grad-CAM and generate coarse localisation maps to help identify discriminative regions on the retinal fundus images. In the following sections, we first present background in Section 2.2, then discuss previous related work in Section 2.3, describe the methods used in this chapter in Section 2.4, describe the experiments in Section 2.5, discuss the results in Section 2.6, and finally conclude in Section 2.7.

2.2 Background

There has been tremendous success in the field of deep learning, especially in the domain of computer vision [Simonyan and Zisserman, 2015]. Convolution neural networks (CNNs) have played a critical role in this success [LeCun *et al.*, 1998]. Krizhevsky *et al.* [2012] built one of the first deep CNNs using the ImageNet dataset and attained state-of-the-art performance on a large-scale image classification benchmark. Since then, several deep CNN models have been introduced.

The use of CNN models has now been extended to specialised fields such as medicine [Raghu *et al.*, 2019]. General use cases of CNNs in the medical domain include classification, image segmentation, localisation, detection, and image generation [Çallı *et al.*, 2021]. This has opened a wide avenue of research for scientists. In addition, the use of transfer learning has further propelled the success of CNNs in the medical domain [De Fauw *et al.*, 2018; Gondal *et al.*, 2017; Gulshan *et al.*, 2016; Raghu *et al.*, 2019]. It is therefore important that we find among the numerous deep CNN models available which one performs best for our task of classifying the severity of diabetic retinopathy in retinal fundus images.

Even though these models are able to attain good performance on medical images, it can be challenging to comprehend the reasoning behind their decisions. Clinicians need more evidence from these models to increase trust and comprehension of the results produced [Gondal *et al.*, 2017]. To resolve this, we consider the use of a post-attention technique called gradient-weighted class activation mapping (Grad-CAM) [Selvaraju *et al.*, 2017] to generate coarse localisation maps on the lesions in retinal fundus images. Unlike other approaches, Grad-CAM is capable of generating class-specific localisation maps on any CNN-based architecture without restructuring the model and requires no additional operations, thus making it efficient [Bazzani *et al.*, 2016; Oquab *et al.*, 2014; Simonyan *et al.*, 2014; Springenberg *et al.*, 2015; Zeiler and Fergus, 2014; Zhou *et al.*, 2016]. The aim of generating the localisation maps is to provide visual explanations

for the decisions made by the model. In addition, generating coarse localisation maps on retinal fundus images (or on medical images, in general) can be categorised as a computer-aided diagnosis tool that can speed up diagnosis. Beede *et al.* [2020] showed that integrating deep learning algorithms in clinician workflow significantly reduces diagnosis time (see Figure 2.1).

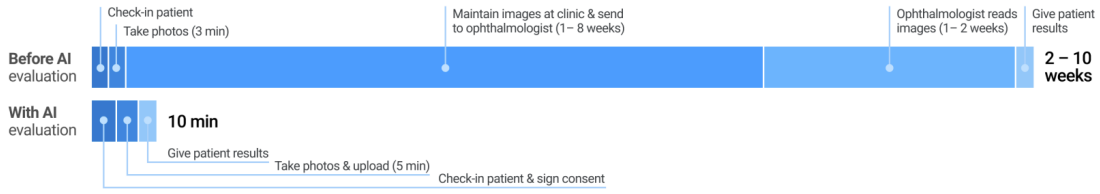


Figure 2.1: A comparison of the diagnostic time before and after using deep learning [Beede *et al.*, 2020].

Before implementing deep learning models, we pre-process images using contrast limited adaptive histogram equalisation (CLAHE). CLAHE is a crucial step because it helps resolve the low-contrast issues, uneven illumination and poor-quality images associated with retinal fundus images Zhou *et al.* [2017]. CLAHE has two hyperparameters, namely the clip limit and the grid size. Most studies select unique values of hyperparameters for all the images. However, the optimal set of hyperparameters varies from image to image [Kuran and Kuran, 2021]. Hence, it is recommended that each image has its own set of hyperparameters. Moreover, the selection process of these hyperparameters is hardly discussed in studies that employ CLAHE as a pre-processing technique [Campos *et al.*, 2019].

In this chapter, we estimate optimal hyperparameters for CLAHE in each image during the pre-processing. We then use pre-trained weights from deep learning models to classify the severity of diabetic retinopathy in retinal fundus images. The deep learning models we consider are ResNet-50 [He *et al.*, 2016], VGG-16 [Simonyan and Zisserman, 2015], Inception-V3 [Szegedy *et al.*, 2015] and InceptionResNet-V2 [Szegedy *et al.*, 2017]. Next, we find the best-performing CNN model for the classification task.

Finally, we generate coarse localisation maps on the retinal fundus images using the best-performing model.

2.3 Related Work

In this section, we present related works on the techniques used in this chapter. They include contrast-enhancing techniques for diabetic retinopathy, transfer learning and localising regions of interest in images.

2.3.1 Contrast Enhancing Techniques for Retinal Fundus Images

Medical images, including retinal fundus images, suffer from low contrast problems [Cao and Li, 2020; Gupta and Tiwari, 2019]. These contrast issues are usually resolved by using histogram equalisation (HE). Salem *et al.* [2019] experimented with four variants of HE methods on five medical images. Their goal was to determine which variant of HE works best with medical image datasets. In the end, they noted contrast limited adaptive histogram equalisation (CLAHE) as the best enhancement technique for retina images, while quadrant dynamic histogram equalisation (QDHE) worked best with brain, endometrium, breast, and knee images.

Other studies have employed CLAHE in different colour spaces [Setiawan *et al.*, 2013; Zhou *et al.*, 2017]. Zhou *et al.* [2017] resolved the retina image enhancement issue by enhancing the luminosity channel in the LAB colour space using CLAHE. They evaluated their technique on poor-quality retina images on a quality scale from 0 to 1. They observed improved colour quality averages from 0.0404 to 0.4565. Setiawan *et al.* [2013] noted that applying CLAHE on the individual channels of an RGB retinal fundus image reveals different attributes of an image. Particularly, they observed that applying CLAHE only to the red channel diminished the intensities of the blood vessels while

increasing the background structure of the retina. When CLAHE was applied only to the green channel, there was a significant improvement in the visibility of blood vessels on the retinal fundus images. The blue channel, on the other hand, produced hazy outputs.

Mohan and Mahesh [2013] used particle swarm optimisation (PSO) for tuning the hyperparameters for CLAHE. Notably, PSO is susceptible to premature convergence, displays partial optimism capabilities, suffers computational overhead and high tuning complexity that emanate from its hyperparameters including the number of particles, acceleration coefficients and termination conditions [Juneja and Nagar, 2016; Li *et al.*, 2014; Rahman *et al.*, 2016]. More *et al.* [2015] and Kuran and Kuran [2021] both utilised a meta-heuristic to tune the parameters of CLAHE, which is typically parameter sensitive [Huang *et al.*, 2019]. Campos *et al.* [2019] also presented a machine learning approach to select the best hyperparameters for CLAHE. However, this approach requires additional data labelling of well-contrasted instances of the dataset and training a model to select hyperparameters for CLAHE.

In this study, we use CLAHE for pre-processing because it works well with retinal fundus images. We estimate the hyperparameters for CLAHE by finding the maximum curvature in each entropy function curve. We use this approach because it is easy to implement and provides a quick evaluation of the hyperparameters [Min *et al.*, 2013].

2.3.2 Traditional Machine Learning

Traditional machine learning models such as random forests, support vector machines (SVMs) and k-nearest neighbours (k-NN) are simple and their results are easy to comprehend [James *et al.*, 2013]. In the past, these traditional machine learning models were used to classify diabetic retinopathy. To start, Dara and Tumma [2018] defined feature extraction as the process of converting data into a set of features. It simplifies data by removing redundant variables and keeping the informative ones. Examples of feature extraction include principal component analysis (PCA), histogram of oriented

gradients (HOG), local binary patterns (LBP), speeded-up robust features (SURF) and many others.

Alzami *et al.* [2019] predicted the severity of DR using a random forest but not without extensive manual feature extraction. They created a pipeline that included seven feature extraction techniques plus a fractal dimension technique (another feature extraction method) as their main contribution. To classify the severity of DR using an SVM, Carrera *et al.* [2017] first isolated blood vessels, microaneurysms and hard exudates in order to extract features. Together with the manual feature extraction techniques, Gandhi and Dhanasekaran [2013] used an SVM to first classify normal and diseased retinal fundus images before classifying the diseased images into the sub-classes. Labhade *et al.* [2016] compared the performance of several traditional machine learning models in classifying DR. They obtained a maximum area under the ROC curve (AUC) score of 93% and the best-performing model was a random forest. Reddy *et al.* [2020] created an ensemble model consisting of five traditional machine learning models to classify DR. Additionally, there are several works similar to the ones described above [Roychowdhury and Banerjee, 2018; Tjandrasa *et al.*, 2013; Verma *et al.*, 2011; Zhang and Chutatape, 2005]. In short, traditional machine learning needs manual feature extraction for good results. These manual feature extractions should often be done by domain experts.

2.3.3 Deep Learning

Deep learning models take in the data without manual feature extractions. The base of deep learning is neural networks, which are mathematical models inspired by how the brain functions. They are made up of layers which contain several nodes connected together from layer to layer. The nodes are known as the neurons or units of a layer. A neural network with several stacked layers creates a deep model. In detail, the layers are mathematical operations. A neural network can consist of three parts, namely the input layer, one or more hidden layer and the output layer.

To emphasise, deep learning models perform feature extractions automatically. They are simple in the sense that they do not require manual feature extraction on the input data. They can easily be scaled to large datasets and their learned features can be transferred from one domain to another. Large datasets are crucial for deep learning models because they offer a high-quality data representation that leads to high model performance. This is mainly due to the wide range of examples provided in a large dataset, which helps deep learning models generalise better on unseen data and learn meaningful feature representation [Najafabadi *et al.*, 2015]. Deep learning is possible mainly due to increased computing power and increased memory [Khan *et al.*, 2018]. In deep learning for computer vision, the most used mathematical operation is the convolutional operation.

Given a 2-D image I of size $h \times w$ and a 2-D kernel K of size $p \times q$, the convolutional operation is defined as

$$O_{(r,c)} = (I \odot K)_{(r,c)} = \sum_{a=0}^{p-1} \sum_{b=0}^{q-1} K_{(a,b)} I_{(r-a,c-b)}, \tag{2.3.1}$$

where $\odot$ represents the convolutional operation and O denotes the output. $r \in \mathbb{Z}^+ | 1 \leq r \leq (h-p)$ and $c \in \mathbb{Z}^+ | 1 \leq c \leq (w-q)$ are the row and column indices respectively of the image and output. A neural network with a convolutional operation is called a convolution neural network (CNN), and requires fewer parameters for training. One of the earliest deep CNN models was introduced by Krizhevsky *et al.* [2012]. They stacked five convolutional layers and three fully connected layers, and won the ImageNet Large-Scale Visual Recognition Challenge in 2012. Since then, CNN models have become deeper and several studies have invented unique architectures. We briefly describe the models used in our study below.

VGG-16. VGG-16 is a model introduced by Simonyan and Zisserman [2015]. Its main characteristics are the high depth of convolutional layers and small kernel size. In Figure 2.2, we see that several convolutional layers, each with a 3×3 kernel size, are stacked together. Also, the feature maps decrease in size while the channel size grows

as the network goes deeper. This is because a pooling layer of size 2 is used to summarise the feature maps. VGG stands for Visual Geometry Group and the 16 represents the number of layers in the model.

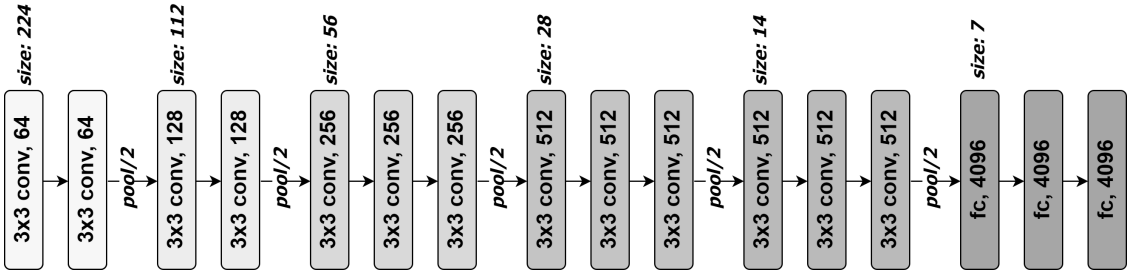


Figure 2.2: The VGG-16 architecture. The numbers inside a layer represents the number of channels. In this model, the feature map size decreases as the depth of the model increases.

Inception-V3. Szegedy *et al.* [2015] introduced a deep CNN model with a unique module called inception. A module in the context of deep learning is composed of known mathematical structures that perform specific computations and is designed to be modular and reusable [Abadi *et al.*, 2015; Paszke *et al.*, 2019]. The inception module branches its input into three convolutional operations with different kernel sizes and a max pooling of size 3. It then concatenates the outputs of these operations for subsequent layers. We illustrate the inception module in Figure 2.3. Inception-V3 consists of several inception modules stacked on top of each other. In total it has 42 layers including 11 blocks of the inception module.

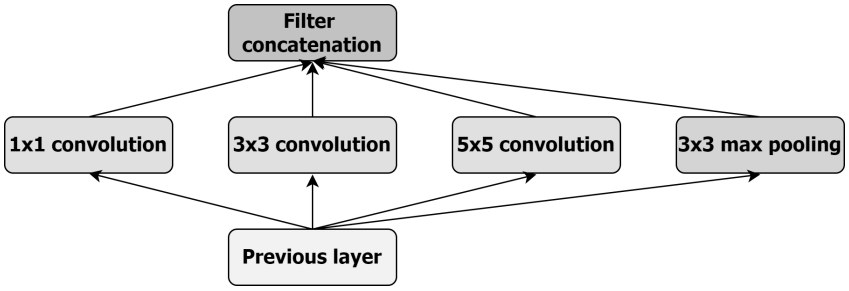


Figure 2.3: Illustrating the inception module. The module concatenates outputs from three convolutional operations with different kernel sizes and a max pooling operation with a 3×3 kernel size.

ResNet-50. An even deeper model is the ResNet-50 model introduced by He *et al.* [2016]. Inspired by the significance of the depth of a model, they explicitly formulated their layers as learning residual functions making reference to the layer inputs. This approach of referring to layer inputs is termed skip connections. Figure 2.4 shows the layout of a residual module. Similar to the inception model, ResNet stacks several residual modules together. The number at the end of the name denotes the number of layers in the model. For example, ResNet-152 means there are 152 layers in the model.

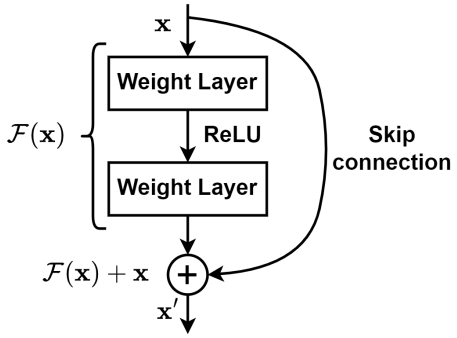


Figure 2.4: Showing the layout of the residual module. The unique feature of the residual module is the skip connection.

InceptionResNet-V2. Szegedy *et al.* [2017] creatively mixed the inception and ResNet modules to create the InceptionResNet module. The module works like an inception module while making reference to the layer input. We show the layout in Figure 2.5. Several InceptionResNet modules can be stacked together to create a model.

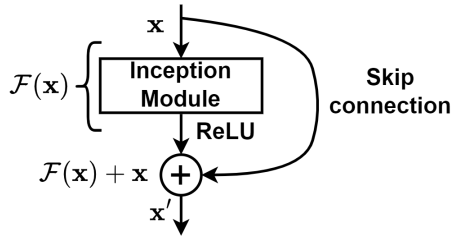


Figure 2.5: The InceptionResNet module mixes the inception module with the residual module.

These CNN models have revolutionised the field of computer vision [O'Mahony *et al.*, 2019; Raghu *et al.*, 2019]. The success has been extended to the medical field [Cai *et al.*, 2020; Raghu *et al.*, 2019; Sarvamangala and Kulkarni, 2022; Yadav and Jadhav, 2019]. The models described above have been used on various medical datasets including computed tomography (CT) [Liu *et al.*, 2017; Marcos *et al.*, 2022; Power *et al.*, 2016], radiography (X-ray) [Chouhan *et al.*, 2020; Jain *et al.*, 2020], magnetic resonance imaging (MRI) [Khan *et al.*, 2020; Lu *et al.*, 2020; Majib *et al.*, 2021], histopathological images [Alom *et al.*, 2019; Jiang *et al.*, 2019; Toğaçar *et al.*, 2020], and retinal fundus images [Gulshan *et al.*, 2016; Raghu *et al.*, 2019].

Focusing on retinal fundus images, Gulshan *et al.* [2016] conducted one of the earliest studies on classifying DR with CNNs. In detail, they used an Inception-V3 model to classify a retinal fundus image as either non-proliferative diabetic retinopathy or proliferative diabetic retinopathy. Mateen *et al.* [2018] incorporated singular value decomposition (SVD) and principal component analysis (PCA) to the final layers of VGG-16 to predict the severity of DR. In our study, we compared the performance of the models described above in predicting the severity of DR [Mensah *et al.*, 2021].

2.3.4 Localising Regions of Interest

It is possible to extract discriminative regions of an image with a trained model [Zhu *et al.*, 2019] and particularly with a CNN [Angeletti *et al.*, 2018]. Discriminative regions are usually objects of interest in an image. The process of identifying the location of objects of interest in an image is called localisation. While segmentation outputs a pixel-wise mask of each object in the image, it does not give information about the location of the objects of interest [Zhu *et al.*, 2019]. Localisation can be seen as another form of classification, but classification does not produce information about the location of objects of interest. However, localisation maps can be generated from classification models [Hui *et al.*, 2022]. Objects of interest are relevant to increase the trust and comprehension of a model.

Several studies have been conducted on the topic of localisation [Gan *et al.*, 2015; Simonyan *et al.*, 2014; Springenberg *et al.*, 2015; Zeiler and Fergus, 2014]. A popular example for CNNs is class activation mapping (CAM) [Zhou *et al.*, 2016]. CAM is used to generate maps which indicate discriminative regions of an image. However, CAM can only utilise the penultimate layer of a CNN model to generate localisation maps. Selvaraju *et al.* [2017] introduced gradient-weighted class activation mapping (Grad-CAM) which is capable of generating localisation maps from all the layers in a model. Hence, Grad-CAM is a more general version of CAM [Selvaraju *et al.*, 2017].

Explaining the decisions deep learning models make on medical images has improved [Singh *et al.*, 2020]. Grad-CAM has been used to locate brain tumour in MRI images [Pereira *et al.*, 2018]. Bhusal *et al.* [2022] used Grad-CAM to detect pathology in chest X-ray dataset. Barnett *et al.* [2021] applied Grad-CAM to mammograms to calculate activation precision, which is their measure of interpretability. To shed more light on CNNs, Young *et al.* [2019] generated localisation maps on skin images for the task of melanoma detection. In our study, we use Grad-CAM to localise lesions in retinal fundus images.

2.4 Methodology

In this section, we present the methods used in the chapter. We expand on contrast limited adaptive histogram equalisation (CLAHE) and how we choose its hyperparameters in Section 2.4.1. Next, we present transfer learning and gradient-weighted class activation mapping (Grad-CAM) in Sections 2.4.2 and 2.4.3 respectively.

2.4.1 Contrast Limited Adaptive Histogram Equalisation

Retinal fundus images commonly suffer from uneven illumination and low contrast, resulting in unsatisfactory and poor-quality images [Cao and Li, 2020; Gupta and Tiwari, 2019]. Poor-quality retinal fundus images may lead to difficulty in interpreting signifi-

cant features and structures in the retina [Cao and Li, 2020; Zhou *et al.*, 2017]. Usually, poor-quality retinal fundus images result from the acquisition process which generally involves the type of camera used, the skill of the photographer (or ophthalmologist), the room lighting and the general environment [Cao and Li, 2020]. High-quality retinal fundus images are relevant for clinical purposes and suitable for subsequent accurate computer-aided diagnosis [Gupta and Tiwari, 2019; Zhou *et al.*, 2017], while poor retinal fundus images can affect the sensitivity and specificity of a model [Sahu *et al.*, 2019]. Thus, it is necessary to resolve the challenges related to poor-quality retinal fundus images [Zhou *et al.*, 2017].

In this study, we use contrast enhancement techniques to overcome the challenges related to poor-quality retinal fundus images. Specifically, we use contrast limited adaptive histogram equalisation (CLAHE) to resolve these challenges. The base of CLAHE is histogram equalisation (HE). HE is a nonlinear function [Zuiderveld, 1994] which uses a global approach to adjust image intensities for contrast enhancement. It uses the cumulative distribution function to transform the intensity levels of an image.

For an image f of size $M \times N$, with histogram of intensities denoted as H_f, the normalised image histogram is given by

$$p_f(n) = \frac{H_f(n)}{MN}, \qquad n = 0, 1, \ldots, L-1, \tag{2.4.1}$$

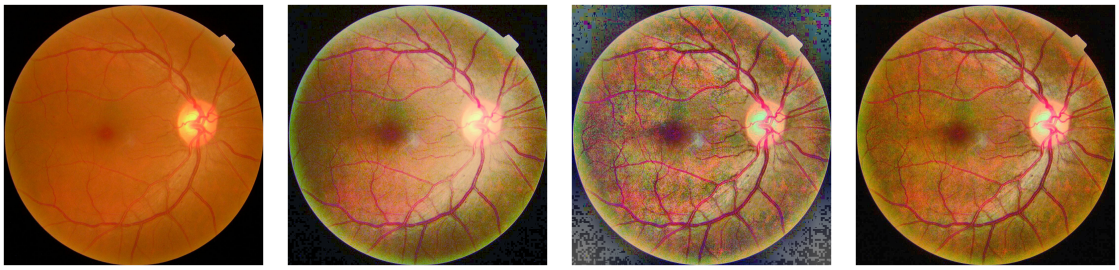
where n represents intensity. L is the maximum intensity level of the image, for example 256 for an 8-bit image. HE of an image is defined by

$$g(n) = \left\lfloor (L-1) \sum_{j=0}^{n} p_f(j) \right\rfloor, \tag{2.4.2}$$

where $g(n)$ is the transformed intensity of the output image corresponding to input intensity n.

Unfortunately, HE can lead to the amplification of noise in images [Zuiderveld, 1994], especially in homogeneous regions [Ma *et al.*, 2018]. The intuition behind this

is that HE attempts to make the intensity range of an image evenly distributed, consequently highlighting subtle differences in homogeneous regions as noise. The local variant of HE is called adaptive histogram equalisation (AHE). First, AHE splits an image into patches of specific dimensions. These patches and their dimensions are known as tiles and grid size respectively. After applying HE on the individual tiles, they are stitched together using bilinear interpolation. In doing so, AHE adjusts image intensities locally and reduces image noise. Even though AHE improves the contrast of an image, it may still have a tendency of noise amplification in relatively homogeneous regions [Ma *et al.*, 2018]. CLAHE is a variant of AHE that resolves the noise amplification problem in contrast enhancement techniques (see Figure 2.6).



(a) **(b)** **(c)** **(d)**

Figure 2.6: Comparing the output of different contrast enhancement techniques applied to the retinal fundus image in Figure 2.6a. In this example, we used the same hyperparameters for all the techniques, namely a grid size of 18×18 and a clip limit of 4. Figure 2.6b shows the impact of applying HE, where the blood vessels are more visible now. The blood vessels and fine details are even more conspicuous in Figure 2.6c, where we applied AHE. Unfortunately, noise is also amplified. This is resolved in Figure 2.6d after applying CLAHE. The blood vessels alongside the optic disk and the macula region are now noticeable.

CLAHE imposes a threshold on the histogram for each local region [Pisano *et al.*, 1998]. This threshold is called a clip limit, and clipped pixels are redistributed equally over the whole histogram. Its purpose is to prevent the local intensity values of a histogram from exceeding the clip limit [Kuran and Kuran, 2021], consequently resolving the problem of noise amplification [Pisano *et al.*, 1998]. One advantage of using CLAHE is that it produces images with high entropy [Kuran and Kuran, 2021]. An im-

age with high entropy is necessary because it can affect the performance of subsequent tasks [Campos *et al.*, 2019]. The process involved in CLAHE is summarised below.

i. Split an image into non-overlapping tiles.

ii. Apply a clip limit on the individual tiles and redistribute clipped pixels to the whole histogram.

iii. Apply HE on each tile.

iv. Merge transformed tiles using bilinear interpolation.

The clip limits and grid size in CLAHE are considered hyperparameters, and the efficiency of CLAHE depends on them. An improper choice of these hyperparameters can produce poor-quality images, but a manual selection of ideal hyperparameters over a wide range of possibilities can be tedious [Campos *et al.*, 2019; Kuran and Kuran, 2021]. The choice of the hyperparameters of CLAHE affects the contrast of the image. The resulting contrast-enhanced image is characterised by visible features and patterns, which are instrumental for CNNs for accurate predictions [**?**]steffens2019contrast). Moreover, it would be beneficial to have different hyperparameters for different images since each image is captured under different conditions.

In this study, we use image entropy to determine proper hyperparameters for each image. Entropy is defined as the measure of the average information content of an outcome [MacKay, 2003]. In the context of an image, entropy is defined as a measure of uncertainty or randomness in the pixel intensities (also known as a distribution) of an image [Dey, 2018]. We employ entropy as a metric to evaluate the performance or assess the quality of CLAHE [Aurangzeb *et al.*, 2021]. Maximum entropy is desired in this context because it suggests information-richness and diversity in the pixel values [Min *et al.*, 2013]. We compute entropy using the following expression:

$$H = -\sum_{i=0}^{255} p_i \log_2(p_i), \tag{2.4.3}$$

where H is the entropy and p_i represents the probability of occurrence of an intensity i in an image (equivalent to the normalised histogram mentioned before). We obtain hyperparameters by finding the maximum curvature on entropy curves. In our case, the entropy curves consist of a plot of the clip limits versus image entropies obtained from changing the hyperparameters. We fit a nonlinear function to find the maximum curvature on the entropy curves. First, we select the grid size with the highest entropies and use its corresponding entropies together with the clip limits as data to fit a nonlinear function of the form

$$f(x) = ae^{-bx} + c, \tag{2.4.4}$$

where x represents clip limits, $f(x)$ the entropy and $a, b,$ and c fitted coefficients.

2.4.2 Transfer Learning

Generally, CNNs require a large amount of data to construct useful latent representations [Tan *et al.*, 2018; Tripuraneni *et al.*, 2020; Zhao, 2017]. In the medical field, especially in medical imaging, large datasets are less readily available [Raghu *et al.*, 2019]. This is due to the high cost of data acquisition and annotation, resulting in insufficient data for deep learning-related tasks [Tan *et al.*, 2018; Zhuang *et al.*, 2020]. We can use transfer learning to resolve the issue of insufficient data because it relaxes the hypothesis that training and test datasets are required to be independent and identically distributed (i.i.d.) [Tan *et al.*, 2018]. Hence, a model (source model) can be trained on data (source data), and we can leverage the weights (knowledge) of that model to the new model (target model) and fine-tune on new data (target data). Figure 2.7 illustrates. For example, the knowledge obtained from training on natural images can be transferred to perform classification tasks in the medical field [Gulshan *et al.*, 2016; Raghu *et al.*, 2019]. The effects of transfer learning are often improved performance, lower training data requirements, and reduced training time [Tan *et al.*, 2018; Zhuang *et al.*, 2020].

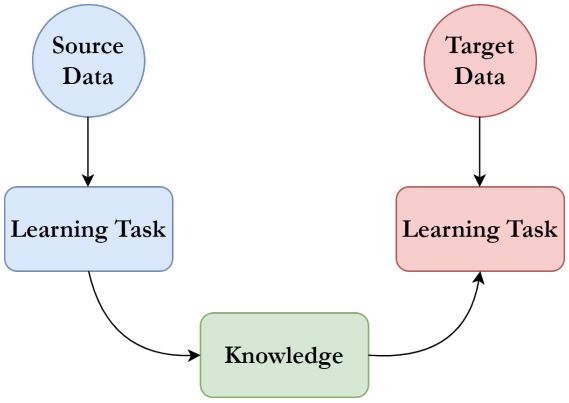


Figure 2.7: Showing the setup for transfer learning. A learning task generates knowledge from source data and transfers the knowledge to another learning task. The new learning task together with the knowledge obtained is used for subsequent tasks such as classification.

In this study, we show the effect of transfer learning by using pre-trained weights from open-sourced state-of-the-art deep learning models. The models considered for the study include ResNet-50 [He *et al.*, 2016], VGG-16 [Simonyan and Zisserman, 2015], Inception-V3 [Szegedy *et al.*, 2015] and InceptionResNet-V2 [Szegedy *et al.*, 2017]. These models are pre-trained on the ImageNet dataset which consists of 1.2 million labelled natural images belonging to 1,000 classes [Deng *et al.*, 2009].

To express transfer learning mathematically, let us consider a domain $\mathcal{D}$ that consists of a feature space $\mathcal{X}$ and a marginal probability distribution $P(X)$ over the feature space, where $X = \{x_1, \ldots, x_n\} \in \mathcal{X}$. Thus $\mathcal{D} = \{\mathcal{X}, P(X)\}$. Then for a domain $\mathcal{D}$, we consider a task $\mathcal{T}$ that consists of a labelled space $\mathcal{Y}$ and a conditional probability distribution $P(Y|X)$ learned from the training data consisting of pairs $x_i \in X$ and $y_i \in Y$. $P(Y|X)$ is the same as the target prediction function $f(x)$. The transfer learning is defined below.

Definition 2.4.1 ([Ruder, 2017*b*])**.** *Given a source domain $\mathcal{D}_S$, a corresponding source task $\mathcal{T}_S$, as well as a target domain $\mathcal{D}_T$ and a target task $\mathcal{T}_T$, the objective of transfer learning is to learn the target conditional probability distribution $P(Y_T|X_T)$ in $\mathcal{D}_T$ with the information gained from $\mathcal{D}_S$, where $\mathcal{D}_S \neq \mathcal{D}_T$ or $\mathcal{T}_S \neq \mathcal{T}_T$. In most cases, the data size of D_S is much larger than the data size of D_T.*

As both the domain $\mathcal{D}$ and the task $\mathcal{T}$ are defined as tuples, the inequalities in Definition 2.4.1 give rise to four transfer learning scenarios.

- $\mathcal{X}_S \neq \mathcal{X}_T$: This suggests that the feature spaces of the source and target domain are different. For example, the source domain may be natural images while the target domain is medical images.
- $P(X_S) \neq P(X_T)$: The marginal probability distributions of the source and target domain are different. For example, the objects of interest for natural images are different from that of medical images.
- $\mathcal{Y}_S \neq \mathcal{Y}_T$: The label spaces between the source and the target are different. For example, the natural images in ImageNet are grouped into 1,000 classes, while the severity of diabetic retinopathy may correspond to five classes.
- $P(Y_S|X_S) \neq P(Y_T|X_T)$: The conditional probability distributions of the source and target tasks are different. For example, the source and target data are unbalanced with respect to their classes.

2.4.3 Gradient-weighted Class Activation Mapping (Grad-CAM)

CNNs demonstrate good performance in computer vision tasks but they are unable to provide intuitive components, making it difficult to explain and interpret their decisions [Selvaraju *et al.*, 2017]. Explainability and interpretability are necessary to increase trust in models. We address explainability and interpretability by generating a coarse localisation map to highlight discriminative regions of an image. We generate the localisation map by using a post-attention technique called gradient-weighted class activation mapping (Grad-CAM) [Selvaraju *et al.*, 2017].

In detail, Grad-CAM generates localisation maps by passing a linear combination of neuron importance weights and feature map activations through a rectified linear unit

(ReLU) [Nair and Hinton, 2010] function. ReLU is used in this case to discard any negative influence on the class of interest. The significance of the neuron importance weight is to assign importance values to the neurons in the layer of interest. It is computed by spatially global average pooling gradients of the activations in the layer of interest with respect to the output score of a class. For a class c, we denote the score of the class as y^c. Also, for a feature map k, we denote feature map activations as $A^k \in \mathbb{R}^{u \times v}$ where u is the width and v is the height of the layer of interest. A^k is indexed by i, j, hence A^k_{ij} represents the location of activations at (i, j). We express the neuron importance weight α^c_k as

$$\alpha^c_k = \overbrace{\frac{1}{Z}\sum_i\sum_j}^{\text{global average pooling}} \underbrace{\frac{\partial y^c}{\partial A^k_{ij}}}_{\text{gradients via backprop}}, \tag{2.4.5}$$

where Z represents the number of pixels of the feature map activations. Next, we compute the Grad-CAM, $L^c_{\text{Grad-CAM}} \in \mathbb{R}^{u \times v}$, as

$$L^c_{\text{Grad-CAM}} = \text{ReLU} \underbrace{\left(\sum_k \alpha^c_k A^k\right)}_{\text{linear combination}}. \tag{2.4.6}$$

2.4.4 Components of the Models

A deep learning model has several components that work together to achieve desirable results. They range from the activation function, the kind of weight initialisation used, the type of optimisation algorithm for learning, the objective function, the set of hyperparameters and the evaluation metric used to assess the performance of the model. The choice of these components affects the performance of a model. In this section, we introduce the various components of deep learning models used in our experiments.

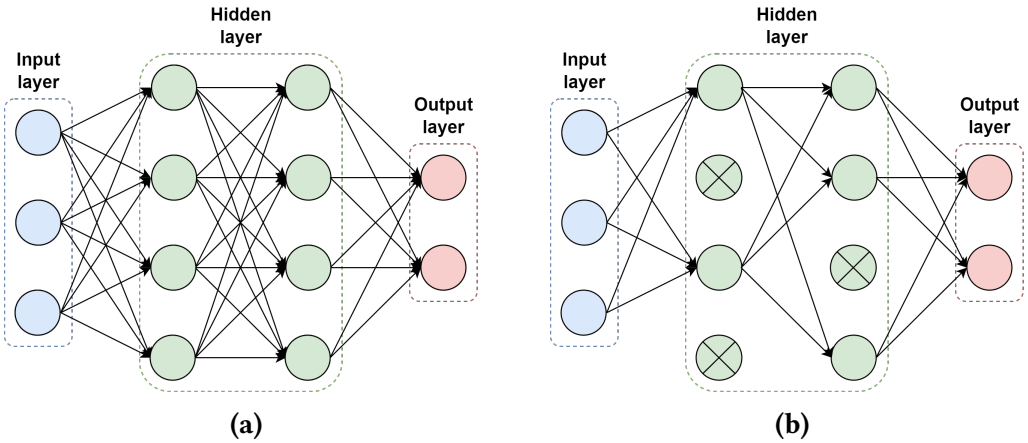
Data splitting. Data splitting is the process of partitioning available data into training, validation and testing sets. These are subsets of the full data created for developing a model, fine-tuning hyperparameters of the model and evaluating the performance of

a trained model. The training set is used to train the model and it usually contains most of the data. The validation set is used to validate and fine-tune hyperparameters for the model to obtain better results. The testing set is only used to test a trained and validated model. These subsets contain unique data points and they do not overlap. A typical split of the available data might be 70% for training, 15% for validation, and 15% for testing.

Batch size and epochs. The batch size is the number of instances that are fed to the model at one time before updating a model's parameters. Ideally, a model should be able to train on all the data at once. However, this can be impossible due to insufficient processing memory. Rather, batches of the dataset are created and fed to the model one at a time. For example, one can create 100 batches (each having a batch size of 50) from a dataset consisting of 5,000 examples. An epoch is when all 100 batches have passed through the model for training.

Dropout. As the name suggests, dropout randomly drops a set of units in a layer with a given probability (see Figure 2.8). The dropped set of units is set to zero, consequently nullifying their contribution to the next layer. For example, in a given layer with 40 units, if the dropout probability is set to 0.5, then 20 randomly chosen units in the layer will be set to zero and the remaining 20 units will proceed to the subsequent layer. By applying dropout to a layer, the units of the layer are forced to avoid sensitive co-dependencies on each other. Hence, dropout can help to reduce overfitting problems in deep learning [Srivastava *et al.*, 2014].

ReLU. In deep learning, activation functions are relevant to introduce nonlinearity to the model. In this study, we use the rectified linear unit (ReLU) as the activation function [Nair and Hinton, 2010]. There are no saturation which could lead to vanishing gradients (a situation where the gradients of a deep learning model approach values close to zero, making parameter updates insignificant) in ReLU except if its in-



(a) (b)

Figure 2.8: Showing an example of a neural network without dropout in Figure 2.8a and a neural network with dropout in the hidden layer in Figure 2.8b. Some neurons (those with crosses) in Figure 2.8b do not have connections to other neurons, indicating that they have been dropped in the current training epoch.

put is negative. It is mostly used because it preserves many properties of a model and generalises well [Goodfellow *et al.*, 2016]. ReLU is defined as

$$f(x) = \max(0, x). \tag{2.4.7}$$

Learning rate. The learning rate is used to regulate the speed at which an optimiser updates the weights of a model. An ideal learning rate should be reasonably low for the model to converge but must be high enough to reduce training time. The choice of the learning rate is important because it can hamper the output of a model. A very small learning rate means the weights are updated by very small amounts, leading to longer training times. On the other hand, a larger learning rate makes a model train faster but could lead to divergence in the weight updates.

Weight initialisation. In deep learning, weights are initialised at the beginning of training and are updated during the training process. Hence, an improper weight initialisation can affect the performance of a model. Training may struggle to converge due to poor weight initialisation [He *et al.*, 2015]. This is because the variance of weights and gradients may explode (tend to infinity) or vanish (become very small) [Glorot and

Bengio, 2010]. To resolve this problem, different distributions have been used to initialise weights, including the truncated normal distribution, Xavier normal distribution and the Kaiming He normal distribution. The truncated normal distribution and the Kaiming He normal distribution are presented below.

The truncated normal distribution is simply a normal distribution with mean μ and variance σ^2 but with a truncated range (a, b). The range used for the experiments in this work is $a = -2$ and $b = 2$.

The Kaiming He normal distribution is a normal distribution with mean 0 and variance $\frac{2}{n^l}$, where n^l is the number of neurons in layer l.

Cross-entropy loss function. In this study, we use the cross-entropy loss function to measure the deviation of the model's prediction from the true values. For k classes in a dataset, the cross-entropy loss function for a single example is defined as

$$\mathcal{L} = -\sum_{i=1}^{k} y_i \log(\hat{y}_i), \tag{2.4.8}$$

where y_i represents the true probability of class i and $\hat{y}_i$ represents the predicted value.

Adam optimiser. Adam is short for adaptive moments. Its main feature is the learning rate adaptation throughout training. Adam incorporates momentum (a technique designed to speed up learning) as an estimate of the first-order moment of the gradients $\boldsymbol{g}$. Next, it introduces bias corrections to both the first- and second-order moments. Finally, it updates the parameters by using the scaled gradients from the moment estimates [Goodfellow *et al.*, 2016]. Adam is efficient and robust, and works well for tasks with large datasets and many parameters [Kingma and Ba, 2014]. The algorithm requires a learning rate denoted by δ, two exponential decay rates for the moment estimates denoted by $\beta_1, \beta_2 \in [0, 1)$ (suggested default: 0.9 and 0.999 respectively), initial parameters denoted by $\boldsymbol{\theta}$ and a small constant denoted by ϵ for numerical stabilisation (suggested default: 10^{-8}). The Adam algorithm is summarised as pseudocode in Algorithm 1.

Algorithm 1: The Adam optimiser [Kingma and Ba, 2014].

Require: Learning rate δ
Require: Exponential decay rates for moment estimates, β_1 and β_2 in $[0, 1)$
Require: Small constant ϵ used for numerical stabilisation
Require: Initial parameter $\boldsymbol{\theta}$
Initialise 1st and 2nd moment variables $\boldsymbol{s} = \boldsymbol{0}, \boldsymbol{r} = \boldsymbol{0}$
Initialise time step $t = 0$
while stopping criterion not met **do**
Sample a mini-batch of m examples from the training set $\{\boldsymbol{x}^{(1)}, \ldots, \boldsymbol{x}^{(m)}\}$ with corresponding targets $\boldsymbol{y}^{(i)}$.
Compute first-order moment of the gradient: $\boldsymbol{g} \leftarrow \frac{1}{m} \nabla_{\boldsymbol{\theta}} \sum_i L(f(\boldsymbol{x}^{(i)}; \boldsymbol{\theta}), \boldsymbol{y}^{(i)})$.
$t \leftarrow t + 1$.
Update biased first moment estimate: $\boldsymbol{s} \leftarrow \beta_1 \boldsymbol{s} + (1 - \beta_1)\boldsymbol{g}$.
Update biased second moment estimate: $\boldsymbol{r} \leftarrow \beta_2 \boldsymbol{r} + (1 - \beta_2)\boldsymbol{g} \odot \boldsymbol{g}$.
Correct bias in first moment: $\hat{\boldsymbol{s}} \leftarrow \frac{\boldsymbol{s}}{1 - \beta_1^t}$.
Correct bias in second moment: $\hat{\boldsymbol{r}} \leftarrow \frac{\boldsymbol{r}}{1 - \beta_2^t}$.
Compute update: $\Delta\boldsymbol{\theta} = -\delta \frac{\hat{\boldsymbol{s}}}{\sqrt{\hat{\boldsymbol{r}}} + \epsilon}$. {operations applied element-wise}
Apply update: $\boldsymbol{\theta} \leftarrow \boldsymbol{\theta} + \Delta\boldsymbol{\theta}$.
end while
return $\boldsymbol{\theta}$

Evaluation metric. The receiver operating characteristic (ROC) curve has become a de facto technique for evaluating the performance of medical models [Hajian-Tilaki, 2013]. It was originally designed to study and analyse the characteristics of radar signals. Since then, ROC has become a relevant tool to evaluate the performance of medical diagnostic models, and has been used extensively in epidemiological studies [Calì and Longobardi, 2015; Hajian-Tilaki, 2013].

A ROC is capable of measuring the degree to which a model separates two classes. It is presented as a plot of the true positive rate (TPR) versus the false positive rate (FPR) at different decision thresholds. TPR is defined as the probability that a positive example is correctly classified as positive, for example, the percentage of DR retinal fundus images correctly classified as diseased images. It is mathematically expressed as

$$\text{TPR} = \text{Sensitivity} = \frac{\text{TP}}{\text{TP} + \text{FN}}, \tag{2.4.9}$$

where TP and FN are the numbers of true positives and false negatives respectively. FPR is defined as the probability that a negative example is incorrectly classified as positive, for example, the percentage of healthy retinal fundus images that are incorrectly classified as diseased images. It is mathematically expressed as

$$\text{FPR} = 1 - \text{Specificity} = \frac{\text{FP}}{\text{FP} + \text{TN}}, \tag{2.4.10}$$

where FP and TN are the numbers of false positives and true negatives respectively. The decision threshold helps map the probability of outputs to a label, and a goal of using the ROC curve is to find the optimal decision threshold that increases the true positives while reducing the false positives [Zou *et al.*, 2007]. A model produces varied outputs with different decision thresholds [Hajian-Tilaki, 2013]. For instance, a small decision threshold classifies most examples as positive, hence, increasing TPR and FPR. Thus, TPR and FPR are both inversely proportional to the decision threshold (see Figure 2.9).

The ROC curve provides a visual overview of the performance of the model at different decision thresholds [Calì and Longobardi, 2015]. It is defined as

$$\text{ROC}(\cdot) = \{(\text{FPR}(c), \text{TPR}(c)), \quad \text{for } c \in \mathbb{R}\}, \tag{2.4.11}$$

where c is the decision threshold. Since $\text{FPR}(c)$ and $\text{TPR}(c)$ range between 0 and 1, we can rewrite Equation 2.4.11 as

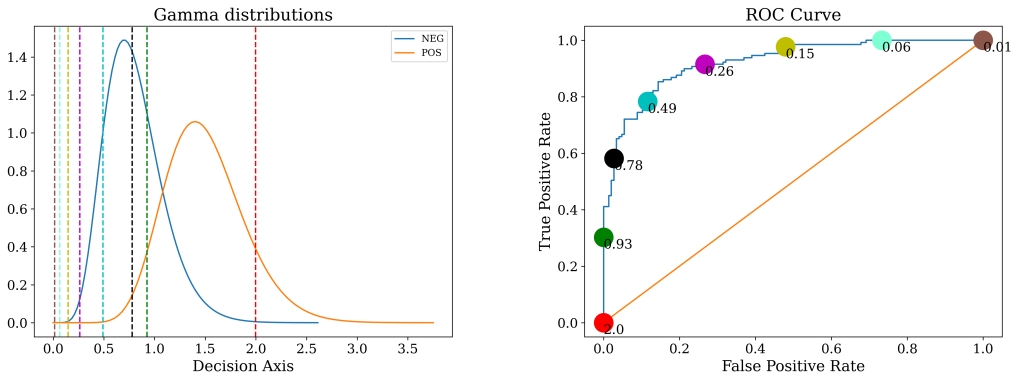
$$\text{ROC}(\cdot) = \{(t, \text{ROC}(t)), \quad \text{for } t \in (0, 1)\}. \tag{2.4.12}$$

The area under the ROC curve (AUC) is a numerical index used to summarise the behaviour of the ROC curve [Calì and Longobardi, 2015]. It is the area underneath the entire ROC curve and it ranges between 0 and 1. Thus, AUC is defined as

$$\text{AUC} = \int_0^1 \text{ROC}(t)\, \mathrm{d}t. \tag{2.4.13}$$

The higher an AUC score, the better the model is at predicting positive classes. An AUC score of 1 indicates a perfect classifier while an AUC score of 0.5 means the model

makes random guesses. Also, an AUC score of 0 means the classifier always predicts a negative example as positive. In Figure 2.9, we see an example of a model obtaining an AUC score of 92%.



(a) **(b)**

Figure 2.9: In this example, we used a logistic regression model to find the optimal decision threshold between a positive (orange) and negative (blue) data distribution. The vertical lines in Figure 2.9a are different decision thresholds. The colours of these vertical lines match the corresponding points on the ROC curve in Figure 2.9b. We see that as the decision threshold increases, the TPR and the FPR decrease.

2.5 Experiments

In this section, we present details of the experiments carried out. First, we give details of the dataset used for the experiments, and the pre-processing and augmentation techniques employed, in Sections 2.5.1 and 2.5.2 respectively. Then, we present the workflow of the experiments for classification and localisation under the model and implementation section (Section 2.5.3), where we also present details of the pre-training and random initialisation techniques used.

2.5.1 Data

We use a publicly available dataset from the Asian Pacific Tele-Ophthalmology Society (APTOS)[1] to classify the severity of diabetic retinopathy (DR). In total, there are 3,662 retinal fundus images in the dataset. The distribution of different levels of severity is highly imbalanced (see Figure 2.10). The dominating class, which is the normal class (that is, retinal fundus images with no DR), represents almost half the dataset (49.29%). The imbalance problem exists also among the remaining classes. The class with the smallest number of images is the severe class (5.27%). We resolve the class imbalance problem by generating weights for the classes and incorporate these weights into the loss function. We describe this solution in more detail in later sections.

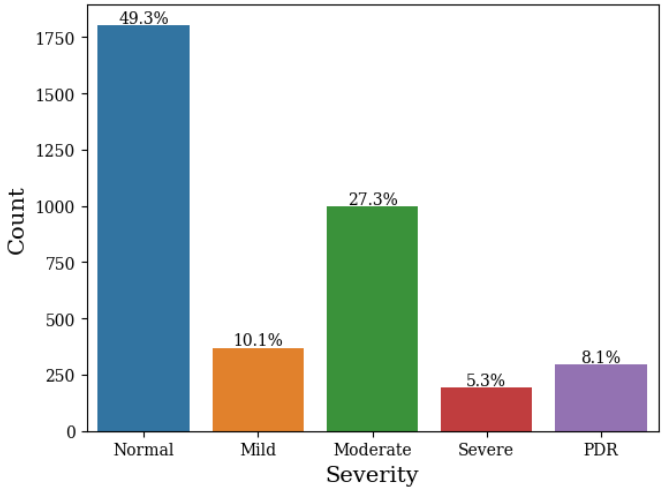


Figure 2.10: Showing the distribution of the various classes in the APTOS dataset.

2.5.2 Pre-processing and Augmentation

During pre-processing, we perform a circle crop and remove black borders in retinal fundus images. Next, we resize all images to 224×224, except when they are used with Inception-V3 which requires an input size of 299×299. Next, we apply CLAHE to all the images. For CLAHE, we estimate optimal hyperparameters for each image.

[1] www.asiateleophth.org

As discussed previously, we desire a combination of hyperparameters with the highest entropy of an image. The clip limit we consider for the study ranges from 2 to 8, and the grid sizes from 2×2 to 32×32.

We show an example and illustrate the impact of CLAHE in Figure 2.11. In this example, we randomly select a retinal fundus image (see Figure 2.11a), which has an entropy of 6.21 and show its corresponding histogram in Figure 2.11b. The histogram peaks at certain colour intensities with a steep slope. This suggests that the image has contrast issues. Next, we compute the entropies for the various hyperparameters and plot the obtained entropies against the clip limit in Figure 2.11c. We observe in this example that a grid size of 2×2 has the highest entropy. Finally, we fit a nonlinear function to find the maximum curvature on the highest entropy curve (see Equation 2.4.4). In this example, a grid size of 2×2 and a clip limit of 3.1 produce the highest entropy of 6.78 (Figure 2.11e). We notice that there are fewer significant peaks in the colour intensity histogram after applying CLAHE, and it is now more uniform (Figure 2.11f). Moreover, we see that the image quality in Figure 2.11e has significantly improved. In addition, the optic disk, the macula region and the blood vessels are more visible than in the original image.

For augmentation, we randomly flip the images horizontally and vertically. Also, we randomly rotate the images by $30°$, and randomly jitter the brightness, hue, saturation and contrast. These augmentation techniques are applied only to the training dataset. Lastly, we normalise the pixel values of images in the training, validation and testing sets to be between -1 and 1.

2.5.3 Models and Implementation

In this section, we compare the performance of four state-of-the-art CNN models in predicting the severity of diabetic retinopathy. The models include ResNet-50, VGG-16, Inception-V3, and InceptionResNet-V2. We compare the performance of these models when trained from a random initialisation of weights (see Section 2.5.3.2 for details)

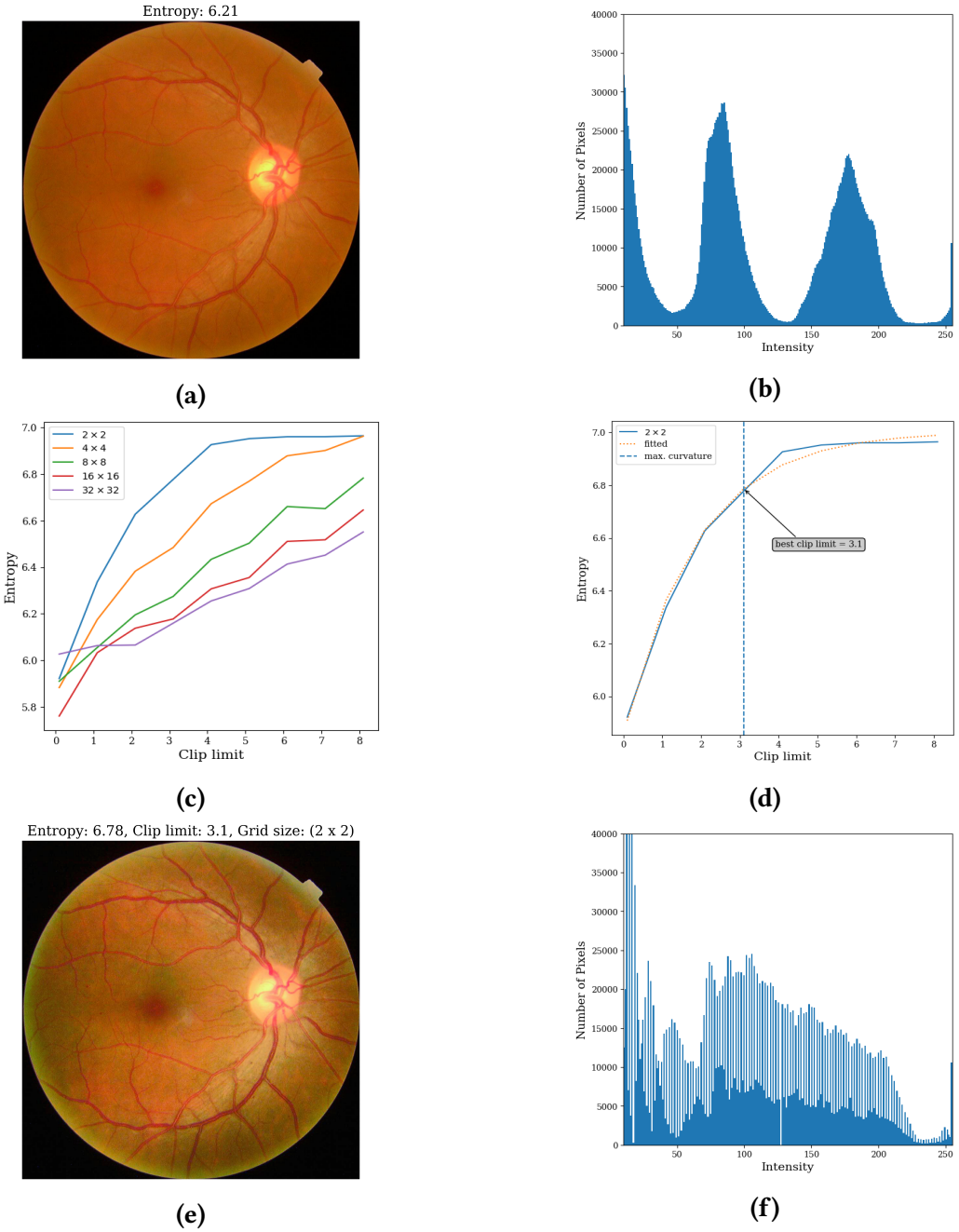


Figure 2.11: The impact of CLAHE on an example of a retinal fundus image (Figure 2.11a) with contrast issues (Figure 2.11b). In this example, we compute entropies for various hyperparameters (Figure 2.11c) and find that 2×2 grid size and a clip limit of 3.1 results in the highest entropy (6.78) (Figure 2.11d), which significantly improves image quality (Figure 2.11e) and enhances uniform distribution in the histogram (Figure 2.11f). The reader is referred to the text in Section 2.5.2 for details.

and when transfer learning is applied (see Section 2.5.3.1 for details). For each model, we replace the final layer with a global average pooling layer, a dropout layer with 50% dropout probability and a linear layer with five units (each representing a class in the dataset) for classification. Finally, we also apply Grad-CAM to the extracted feature maps (see Figure 2.12).

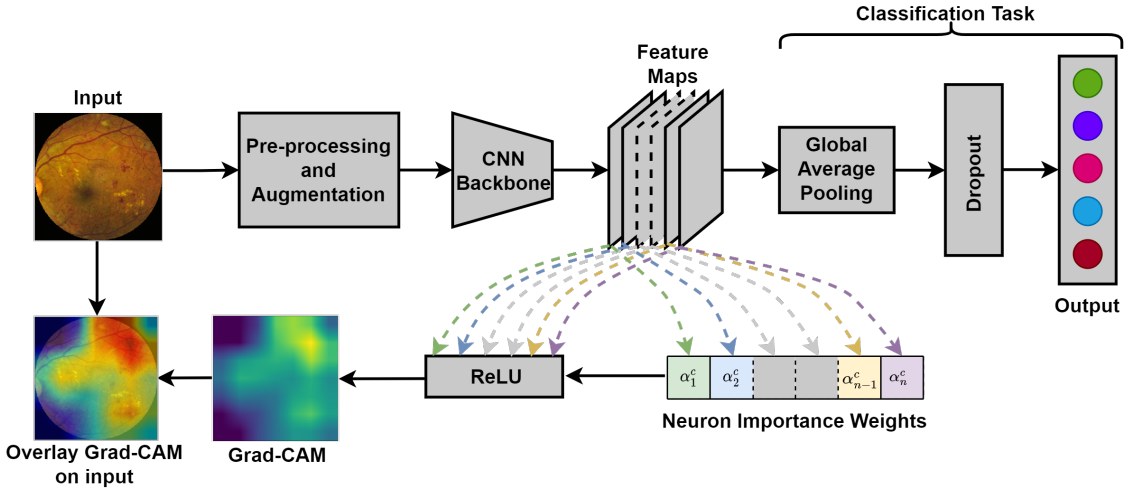


Figure 2.12: The general layout of our model for DR classification and lesion localisation. First, retinal fundus images are fed to pre-processing and augmentation techniques. Next, the images are fed to a CNN backbone for feature extraction. The CNN backbones considered in this study include ResNet-50, Inception-V3, VGG-16, and InceptionResNet-V2. Next, we classify the severity of DR using the extracted features. The output layer in this diagram has five units because there are five classes in the dataset. Finally, we use the extracted features also to generate coarse localisation maps by Grad-CAM.

Regarding implementation, we feed CNN models with pre-processed and augmented retinal fundus images to obtain feature maps. We then feed the feature maps to subsequent layers for classification. During the localisation process, we use the feature maps to generate neuron importance weights, multiply them with the feature maps and pass the results through a ReLU function to obtain localised outputs. Finally, we overlay the output of the Grad-CAM on the retinal fundus image to visualise discriminative regions.

2.5.3.1 Pre-training and Fine-tuning

We split the dataset into training, validation and testing sets. Specifically, we use 70% of the dataset to train the model, and 15% for validation. After training, we test the model using the remaining 15% of the data. We train each model with a batch size of 32 images. We use the Adam optimiser [Kingma and Ba, 2014] to optimise each model with a learning rate of 10^{-4} and categorical cross-entropy as the loss function. We initialise the models with their corresponding pre-trained weights from the ImageNet dataset. Finally, we fine-tune each model on the APTOS dataset for 50 epochs. We save the model with the best validation AUC (refer to Section 2.4.4).

2.5.3.2 Random Initialisation

For random initialisation, we maintain the same split partition of the data, batch size, learning rate, optimiser and the loss function as above. We initialise ResNet-50 and VGG-16 with Kaiming He normal initialisation [He *et al.*, 2015] and truncated normal initialisation [Burkardt, 2014] for Inception-V3 and InceptionRes- Net-V2. Even though we initially maintain the learning rate, we halve it after training the model for 2,000 iterations (24 epochs). In total, we train each model for 100 epochs. Similar to the above, we save the model with the best validation AUC metric.

2.6 Results

In this section, we present the results obtained before and after applying transfer learning to four different models trained on the APTOS dataset (Section 2.6.1). We aim to identify which model best predicts DR severity from retinal fundus images. We also report on the impact of CLAHE and show some results of Grad-CAM (Sections 2.6.2 and 2.6.3). We evaluate the models using the testing dataset and employ AUC as a performance metric.

2.6.1 Impact of Transfer Learning

We observe in Table 2.1 that the four models all perform similarly. For all the models, we note that models initialised with pre-trained weights outperform their randomly initialised counterparts. Interestingly, we performed only 50 epochs on the pre-trained models, which is half the number of epochs (100) used when training models initialised with random weights. This suggests that transfer learning requires less time to attain better results.

Table 2.1: Evaluating the performance of the models using AUC.

Model	Training		Validation		Testing	
	Random Init.	Pre-trained	Random Init.	Pre-trained	Random Init.	Pre-trained
ResNet-50	95.09	**97.48**	94.40	**96.87**	94.96	**96.84**
Inception-V3	92.92	**98.10**	93.32	**96.91**	93.89	**96.19**
VGG-16	96.99	**97.65**	94.50	**96.99**	95.70	**96.40**
InceptionResNet-V2	98.22	**98.27**	95.46	**96.71**	95.54	**96.20**

Bold digits in each partition highlight the best performing technique.

2.6.2 Impact of CLAHE

The dataset used in Section 2.6.1 was only resized and augmented. In this section, we additionally pre-process the images with CLAHE, before training or fine-tuning the models. We observe in Table 2.2 that there is further improved performance after applying CLAHE. We see that all the models achieve an increase in their AUC scores, except for the validation AUC in the InceptionResNet-V2 model. Moreover, we observe that the ResNet-50 model outperforms the rest of the models by attaining the best AUC scores in both the validation and testing sets.

It is important to note that in this experiment we fine-tune the models for only 20 epochs, which is fewer than half the number of epochs in the previous experiments when we did not apply CLAHE. This reveals that applying CLAHE further reduces the

time necessary to train a model, and improves performance. This may be explained by the fact that the application of CLAHE results in high entropy values, consequently reducing the effort required to find important areas of an image.

Table 2.2: Evaluating performance of the models after applying CLAHE.

Model	Training	Validation	Testing
ResNet-50	98.44	97.15	97.40
Inception-V3	98.27	97.01	96.84
VGG-16	97.50	96.91	96.76
InceptionResNet-V2	98.89	96.54	96.80

Highlighted row represents the best performing model.

2.6.3 Visualising Localisation Maps

In this section, we visualise coarse localisation maps generated by Grad-CAM for the various classes. After training, we use the feature maps from the trained model to localise discriminative regions in the images. Specifically, we use feature maps from ResNet-50 fine-tuned on the APTOS dataset, since it performs best among all models evaluated (see Table 2.2). Figure 2.13 shows Grad-CAM applied on randomly sampled retinal fundus images. In these examples, we observe that Grad-CAM is able to identify important regions in the images. For the diseased classes, we observe that the model highlights the lesions in the retina images. Since a normal class (non-diseased class) has no lesions, the model tends to highlight regions such as the optic disc and the blood vessels.

2.7 Conclusion

In summary, we presented a process to classify the DR severity of, and localise lesions in retinal fundus images. First, we evaluated the performance of different models in predicting DR severity. We noted similar results between the models trained from scratch,

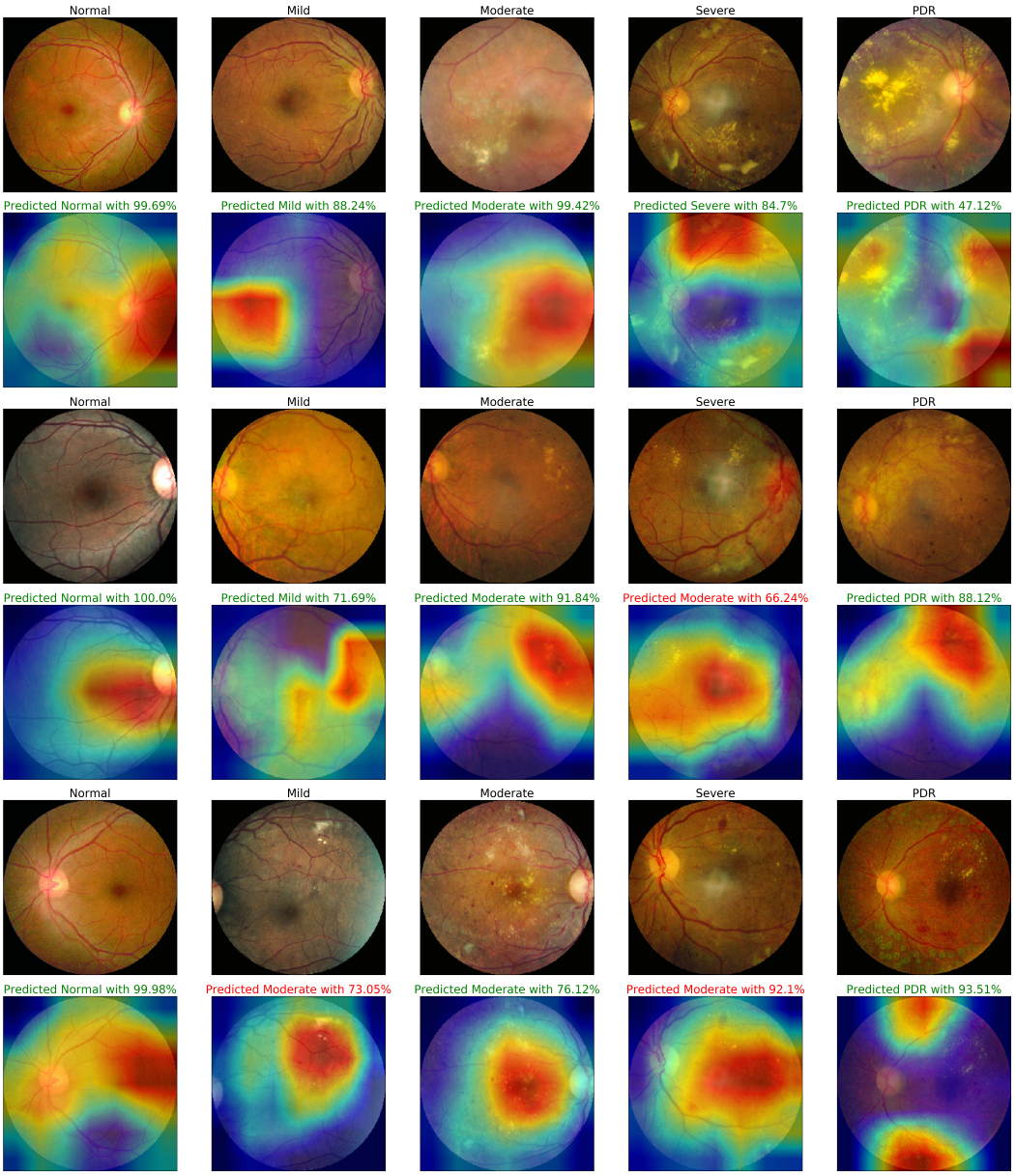


Figure 2.13: Showing the results of localisation maps generated by Grad-CAM on randomly selected retinal fundus images using Grad-CAM. Each column represents a class in the data: normal, mild, moderate, severe and proliferative diabetic retinopathy.

and the pre-trained models. However, the results revealed remarkable performance for models initialised with pre-trained weights. Besides, we trained the models for only 50 epochs when we applied transfer learning, which is half the number of epochs when we trained from scratch. Initially, we used only the augmented datasets for the experi-

ments.

Next, we employed CLAHE as a pre-processing technique to resolve contrast issues in retinal fundus images, and later augmented them for training. After applying CLAHE, we observed improved performance in the models. The number of epochs further reduced to 20, showing the benefits of CLAHE for this task. We noted this as a positive direction towards developing a fast and robust computer aided diagnostic tool.

Furthermore, the ResNet-50 model attained the best performance among the models tested, with an AUC score of 97.15% and 97.40% on the validation and testing sets respectively. We argue that ResNet-50 outperforms the other models because of its skip connection feature which increases the convergence speed and better learns lower semantic information [Bello *et al.*, 2021]. Hence, we used feature maps from the ResNet-50 model to generate coarse localisation maps, in order to visually explain the decisions made by the model. We noticed that the localisation maps generated were mainly over lesions present in the retinal fundus images.

Chapter 3

Prediction of Cardiovascular Risk Factors

3.1 Overview

This chapter aims to create a multi-task learning model to predict risk factors for cardiovascular disease using retinal fundus images. Using the ResNet-50 model as a backbone, we compare the performance of two kinds of multi-task learning (MTL) models. The dataset we use in this chapter contains several attributes, such as the view of the retina, the patient's ethnicity, the camera used to capture the retina, and many others. As we observe towards the end of the chapter, the models perform better when we use retinal fundus images that contain both the optic disc and the macula region. Also, we find significant evidence of discriminating regions in the retina that separate males and females. In the rest of this chapter, we first present background for the study in Section 3.2. Next, we present related work and our methodology in Section 3.3 and Section 3.4, respectively. This is followed by Section 3.5, which details the experiments carried out. We present the results obtained in Section 3.6 and conclude in Section 3.7.

3.2 Background

Cardiovascular diseases (CVD) are a group of disorders involving the heart and blood circulation system, such as hypertension, stroke, coronary heart disease, and peripheral vascular disease [Salkind, 2005]. According to the World Health Organisation, they are the leading causes of death globally and have contributed to an estimated 17.9 million deaths in 2019 [WHO, 2021]. CVD continues to be on the rise worldwide, especially in low-income countries. Thus, they are considered a global burden of disease [Roth *et al.*, 2020].

There are factors including age, sex, blood pressure, smoking status, and diabetes that significantly contribute to the development of CVD [Poplin *et al.*, 2018]. These factors are known as cardiovascular risk factors (CVFs). There are several algorithms available to assess the risk of a CVD event occurring in an individual [D'Agostino Sr *et al.*, 2008]. An example is the well-known Framingham risk score, which incorporates several CVFs into a Cox proportional hazard regression model to estimate the 10-year cardiovascular risk of an individual [D'Agostino Sr *et al.*, 2008; Jahangiry *et al.*, 2017].

Studying CVFs improves our understanding of the early development of CVD. Moreover, different diseases can be observed on the retina as they follow a unique pathophysiological process. By monitoring and observing the retina, scientists are able to identify early signs of different diseases, including CVD [Nguyen and Wong, 2009]. This has consequently created new avenues to explore and understand the pathophysiology of CVD [MacGillivray *et al.*, 2014].

Major advancements in retinal imaging technology have made it possible for scientists to monitor and observe the retina. Retinal imaging technology uses special cameras or sensors to capture the retina in a multi-spectral approach or at a high spatial resolution. For example, the cameras used for this study capture the retinal fundus images in a 2D high spatial resolution. They include Canon, Centrevue, Crystalvue, Optovue, Topcon, and Zeiss Visucam fundus cameras. These advancements have resulted in pre-

cise in vivo observation of the retina and have made it possible to predict CVFs using retinal images [Nguyen and Wong, 2009].

Even though imaging technologies have advanced, retinal fundus image analysis require specialists [Date *et al.*, 2019]. Prediction of risk factors such as age and sex may not be clinically relevant but may provide new retinal insights that are not apparent to specialists [Korot *et al.*, 2021]. In recent times, several techniques have been introduced to automatically analyse retinal fundus images. Among them are deep learning models, which are mainly applied to tasks such as classification [Mensah *et al.*, 2021; Poplin *et al.*, 2018], vascular segmentation [Jin *et al.*, 2019; Oliveira *et al.*, 2018; Zhang *et al.*, 2019], recognition [Li *et al.*, 2019; Mo *et al.*, 2018], and so on.

In this study, we predict CVFs from retinal fundus images using a multi-task deep learning model. Specifically, we use only age, sex, and hypertension as CVFs for the study because of limited information available in the data for the other factors. In detail, we use retinal fundus images for the study because the retina is a unique part of the body that allows for non-invasive observation of the retinal vasculature relating to the development of CVD [Liew and Wang, 2011; MacGillivray *et al.*, 2014; Nguyen and Wong, 2009; Zhang *et al.*, 2020]. In summary, we build a multi-task deep learning model to predict CVFs, and we perform an in-depth analysis of attributes in the data.

3.3 Related Work

In this section, we present related work on risk factor classification and multi-task learning models employed for medical tasks.

3.3.1 Risk Classification

Several risk factor prediction models, such as the Framingham risk score [D'Agostino Sr *et al.*, 2008], the Reynolds risk score [Ridker *et al.*, 2007, 2008], and the SCORE (systematic coronary risk evaluation) [Mach *et al.*, 2019], were used to calculate the risk of CVD

occurring. These models mostly use regression models and several CVFs as variables to compute the risks [Goldstein *et al.*, 2017]. Even though these models have become standard for predicting cardiovascular risks, the advent of machine learning has opened new avenues of research. For example, Kakadiaris *et al.* [2018] developed a risk factor calculator based on support vector machines and 13-year follow-up data. Goldstein *et al.* [2017] used different classification tree algorithms, nearest neighbour algorithms, and neural networks to calculate the cardiovascular risk of an individual. Weng *et al.* [2017] compared the performance of cardiovascular risk prediction using four machine learning algorithms with an established algorithm, and observed better results.

Inspired by the results obtained from machine learning, several studies have developed deep learning models to predict CVFs. The most notable study done on predicting CVFs is by Poplin *et al.* [2018], who based their model on the Inception-V3 to predict CVFs from retinal fundus images. However, it is not mentioned whether each CVF was independently predicted or whether they were simultaneously predicted. Although several studies have predicted single CVFs, none provide a multi-task learning approach [Betzler *et al.*, 2021; Dieck *et al.*, 2020; Korot *et al.*, 2021; Nusinovici *et al.*, 2022; Yamashita *et al.*, 2020; Zhang *et al.*, 2020]. In our study, we develop a multi-task learning model based on ResNet-50 to predict various CVFs simultaneously.

3.3.2 Multi-Task Learning

Crawshaw [2020] compared multi-task learning (MTL) to the way humans learn and perform multiple tasks. MTL has been applied to computer vision [Thung and Wee, 2018; Vu *et al.*, 2019], natural language processing [Chen *et al.*, 2021; Collobert and Weston, 2008], and speech recognition [Krishna *et al.*, 2018; Pironkov *et al.*, 2016; Shinohara, 2016]. For example, Vu *et al.* [2019] used MTL to predict age and gender using a dataset that contained human faces. Lee and Liu [2021] developed an MTL model to improve path prediction in autonomous driving.

In the medical domain, MTL has been employed on medical images [Zhao *et al.*,

2022]. Moeskops *et al.* [2016] employed MTL to segment three datasets, namely brain magnetic resonance imaging (MRI), breast MRI, and cardiac computed tomography angiography. Xie *et al.* [2022] studied weakly supervised medical image segmentation based on MTL. Zhou *et al.* [2021] used MTL to segment and classify tumours for 3D automated breast ultrasound images. Regarding retinal fundus images, Pascal *et al.* [2022] used MTL for glaucoma detection, while Ayhan *et al.* [2023] developed an MTL model for activity detection in neovascular age-related macular degeneration. Again, in our study, we employ MTL based on ResNet-50 to predict CVFs.

3.4 Methodology

In this section, we present details of our multi-task learning model and the two approaches it uses to share representations (Section 3.4.1). We then present the evaluation methods used for this study (Section 3.4.2).

3.4.1 Multi-Task Learning Models

Multi-task learning (MTL) is a technique where several tasks are simultaneously learned using a shared model [Crawshaw, 2020]. The aim of an MTL model is to encourage domain-specific representations between related tasks, thereby enforcing generalisation, reducing the risk of overfitting, and improving performance [Crawshaw, 2020; Dobrescu *et al.*, 2020; Ruder, 2017*a*; Zhang and Yang, 2018] (see Figure 3.1). MTL is considered a special case of transfer learning because there is no distinction between tasks [Dobrescu *et al.*, 2020].

An MTL model learns shared, common space parameters or representations of related tasks by using a hard or soft approach. The most used and well-known among the two is hard-parameter sharing (HPS) [Ruder, 2017*a*]. HPS shares parameters between all tasks such that each parameter is trained to jointly minimise multiple loss functions for the corresponding tasks [Crawshaw, 2020; Ruder, 2017*a*] (see Figure 3.2a).

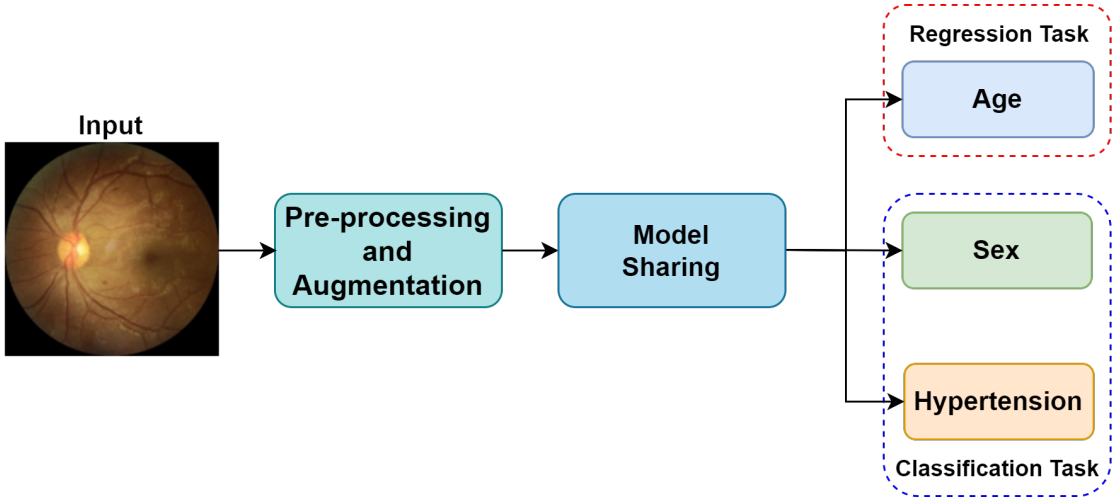


Figure 3.1: The pipeline of the proposed MTL model. Our model takes pre-processed retinal fundus images as input and returns predictions of a patient's age, classification of a patient's sex and their hypertension status.

Soft-parameter sharing (SPS) on the other hand has a separate model, with separate parameters, for each task [Crawshaw, 2020; Ruder, 2017*a*]. Thus, SPS successfully learns representations by introducing a constraint that penalises the distance between the individual models [Ruder, 2017*a*] (see Figure 3.2b).

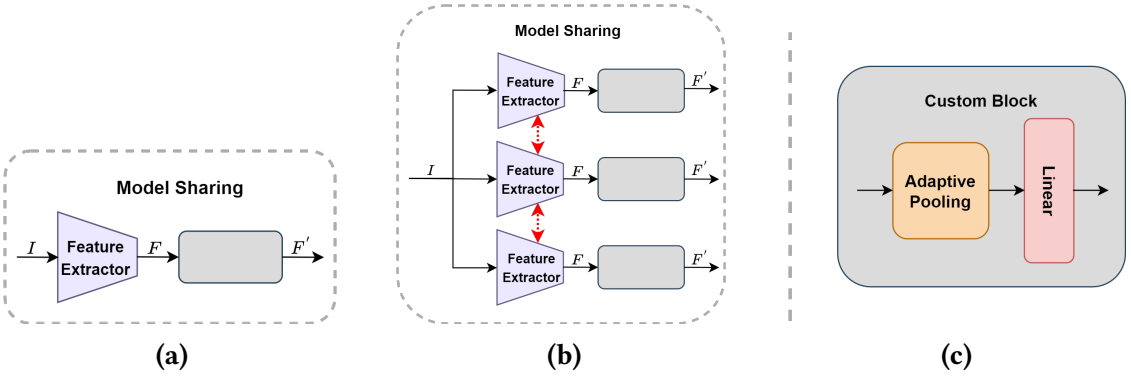


Figure 3.2: Depicting the various multi-task learning techniques considered in this study. A hard-parameter sharing technique uses a single feature extractor for predictions (Figure 3.2a) while a soft-parameter sharing technique uses independent feature extractors (Figure 3.2b). Figure 3.2c shows the custom block used in the models.

Consider a dataset with N retinal fundus images, and each with T risk factors (tasks). Let us denote the dataset as $\mathcal{D} = \{\mathcal{X}, \mathcal{Y}\}$, where $\mathcal{X} = \{X_i\}_{i=1}^{N}$ with $X_i \in \mathbb{R}^d$, and

$\mathcal{Y} = \left\{ \{y_i^j\}_{j=1}^T \right\}_{i=1}^N$ such that $y_i^j \in \mathbb{R}$ if task j is a regression task and $y_i^j \in \{0,1\}$ if task j is a binary classification task. In MTL, we seek to minimise the error function

$$\underset{\{W^j\}_{j=1}^T}{\operatorname{argmin}} \sum_{j=1}^{T} \sum_{i=1}^{N} \mathcal{L}_j(y_i^j, \mathcal{F}(X_i, W^j)), \tag{3.4.1}$$

where $\mathcal{F}(\cdot,\cdot)$ is the model with input X_i and weight matrix W^j, $\mathcal{L}_j(\cdot,\cdot)$ is the loss function for task j which we configure as the mean squared error (MSE) for regression tasks (age) and cross entropy (see Equation 2.4.8) for classification tasks (sex and hypertension). The MSE is defined as

$$\mathcal{L}_{\text{MSE}} = \frac{1}{N} \sum_{i=1}^{N} (y_i - \hat{y}_i)^2, \tag{3.4.2}$$

where N represents the total number of instances, y_i denotes the ground truth values, and $\hat{y}_i$ the predicted values. For $T = 1$, MTL is reduced to single task learning. Regarding SPS in a two-task setting with task A and B, we seek to minimise the loss

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{main}} + \lambda \sum_{L} \left\| W^{(A)} - W^{(B)} \right\|_F^2, \tag{3.4.3}$$

where $\mathcal{L}_{\text{total}}$ and $\mathcal{L}_{\text{main}}$ are the overall loss and the sum of losses for all the tasks, respectively. The second term in Equation 3.4.3 represents the constraint used to penalise the distance between the parameters of the individual models. The factor $\lambda > 0$ in the second term is a hyperparameter controlling the relative importance of the second term, and L represents the corresponding layers for the individual models. $W^{(A)}$ and $W^{(B)}$ represent the individual parameters for the corresponding tasks, and $\|\cdot\|_F^2$ represents the squared Frobenius norm [Duong *et al.*, 2015].

3.4.2 Evaluation Metrics

The metric used to measure the performance of our models is the accuracy evaluation metric. In particular, this metric is used to evaluate the classification components (that

is, sex and hypertension) of the MTL models. Later, we use precision, recall and the F1-score to evaluate the performance of other attributes of the data on sex classification. We present the definitions for the various evaluation metrics below.

Accuracy. Accuracy measures the percentage of the number of correct predictions. It is defined as

$$\text{Accuracy} = \frac{\text{correct classifications}}{\text{all classifications}} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}. \tag{3.4.4}$$

Precision. Precision measures the proportion of positive classifications that are actually positives. It is defined as

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}. \tag{3.4.5}$$

Recall. Recall measures the proportion of actual positives that were correctly classified:

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}. \tag{3.4.6}$$

F1-Score. The F1-score computes an evaluation score from the harmonic mean of precision and recall metrics:

$$\text{F1-Score} = 2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}}. \tag{3.4.7}$$

3.5 Experiments

In this section, we present details of the dataset used for the study in Section 3.5.1. Next, we present the pre-processing and augmentation techniques used in Section 3.5.2. Finally, we present the details of the multi-task learning models and how they are implemented in Section 3.5.3.

3.5.1 Data

In this study, we use the Eye Picture Archive Communication System (EyePACS)[2] dataset which contains eye-related patient metadata and retinal fundus images. The metadata includes patients' hypertension status, their ethnicity, and the kind of camera used to capture the retinal fundus image. In total, the EyePACS dataset contains 476,545 images obtained from 42,296 patients. Some information on the age, sex and hypertension status of patients is missing. In detail, there are 1,591 missing entries for patients' sex, 54 for their age and 18,379 for their hypertension status. 41.63% of the 40,705 patients are males and 58.37% are females. The average age is 57.51 years with a standard deviation of 11.35. Patients are labelled either as having *controlled hypertension* or as having *no hypertension*. The controlled hypertension group dominates the hypertension status of patients with a percentage of 62.40% compared to 37.60% for the no hypertension class. Regarding ethnicity, a little over half the number of patients are Latin American. The remaining ethnic groups (Asians, Caucasians, Indian subcontinent, Multi-racial and Native Americans) make up just under half the number of patients (see Table 3.1).

There are at least six retinal fundus images for each patient, captured between the years 2013 and 2021. Three images are captured for the left eye, and three for the right, and each is taken from a particular field of view of the retina. The first field of view, named *Field 1*, shows both the optic disc and macula region at the centre of a retinal fundus image (Figure 3.3a). The second, named *Field 2*, shows only the optic disc at the centre of a retinal fundus image (Figure 3.3b). The third field of view, named *Field 3*, shows only the macula region at the centre of a retinal fundus image (Figure 3.3c). The data is made up of 142,856 *Field 1*, 145,470 *Field 2* and 188,219 *Field 3* images. These sum up to 476,545 retinal fundus images in total (as indicated in Table 3.1).

There are two types of fundus cameras used to capture the images in the EyePACS dataset, namely desktop fundus cameras and handheld fundus cameras. Desktop fundus cameras are traditional desktop-mounted retinal fundus cameras. They produce

[2] www.eyepacs.com

Table 3.1: Descriptive statistics for the data used in the study.

Description	Value
Number of patients	42,296
Number of images	476,545
Age in years: mean (SD)	57.51 (11.35)
Sex	
Male	16,945
Female	23,760
Hypertension	
yes	12,097
no	6,840
Ethnicity	
African Descent	2,439
Asian	1,754
Caucasian	3,524
Indian subcontinent	1,518
Latin American	21,529
Multi-racial	266
Native American	274

The number of patients in the individual tasks (risk factors) do not add up to the total number of patients due to missing information.

(a) **(b)** **(c)**

Figure 3.3: Showing examples of the three fields of view from one side of the eye of the same patient. Field 1 (Figure 3.3a) shows both the optic disc and macula region, Field 2 (Figure 3.3b) shows only the optic disc, and Field 3 (Figure 3.3c) shows only the macula region.

high-quality images but are expensive, bulky and require a specialist to operate [Chalam *et al.*, 2022; Panwar *et al.*, 2016; Rajalakshmi *et al.*, 2021]. Handheld fundus cameras,

as the name suggests, are handheld, portable, lightweight, inexpensive and require minimal training to operate [Panwar *et al.*, 2016]. More recently, handheld fundus cameras have improved in their image quality [Chalam *et al.*, 2022]. All the fundus cameras reported in this study are non-mydriatic and have a 45° field of view, except for Volk Pictor Plus which has a 40° field of view. Due to differences in operating fundus cameras, the camera information is categorised into several quality levels. The quality levels of the retinal fundus images used for the study include *Adequate*, *Good*, and *Excellent*. We show the distribution for the camera quality levels in Table 3.2.

Table 3.2: Presenting image quality from the various cameras used in EyePACS.

Camera	Type	Adequate	Good	Excellent	Total (%)
Canon CR1	D	14,244	14,946	10,738	39,928 (8.63%)
Canon CR2	D	23,834	25,800	15,102	64,736 (13.99%)
Canon DGi	D	15,853	17,612	7,638	41,103 (8.88%)
Centrevue DRS	D	8,652	4,463	1,090	14,205 (3.07%)
Crystalvue NFC 700	D	3,859	2,574	684	7,117 (1.54%)
Optovue Vivicon	D	379	827	655	1,861 (0.40%)
Optovue iCam	D	65,412	67,974	21,465	154,851 (33.47%)
Topcon NW 200	D	1,999	2,184	1,630	5,813 (1.26%)
Topcon NW 400	D	54,687	38,769	18,428	111,884 (24.18%)
Topcon NW 700	D	4,098	3,147	384	7,629 (1.65%)
Volk Pictor Plus	H	9,832	80	0	9,912 (2.14%)
Zeiss Visucam	D	2,987	666	16	3,669 (0.79%)

For the camera type, D represents **D**esktop and H represents **H**andheld.

3.5.2 Pre-processing and Augmentation

Again, we employ similar pre-processing and augmentation techniques as in the previous chapter (see Section 2.5.2). The main difference here is the image size. The EyePACS retinal fundus images are roughly 3000×3000 pixels [MacGillivray *et al.*, 2014]. In our experiments, we resize the images to 512×512.

3.5.3 Model and Implementation

Besides experimenting on the multi-task learning models, we also create single-task learning models for individual tasks. This is done in order to compare the performance of the single-task and multi-task learning models. We modify the ResNet-50 model by replacing its final layer and the penultimate layer with a custom block of layers. In detail, the custom block consists of an adaptive pooling layer and a linear layer with 512 units. We feed the output of the custom block to the task-specific output layers for the various cardiovascular risk factors considered in this study.

We train the models using the Adam optimiser with a batch size of 32. We set the learning rate to 10^{-4} and train each model for 50 epochs. We save the best models depending on whether there is an improvement in either validation loss or validation accuracy. That is, if the new validation accuracy of an epoch is the same as the best validation accuracy, we save the one with the lowest validation loss.

3.6 Results

We simultaneously predict cardiovascular risk factors (CVFs) using a multi-task learning (MTL) model. In detail, we use ResNet-50 as a backbone to generate shared representations for the prediction of the CVFs [Ayhan *et al.*, 2023]. First, we explore single-task models trained to predict CVFs independently. We start with *Field 1* retinal fundus images as they contain both the optic disc and the macula region, which are the most central parts of the retina [Kolb, 2011], in Section 3.6.1. Later, we explore the performance of the MTL model on *Field 2* and *Field 3* images in Section 3.6.2. Finally, we present the performance of some attributes in the data in Section 3.6.3.

3.6.1 Results for Field 1

In a supervised learning setting, a multi-task learning model requires the data to possess information on all the various tasks involved. For example, if an instance has information about sex and hypertension status but has missing information about age, then the model will be unable to verify its prediction on age for that instance. Due to the missing information mentioned in Section 3.5.1, we can only use 65,877 *Field 1* retinal fundus images for the multi-task learning experiments. For the single-task model, there are 76,887 additional records for the age data, 72,023 for the sex data and 1,463 for the hypertension data.

Table 3.3: Performance of single-task learning and multi-task learning for *Field 1* images.

Model	Training			Validation			Testing		
	Age (MSE)	Sex (Acc. %)	Hyp. (Acc. %)	Age (MSE)	Sex (Acc. %)	Hyp. (Acc. %)	Age (MSE)	Sex (Acc. %)	Hyp. (Acc. %)
STL	0.0016	89.38	77.22	0.0021	87.61	70.24	0.0022	87.30	68.73
MTL - HPS	0.0041	91.95	80.57	0.0054	87.35	70.39	0.0039	85.55	68.15
MTL - SPS	0.0029	91.52	73.50	0.0274	86.59	69.81	0.0286	85.37	68.21

We observe on-par performance between the models even though more retinal fundus images are used to train the single-task models (Table 3.3). The multi-task learning models require less data to achieve on-par results, thus confirming the benefits of using a multi-task model [Zhang *et al.*, 2020]. It is important to note that we scaled the age target values by dividing them by 100, which resulted in small MSE values. We observe that the hard-parameter sharing (HPS) model performs better than the soft-parameter sharing (SPS) model. A possible explanation for this is that the number of parameters in the SPS model exceeds that of the HPS by a factor of about 3, consequently reducing the possibility of overfitting in the HPS model[Rosenfeld and Tsotsos, 2019; Zhang *et al.*, 2020]. Generally, the models achieve promising results for the age and sex tasks, with a slight drop in performance for the hypertension task.

We visualise randomly selected retinal fundus images and their predicted CVFs (Figure 3.4). In these examples, we observe that the SPS model struggles to predict a patient's age. This corresponds to the results obtained in Table 3.3; the SPS model gives a significantly larger loss value compared to the other models.

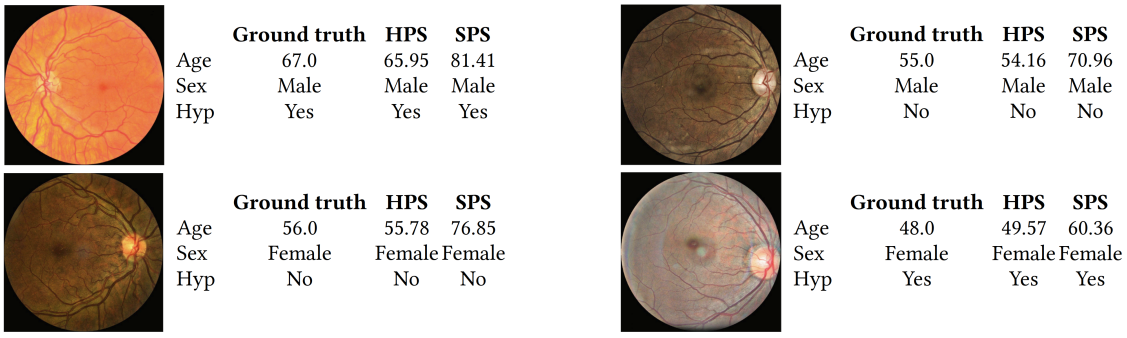


Figure 3.4: Showing predictions of the MTL models for hard- and soft-parameter sharing on randomly selected retinal fundus images. *Hyp* in the figure represents hypertension.

Next, we consider the gradient-weighted class activation map (Grad-CAM) to localise regions of retinal fundus images, in order to provide a visual explanation for decisions made by the best-performing multi-task model regarding sex. We observe that for male predictions, the model localises the optic disc (see Figure 3.5a), while it localises either the macula region (see Figure 3.5b) or both the optic disc and macula region (see Figure 3.5c) for female predictions [Betzler *et al.*, 2021; Dieck *et al.*, 2020; Poplin *et al.*, 2018]. Also, we visualise the localisation of some incorrectly classified examples. The image in Figure 3.5d is labelled as female but the model predicts it as male. We notice that the model even localises the optic disc as done in Figure 3.5a. Similarly, the image in Figure 3.5e is labelled as male but the model predicts female. Again, we notice that the model localises the macula region, suggesting that the model classified the image wrongfully. Finally, the image in Figure 3.5f is falsely classified as male and we observe that the model localises both the optic disc and the macula region.

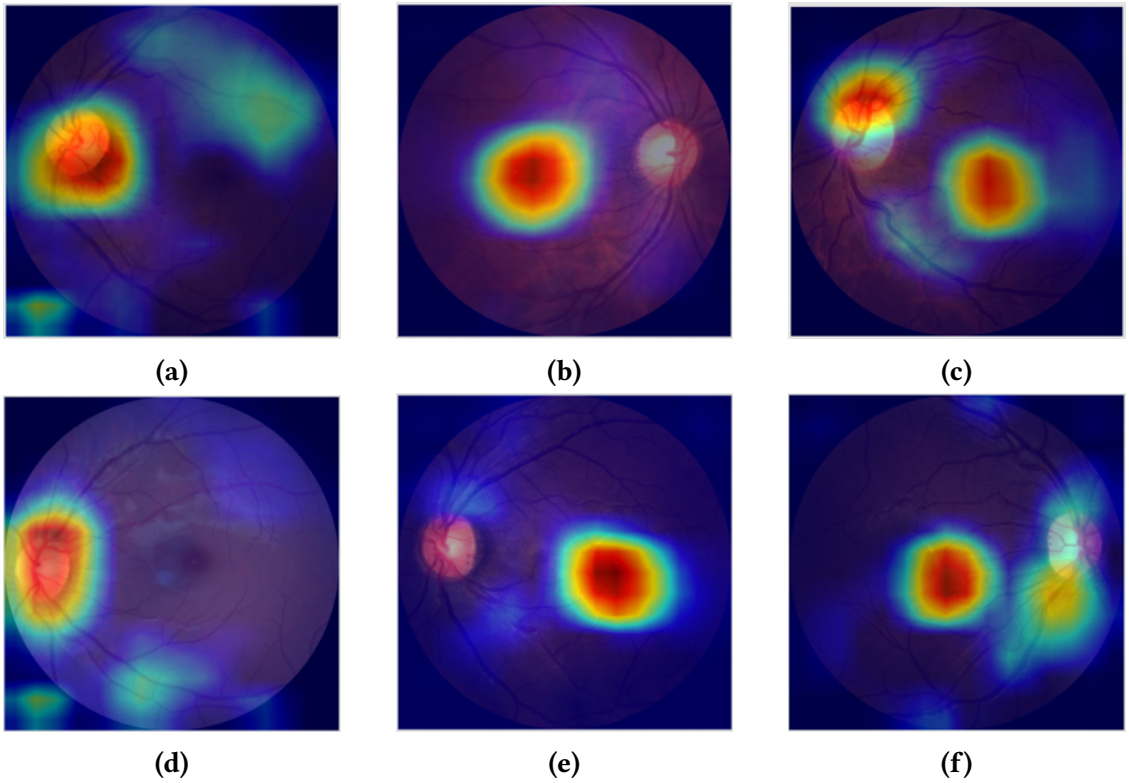


Figure 3.5: Showing localised regions from Grad-CAM on sampled retinal fundus images, for the task of sex prediction. The MTL model localises the optic disc for male predictions (Figure 3.5a) and localises either the macula region only or both the optic disc and the macula region in the same image (Figures 3.5b and 3.5c). For incorrectly predicted retinal fundus images, the model consistently localises the optic disc of an image originally labelled as female (Figure 3.5d). Figures 3.5e and 3.5f are originally labelled as male but the model predicts them as female and localises the macula region for the former, and both the optic disc and macula region for the latter.

3.6.2 Results for Fields 2 and 3

We also investigate how the other fields of view, that is *Field 2* and *Field 3*, perform for the MTL-HPS model. We notice in Table 3.4 that the performances on these fields are not on-par with the *Field 1* results. However, the validation results for the age prediction task and the testing results for the hypertension prediction task obtained for *Field 2* are better than those obtained for *Field 1*. Furthermore, the validation result obtained for the age prediction task for *Field 3* was superior to that obtained for *Field 1*.

We attribute the poor performance to the fact that *Field 2* and *Field 3* images contain

Table 3.4: Performance of the MTL-HPS model for the *Field 2* and *Field 3* views.

View	Training			Validation			Testing		
	Age (MSE)	Sex (Acc. %)	Hyp. (Acc. %)	Age (MSE)	Sex (Acc. %)	Hyp. (Acc. %)	Age (MSE)	Sex (Acc. %)	Hyp. (Acc. %)
Field 2	0.0045	89.76	74.20	0.0041	86.20	69.88	0.0042	81.75	68.41
Field 3	0.0046	84.25	75.17	0.0040	80.03	69.27	0.0056	73.95	59.64

at most one of the critical features needed for the correct prediction of CVFs, which are the optic disc and macula region. For example, we see in Figure 3.5 that the model focuses on either the optic disc or macula region, or both.

3.6.3 Results for Other Attributes

Image quality. We assess the performance of the image quality on the sex prediction task. We do this assessment to check if the quality of the image influences the performance of the model. We use the sex prediction task since it produced better performance in the previous experiments. Although the retinal fundus images categorised as excellent lead to the best accuracy, the differences between the performances were small. This suggests that the quality of the images does not influence the performance of the model (see Table 3.5).

Table 3.5: Test performance of the MTL-HPS model on sex classification, by image quality.

Quality	Images	Accuracy	Precision	Recall	F1-score
Adequate	14,500	85.90	82.11	84.84	83.45
Good	13,078	88.61	85.93	86.61	86.27
Excellent	6,897	89.29	86.04	85.68	85.86

Ethnicity. We assess the performance of the MTL-HPS model in predicting CVFs for the various ethnic groups. We notice that for the age prediction task, Asian, Indian subcontinent and Latin American groups give better results than the remaining groups. We observe a similar pattern for the sex prediction task, except that the African Descent

group gives better results. Regarding the hypertension task, we notice decreased performance in the Multi-racial and Native American groups. This can be due to the small number of training samples for the two groups. Overall, we observe that the results from the Caucasian, Multi-racial and Native American groups are not on-par with the remaining ethnic groups (see Table 3.6).

Table 3.6: Performance of the MTL-HPS model, split according to ethnicity.

Ethnicity	Age (MSE)	Sex (Acc. %)	Hyp (Acc. %)
African Descent	0.0044	85.77	73.36
Asian	0.0032	89.02	69.51
Caucasian	0.0057	79.40	63.52
Indian subcontinent	0.0030	84.18	69.49
Latin American	0.0039	87.36	68.58
Multi-racial	0.0032	78.05	58.54
Native American	0.0053	76.32	47.35

Camera. It is known that different fundus cameras operate under different settings and environments, resulting in different distributions of the fundus images and model performance. Therefore, we assess the performances of the various fundus camera used to capture the retinal fundus images in the EyePACS dataset. We observe that the performances for all the fundus cameras are similar except for the Volk Pictor Plus which gives slightly lower performance metrics. Note that the Volk Pictor Plus is the only handheld fundus camera considered in the study and, as indicated in Table 3.2, contains only images of adequate quality. Using accuracy, we notice that the Topcon fundus camera outperforms all the other fundus cameras. This is followed by the Zeiss Visucam, the Canon cameras, Crystalvue, Centrevue and finally the Optovue cameras (see Table 3.7).

Table 3.7: Test performance of the MTL-HPS model on sex classification, by the camera used.

Camera	Type	Images	Accuracy	Precision	Recall	F1-score
Canon CR1	D	2,778	88.05	82.20	89.08	85.50
Canon CR2	D	4,602	86.85	82.55	82.79	82.67
Canon DGi	D	2,795	88.37	85.32	86.44	85.88
Centervue DRS	D	974	86.86	85.42	82.00	83.67
Crystalvue NFC 700	D	689	87.37	85.36	83.86	84.60
Optovue Vivicon	D	124	84.68	89.16	88.10	88.62
Optovue iCam	D	11,429	85.90	83.17	85.95	84.53
Topcon NW 200	D	432	95.14	93.41	94.97	94.18
Topcon NW 400	D	9,588	89.97	86.94	85.20	86.06
Topcon NW 700	D	429	91.14	93.75	85.94	89.67
Volk Pictor Plus	H	316	76.27	73.33	89.02	80.42
Zeiss Visucam	D	10	90.00	100.00	85.71	92.31

For the camera type, D represents **D**esktop and H represents **H**andheld.

3.7 Conclusion

In summary, we created a multi-task learning (MTL) model to predict cardiovascular risk factors (CVFs), including patients' age, sex, and hypertension status. We observed comparable performance between the MTL and single-task learning (STL) models, even though more images were available for the latter. Within the MTL model, hard-parameter sharing (HPS) is preferred over soft-parameter sharing (SPS) as it obtained superior results. Further analysis revealed discriminating areas of the retina, including the optic disc and the macula region. We observed that the model focuses on the optic disc alone when predicting males, and either the macula region alone or both the optic disc and macula region when predicting females.

The results obtained from experimenting with *Field* 2 and *Field* 3 images stress the importance of the optic disc and the macula region in predicting sex. We also noticed that the quality of an image does not seem to influence the performance of the model. Finally, we investigated the performance of the various fundus cameras on sex prediction. We observed acceptable results for the handheld fundus camera and better performance for the desktop fundus cameras.

A limitation of our MTL model is that there are missing data, especially in case of hypertension. Hence, a reduced number of instances were used to train the MTL models, making it challenging to fairly compare the results with the STL model. Moreover, other risk factors such as smoking status and haemoglobin level are missing from many of the dataset instances, making them unusable. Another limitation of the study regarding camera performance is that the data contain significantly more images from desktop fundus cameras than from handheld fundus cameras. Also, 99% of the handheld fundus images belong to the "adequate" image quality group, and none labelled as "excellent". Hence, we are not able to attribute poor image quality to all handheld fundus cameras, as the only one of its kind in the data is the Volk Pictor Plus. However, other studies report having better retinal fundus image results from desktop fundus cameras than handheld fundus cameras [Rajalakshmi *et al.*, 2021].

While a desktop fundus camera may have many advantages, they are expensive, bulky, require a specialist to operate, and should be in the right environment to produce quality images. Despite their performance, handheld cameras have numerous advantages that make them desirable in the field of retinal imaging technology. For this reason, there has been numerous research conducted on the adaptation of handheld fundus cameras in the space of retinal imaging technology [Palermo *et al.*, 2022]. To this end, we suggest the ideal handheld fundus camera should (1) be comfortable for use (for both patients and operators), (2) have a shorter examination period to prevent patients' adaptation to darkness, (3) have quick adjustment settings, (4) be operable in different environments, (5) have a wide angle-of-view, and (6) produce a high-resolution image.

Chapter 4

Combining CNNs and Vision Transformers

4.1 Overview

The aim of this chapter is to design a model that combines desirable underlying properties from a convolutional operation and a Transformer encoder. First, we present some background on attention and the underlying benefits of Transformers in Section 4.2. Next, we discuss different attention-based models to provide insights into their workings in Section 4.3. We present preliminary work and details for the proposed model in Section 4.4. Finally, we give details of the experiments carried out in Section 4.5, the results obtained in Section 4.6, and conclusion in Section 4.7.

4.2 Background

For over a decade, convolutional neural networks (CNNs) have been the standard deep learning model for computer vision-related tasks [Bello *et al.*, 2019; Krizhevsky *et al.*, 2012]. Following their performance in computer vision, other domains including natural language processing (NLP) [Li and Mao, 2019; Wang and Gang, 2018; Yin *et al.*, 2017]

and speech recognition [Han *et al.*, 2020; Kubanek *et al.*, 2019; Passricha and Aggarwal, 2018] have either adapted CNNs to create hybrid models or built novel models made up of convolutional operators. Yet, convolutions are unable to capture long-range dependencies due to their poor scaling properties with respect to large receptive fields [Ramachandran *et al.*, 2019]. CNNs are mainly characterised by local connectivity and may not capture global information which is necessary for better recognition [Bello *et al.*, 2019].

These issues of CNNs were studied by [Baker *et al.*, 2020]. In their work, they probed several networks to classify shapes with conflicting local and global contour information. For example, they created squares with curved elements and circles with corner features. They observed that the models made predictions based on local contour features instead of the global shape. As expected, when the shapes were augmented to diminish information on local contour features, performance increased. They concluded that there is a converse relationship between human perception and CNNs regarding local and global shape classification. In the end, CNNs fail to spatially combine local features into a global understanding.

Work is being done to alleviate the dependencies on convolution in computer vision tasks and general deep learning. For example, Ramachandran *et al.* [2019] replaced all convolutions in the ResNet model with a module called stand-alone self-attention (SASA). They observed that self-attention is most effective in higher layers. Also, Tolstikhin *et al.* [2021] introduced an all-multi-layer perceptron architecture for computer vision. They acknowledged that while convolutions have attained better results over the years, it is necessary to stir further research. Motivated by human perception, Mnih *et al.* [2014] introduced a recurrent attention model to adaptively select a sequence of regions in the visual space at high resolution by setting the centre of fixation on the object of interest and ignoring irrelevant features. Their technique is non-differentiable and uses reinforcement learning approaches during training.

The concept of “attention” in machine learning is analogous to human perception.

Human perception does not process a whole scene at once, because the brain cannot fully process the information received by the optic nerve, which is in the order of 10^8 bits per second. Rather, human perception efficiently allocates resources to a fraction of information by selectively focusing on parts of the visual space to understand scenes. It combines several fixations on objects of interest to build up a representation of a scene [Mnih *et al.*, 2014; Zhang *et al.*, 2021].

Lindsay [2020] defined attention as an ability to flexibly control limited computational resources. In deep learning, attention is the process of focusing on interesting regions of data. It was first introduced by Bahdanau *et al.* [2014] to inject alignment (attention) between the input and output in a sequence-to-sequence model. The alignment is a score that is used to measure how well the input and the output match. Among a few variants of alignment are content-base [Graves *et al.*, 2014], location-base [Luong *et al.*, 2015], dot-product [Luong *et al.*, 2015] and scaled dot-product [Vaswani *et al.*, 2017]. The alignment score is used to estimate attention.

Current types of attention include hard attention [Xu *et al.*, 2015], soft attention [Xu *et al.*, 2015], global attention [Luong *et al.*, 2015], local attention [Luong *et al.*, 2015] and self-attention [Cheng *et al.*, 2016]. Recently, attention has played a critical role in the development of Transformer models. In particular, self-attention is used in Transformer models due to their robustness and inherent benefits of generalisation [Zhao *et al.*, 2020]. First introduced in the NLP domain by Vaswani *et al.* [2017], Transformers have become the de facto algorithm for NLP-related tasks and generally for sequential datasets [Dosovitskiy *et al.*, 2021], and are gaining popularity [Touvron *et al.*, 2022]. Transformers have been adapted in the computer vision field due to their large dynamic attention properties [Wu *et al.*, 2021], scalability [Dosovitskiy *et al.*, 2021], improved generalisation and long-range capacities [Touvron *et al.*, 2022]. For example, Dosovitskiy *et al.* [2021] developed a model called the vision Transformer (ViT) which has attained impressive results for computer vision tasks. Inspired by their success, other studies have created variants of ViT [Heo *et al.*, 2021; Liu *et al.*, 2021].

Even though ViTs have recently attained success, they lack the locality inherent to CNNs [Huang *et al.*, 2021]. Moreover, ViT models often require large-scale training [Dosovitskiy *et al.*, 2021]. In addition, the number of operations in ViT increases quadratically with the number of pixels in an input image [Chu *et al.*, 2021].

Local interaction and global understanding are imperative for medical image processing tasks, and to attain semantic information from the images [Lin *et al.*, 2022]. CNNs possess local attributes while ViTs have global properties. Some other distinctions noted by Park and Kim [2022] are (1) CNNs are high-pass filters (meaning CNNs emphasise on high-frequency component of an image such as edges) and ViTs are low-pass filters (meaning ViTs emphasise on low-frequency component of an image such as blur effects); and (2) the key components of CNNs (convolutional operators) diversify feature maps while those of ViT (self-attentions) aggregate them.

Lin *et al.* [2022] argued that there are underlying relationships between convolutions and self-attention. To inherit the best of both worlds, we introduce a hybrid model which combines convolutions and self-attention modules. Our model feeds a fully convolutional Transformer module (FCT) (see Section 4.4.3) with intermediate layers from a ResNet CNN model. We train the model in an end-to-end (without pre-trained weights) fashion and attain promising performance. We demonstrate a new perspective for future designs of hybrid models containing convolutions and self-attention modules while using fewer parameters.

4.3 Related Work

Rao *et al.* [2021] defined attention as the ability by which the human brain processes visual information while also evaluating the relevance of input features. Replicating it computationally, attention has become an integral part of deep learning [Bello *et al.*, 2019]. Hu *et al.* [2018] described attention as a technique of biassing the allocation of available computational resources to the most informative components of a signal.

Attention allows a model to dynamically attend to the most relevant regions of an input [Ba *et al.*, 2014], instead of summarising an entire input to a static representation [Xu *et al.*, 2015]. Hence, attention improves representations and suppresses irrelevant parts of an input [Wang *et al.*, 2017; Woo *et al.*, 2018]. It was first introduced in NLP-related tasks but has been adapted to computer vision. Specifically, attention has been adopted in computer vision to address some of the drawbacks of CNNs in vision-related tasks including the inability to capture long-range dependencies [Woo *et al.*, 2018].

In medical images, Sinha and Dolz [2020] used self-attention to capture rich contextual dependencies from abdominal organs, brain tumours and cardiovascular structures to produce precise segmentation. Nie *et al.* [2018] also proposed attention-based deep networks to obtain improved segmentation performance. Hu *et al.* [2021] proposed scale-attention deep networks to generate semantic segmentation of retinal images. Rao *et al.* [2021] studied the effects of self-attention on medical images. In detail, they empirically compared different self-attention models across various medical images and observed improved AUC-ROC scores.

In this section, we present and investigate some of the works that attention is being used for. We first present how attention is used in convolution-based and Transformer models. Finally, we show how studies have combined CNNs and Transformers to inherit the best of the two paradigms.

4.3.1 Convolution-based Attention Models

In the early days, some studies incorporated attention into computer vision tasks by augmenting already existing models. The primary operator used was the convolution operator and the focus was on either the channel or spatial properties of a layer. In this study, we call these setups convolution-based attention. We explore two convolution-based attention models, namely the squeeze-and-excitation network (SENet) [Hu *et al.*, 2018] and the convolution block attention module (CBAM) [Woo *et al.*, 2018]. We highlight SENet due to its novel approach to addressing attention, and CBAM because it

builds upon SENet by exploring spatial attention in addition to channel attention [Rao *et al.*, 2021].

Squeeze-and-Excitation Networks. Hu *et al.* [2018] argued that prior research has given much importance to the spatial component of a CNN model. For this reason, they proposed a model that focuses on the channel component of the model. There are three main steps involved in their model, namely (in order) squeezing, excitation and scaling. The squeezing step summarises a global spatial feature map into a channel descriptor by using a global average pooling. This is followed by the excitation step which employs a gating mechanism to recalibrate the squeezed output. The final output of the module is then obtained by scaling the excitation output with the global spatial feature map to produce a recalibrated channel (see Figure 4.1). SENet is usually created with residual models as the backbone, hence, it is occasionally referred to as SEResNet.

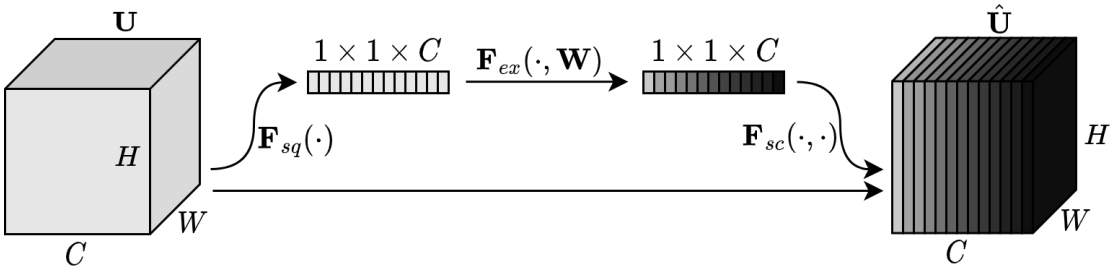


Figure 4.1: Illustrating the workflow of the squeeze-and-excitation network (SENet). The goal is to recalibrate the channels of a global spatial feature map by performing a series of steps: squeezing (sq), excitation (ex), and scaling (sc). $\mathbf{U} \in \mathbb{R}^{H \times W \times C}$ is a feature map used to compute $\hat{\mathbf{U}} \in \mathbb{R}^{H \times W \times C}$. $\mathbf{F}_{sq}$, $\mathbf{F}_{ex}$, and $\mathbf{F}_{sc}$ represent the squeezing function, the excitation function, and the scaling function respectively. $\mathbf{W}$ denotes the weights of the layer.

Convolutional Block Attention Module. Convolutional block attention module (CBAM) [Woo *et al.*, 2018] builds upon SENet by including a spatial attention component. The channel attention is also approached differently. In detail, the channel attention follows a sequential order of max pooling and average pooling in parallel, followed

by a multi-layer perceptron (MLP) and another max pooling and average pooling. The output of the channel attention is fed to the spatial attention module which uses convolution to generate a feature map with only one channel. As an ablation, the authors changed the order of the channel and spatial attention. They observed better performance with channel-first attention (see Figure 4.2).

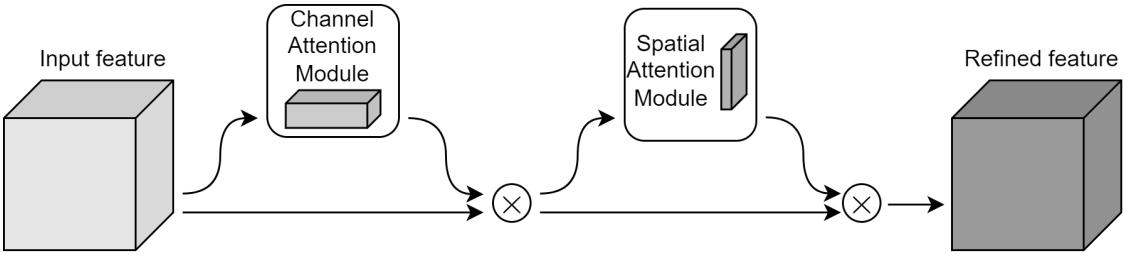


Figure 4.2: The main contribution of the convolutional block attention module (CBAM) is the introduction of the channel and spatial attention module. In sequential order, it first applies channel attention to the global spatial feature map and then spatial attention to the result. $\otimes$ in the diagram represents element-wise multiplication.

4.3.2 Transformer Models

The main components of a Transformer encoder are multi-head self-attention (MHSA) and MLP. Each self-attention is called a head. MHSA is made up of several self-attentions stacked in parallel to generate one output (see Figure 4.7a). MHSA takes in normalised input and feeds the normalised output to an MLP. The Transformer model then stacks several encoders to extract useful features for its task. The vision Transformer (ViT) is the first computer vision-adapted Transformer [Dosovitskiy *et al.*, 2021]. In this section, we explore the ViT model and two of its variants.

Vision Transformer. The ViT model [Dosovitskiy *et al.*, 2021] takes in split images called patches and feeds them to several stacked Transformer encoders for classification. While the structure of the ViT is similar to the original Transformer, the latter has

a decoder component, and the position of the normalisation operation in ViT is also different. ViT applies normalisation before MHSA and MLP, while the Transformer model applies normalisation after MHSA and MLP (see Figure 4.3).

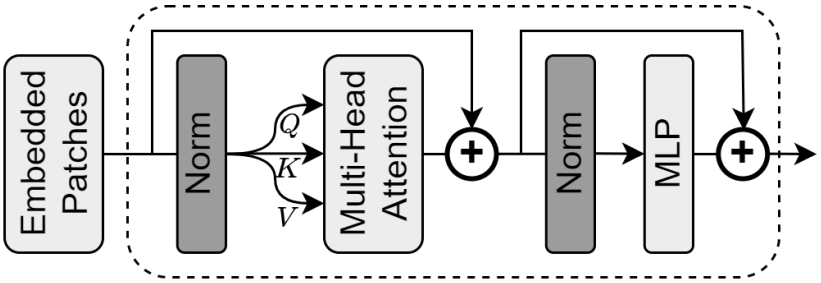


Figure 4.3: Overview of the ViT encoder module. Q, K, and V in the diagram denote representations called the query, key and value (see Section 4.4.1 for details). Norm is a normalising layer in the architecture.

Pooling-based Vision Transformer. Heo *et al.* [2021] created a variant of ViT called the pooling-based vision Transformer (PiT). Inspired by the dimension reduction principle of CNNs, they empirically showed that reducing the spatial dimension of a ViT model improves performance. As the model gets deeper, they applied spatial pooling to reduce the spatial dimension of feature maps of a ViT model.

Shifted Windows Vision Transformer. Liu *et al.* [2021] introduced the shifted window vision Transformer model (Swin). It is at the time of writing the best-performing variant of the ViT model. Swin employs a hierarchical design and a shifted window technique to induce inductive bias and provide efficient computation of the Transformer encoders. Inductive bias refers to the set of assumptions about the underlying distribution of the data that represent the inherent preferences encoded in a learning algorithm to generalise from a finite set of observations to unseen data [Hüllermeier *et al.*, 2013]. The hierarchical design merges 2×2 neighbouring patches in deeper layers. This merging is repeated four times.

The Swin model has two consecutive Transformer encoders. The first has a windowed multi-head self-attention (W-MSA) and the second a shifted window multi-head self-attention (SW-MSA). The computation of the standard Transformer has quadratic complexity since every patch attends to all the other patches. To resolve this, the authors introduced window-based attention such that each window has a fixed number of patches. MSA is only applied among the patches within a window, making the encoder scalable and efficient.

However, W-MSA lacks connections across other windows, which can be imperative for a computer vision task. To add connections across other windows, the authors proposed the SW-MSA, which uses a cyclic approach to permute patches in a window across other windows. In detail, all windows are shifted by half their width and height. At the end of the computations, the windows are reverse-shifted to their original positions. In summary, the Swin model uses a hierarchical design along with a shifted window scheme on a non-overlapping local window to efficiently apply self-attention. The full model has several Swin encoders stacked on top of each other for a corresponding task (see Figure 4.4).

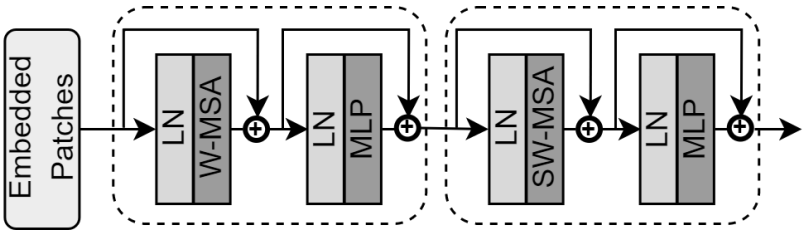


Figure 4.4: An overview of the Swin encoder which consecutively stacks two Transformer encoders with a window multi-head self-attention and a shifted window multi-head self-attention. LN in this diagram means layer normalisation.

4.3.3 Hybrid Models

As mentioned earlier, both convolutions and the self-attention modules in Transformers have desirable properties for computer vision-related tasks [Lin *et al.*, 2022; Wu *et al.*,

2021]. New studies have combined convolutions and Transformer modules into a unified framework [Andreoli, 2019; Cordonnier *et al.*, 2020] to obtain the best of these two paradigms [d'Ascoli *et al.*, 2021; Lin *et al.*, 2022; Wu *et al.*, 2021]. In this study, we refer to these kinds of setups as hybrid models. We present two hybrid models: CVT [Wu *et al.*, 2021] and ConViT [d'Ascoli *et al.*, 2021].

CVT. Wu *et al.* [2021] leveraged convolution to create a convolutional Transformer block. They replaced the linear projection component of a Transformer encoder with convolutional operations. They hypothesised that introducing convolutions in the Transformer encoder progressively improves representation richness. While doing so, the encoder reduces the resolutions of the feature maps and simultaneously increases the width of the feature maps. The MLP component of the Transformer remains in the CVT model (see Figure 4.5).

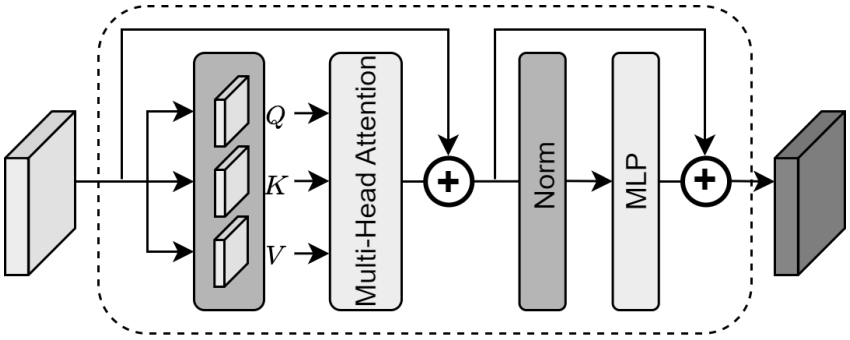


Figure 4.5: Showing the overall CVT encoder which replaces the linear projection component of a Transformer encoder with a convolutional operator. Q, K, and V represent the query, key and value representations for the module (see Section 4.4.1 for details).

ConViT. Self-attention is position agnostic [d'Ascoli *et al.*, 2021]. d'Ascoli *et al.* [2021] proposed gated positional self-attention (GPSA) which introduces soft convolutional inductive bias in Transformer encoders. In addition to the content awareness of a Transformer encoder, the main contribution of their study was the introduction of position awareness, termed as position self-attention (PSA). They observed that initialising PSA

using a convolutional scheme resulted in better performance in early epochs of training. Again, they observed in the later epochs that the attention mechanisms ignore content representations. They resolved this by using a gating scheme with a learning parameter λ which provides relative importance to the content and position component of the module accordingly (see Figure 4.6). Using the theorem below by Cordonnier *et al.* [2020], d'Ascoli *et al.* [2021] concluded that their module behaves as a convolutional layer.

Theorem 4.3.1. *A multi-head self-attention layer with N_h heads of dimension D_h, output dimension D_{out} and a relative positional encoding of dimension $D_p \geq 3$ can express any convolutional layer of kernel size $\sqrt{N_h} \times \sqrt{N_h}$ and $\min(D_h, D_{out})$ output channels.*

In other words, an MSA layer can effectively approximate the behaviour of a convolutional layer with specified dimensions.

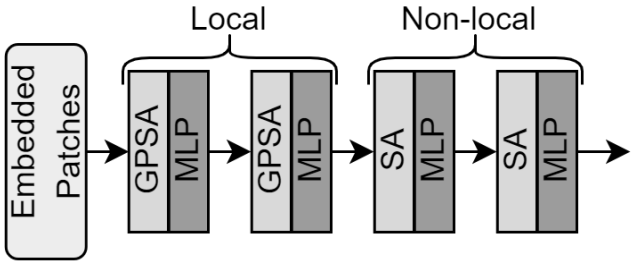


Figure 4.6: The main contribution of a ConViT encoder is the addition of position information to a Transformer encoder.

4.3.4 Summary

In summary, we see a pattern in the design of the various models discussed above, especially for Transformer-related models. The encoders or modules usually perform self-attention which is followed by MLP computations. In our study, we use the same design but with convolutional projections instead of linear projections.

4.4 Methodology

In this section, we propose a model which feeds intermediate layers from a ResNet to a fully convolutional Transformer (FCT) module. The aim is to build a hybrid model that is able to learn long-range dependencies and capture global features of an image. We present some preliminaries in Section 4.4.1, the overall workflow of the proposed model in Section 4.4.2, and the details of the FCT module in Section 4.4.3.

4.4.1 Preliminaries

We consider a dataset $\mathcal{D} = \{\mathcal{X}, \mathcal{Y}\}$, where $\mathcal{X}$ represents the input images and $\mathcal{Y}$ represents the target values. For each image $\mathbf{x} \in \mathcal{X} \subset \mathbb{R}^{H \times W \times C}$, H and W are the height and width of the image, and C represents the number of channels of the image. We seek to learn a model that best approximates the true target values. In this study, we modify a ResNet-50 model (see Section 2.3.2) to have auxiliary layers at certain parts of the network. Specifically, the auxiliary layers are fully convolutional Transformer (FCT) modules (explained in Section 4.4.3), which take in input feature maps from certain layers of the network. By doing so, we expect the model to focus on discriminating regions of the input while paying less attention to regions of less importance. First, we present background on the various components used in the model. These include multi-head self-attention (MHSA), focus layer (FL), batch normalisation (BN), layer normalisation (LN), and Gaussian error linear units.

Multi-Head Self-Attention. We linearly transform an input $\mathbf{x} \in \mathbb{R}^{H \times W \times C}$ to three representations, namely a query (Q), key (K) and value (V). First, $\mathbf{x}$ is represented as a sequence $[x_1, x_2, \ldots, x_n]$, where each $x_i \in \mathbb{R}^{d_{\text{model}}}$ is a vector that represents the embedding of the i-th feature. d_{model} denotes the embedding dimension. The trainable matrices W_{Q_i}, W_{K_i} and W_{V_i} are used to compute the Q, K and V.The representations are computed as

$$Q_i = \mathbf{x}W_{Q_i} \quad K_i = \mathbf{x}W_{K_i}, \quad V_i = \mathbf{x}W_{V_i}. \tag{4.4.1}$$

Next, we define self-attention as

$$\text{Attention}(Q_i, K_i, V_i) = \text{softmax}\left(\frac{Q_i K_i^\top}{\sqrt{d}}\right) V_i, \tag{4.4.2}$$

where d is a scaling factor equal to the dimension size of Q or K. The softmax function maps a vector $\mathbf{a} = (a_1, \ldots, a_N) \in \mathbb{R}^N$ to another vector $\mathbf{a}' \in \mathbb{R}^N$ such that the elements of $\mathbf{a}'$ are positive and sum to 1. The outputs of the softmax function are interpreted as probabilities (p_i), and defined as

$$\text{softmax}(\mathbf{a})_i = p_i = \frac{\exp a_i}{\sum_{j=1}^{N} \exp a_j}, \qquad i \in 1, \ldots, N. \tag{4.4.3}$$

The scaling factor d in Equation 4.4.2 is necessary to avoid the possibility of generating very small gradients from the softmax function due to the large magnitude output from the dot product between Q and K. MHSA employs the softmax function to map real values (which can contain negatives) to values between 0 and 1. A single self-attention is known as a head. That is,

$$\text{head}_i = \text{Attention}(Q_i, K_i, V_i), \qquad i = 1, \ldots, h. \tag{4.4.4}$$

Multi-head self-attention (MHSA) computes several heads in parallel and concatenates the outputs. Hence, MHSA is given as

$$\text{MHSA}(\mathbf{x}) = \textsc{Concat}(\text{head}_1, \text{head}_2, \ldots, \text{head}_h) W_O, \tag{4.4.5}$$

where $\textsc{Concat}(\cdot)$ is a concatenating function and W_O is a trainable weight matrix. MHSA is necessary for the model to jointly attend to different informative regions of an input [Vaswani *et al.*, 2017]. In summary, MHSA generates several scaled dot-product attention heads from three linearly transformed representations and concatenates the outputs (see Figure 4.7).

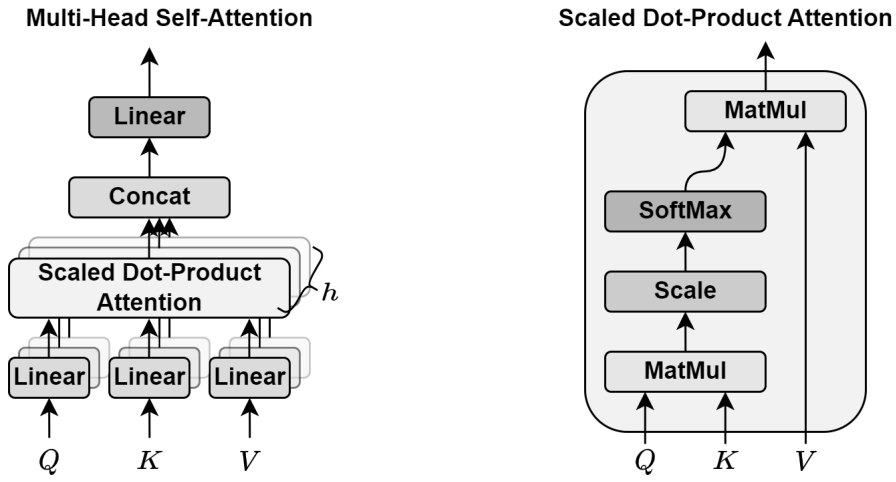


Figure 4.7: Figure 4.7a shows the overall layout of an MHSA (see Equation 4.4.5). Figure 4.7b illustrates the details of a scaled dot-product attention (see Equation 4.4.2).

Focus Layer. The focus layer (FL) is a feature aggregation layer which applies convolutions to extract fine-grained information from the MHSA output. We use convolution operations in the focus layer because they generally have fewer parameters and thus lead to better parameter efficiency [O'Shea and Nash, 2015; Pang *et al.*, 2017; Wu, 2017]. For input $\mathbf{x} \in \mathbb{R}^{H \times W \times C}$, the focus layer is defined as

$$\mathrm{FL}(\mathbf{x}) = \sigma(\textsc{Conv}(\mathbf{x})), \tag{4.4.6}$$

where $\textsc{Conv}(\cdot)$ is a convolutional operator and $\sigma(\cdot)$ is an activation function.

Batch and Layer Normalisation. Normalisation, as the name suggests, is a technique used to normalise the mini-batch or layers of a model to zero mean and unit variance. It is called batch normalisation (BN) [Ioffe and Szegedy, 2015] if applied on a mini-batch, and layer normalisation (LN) [Ba *et al.*, 2016] otherwise. For a sample $\mathbf{x} \in \mathbb{R}^d$, normalisation is defined as

$$\mathrm{N}(x) = \left[(\mathbf{x} - \boldsymbol{\mu}) \times \boldsymbol{\Sigma}^{-1}\right] \otimes \boldsymbol{\gamma} + \boldsymbol{\beta}, \tag{4.4.7}$$

where $\boldsymbol{\mu} \in \mathbb{R}^d$ is the mean and $\boldsymbol{\Sigma} \in \mathbb{R}^{d \times d}$ is the standard deviation of the feature maps, $\otimes$ is element-wise multiplication, and $\boldsymbol{\gamma} \in \mathbb{R}^d$, $\boldsymbol{\beta} \in \mathbb{R}^d$ are learnable parameters.

Gaussian Error Linear Unit. The activation function we use for the FCT is the Gaussian error linear unit (GELU). It was introduced by Hendrycks and Gimpel [2016], and is defined as

$$\begin{aligned} \text{GELU}(x) &= x\Phi(x) \\ &= \frac{x}{2} \cdot \left[1 + \text{erf}\left(\frac{x}{\sqrt{2}}\right)\right] \\ &\approx \frac{x}{2} \cdot \left[1 + \tanh\left(\sqrt{\frac{2}{\pi}} \cdot \left(x + 0.044715x^3\right)\right)\right], \end{aligned} \tag{4.4.8}$$

where $\Phi(x)$ is the cumulative distribution function of the standard normal distribution. The authors used the standard normal distribution because of the normalisation components in the FCT module.

4.4.2 Paying Multiple Attention

Our proposed model is illustrated in Figure 4.8. In detail, we extract feature maps denoted by $\hat{\mathbf{z}}_l$, where $l \in \{1, \ldots, \ell\}$ represents a convolutional layer. Assuming equal dimensions, we add $\hat{\mathbf{z}}_l$ to a global image descriptor $\boldsymbol{g}$ and pass the output to an FCT for attention. A global image descriptor in this case is the output of the penultimate layer of a ResNet-50 model. Finally, the feature maps from the FCT modules are concatenated into a single vector for classification purposes. We use this approach of learning to force earlier layers in the model to learn similar mappings of the global image descriptor of the vanilla model (without attention). We achieve this by using $\hat{\mathbf{z}}_l$ to contribute directly to the classification step [Jetley *et al.*, 2018].

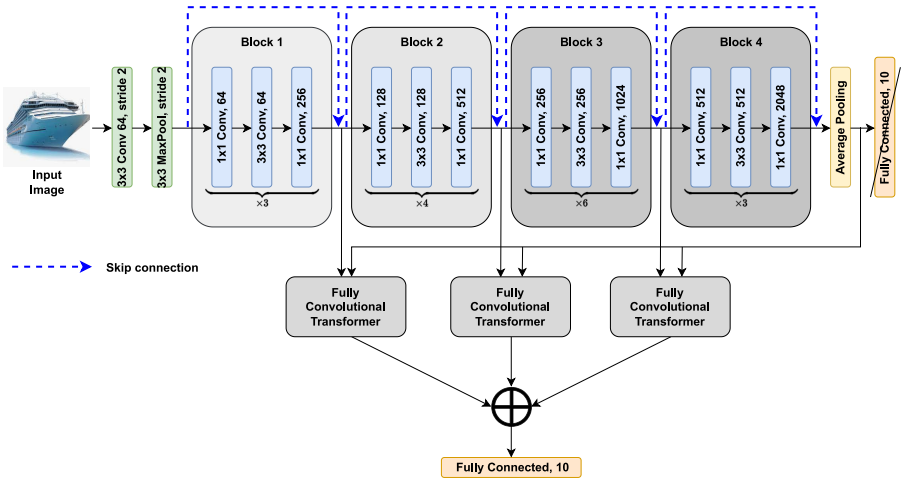


Figure 4.8: Illustrating the architecture of our proposed model. Instead of feeding linear classification layers with feature maps of the final block of a backbone model, we first feed intermediate feature maps to FCT modules to capture long-range dependencies, then concatenate the output for later classification.

4.4.3 Fully Convolutional Transformer

The fully convolutional Transformer (FCT) module is a special case of the Transformer module (Figure 4.9). In FCT, transformations are done using convolutional functions instead of position-wise linear projections for the attention operation inherent in Transformer [Wu *et al.*, 2021]. The motivation behind using convolutions is to keep local relations between pixels or features while simultaneously maintaining the Transformer structure.

In our model, the input to the FCT module is a feature map extracted from intermediate layers of the ResNet-50 model. First, we convert the feature maps into overlapping patches using convolution. The generated patches are analogous to tokens in NLP [Wu *et al.*, 2021]. Next, we feed the generated patches to a depth-wise convolution to generate Q, K, and V. We normalise the outputs and apply MHSA to generate attention. Finally, we fuse the outputs with the patches and feed a normalised resultant to the focus layer which aggregates features using convolution. We summarise FCT mathematically as follows:

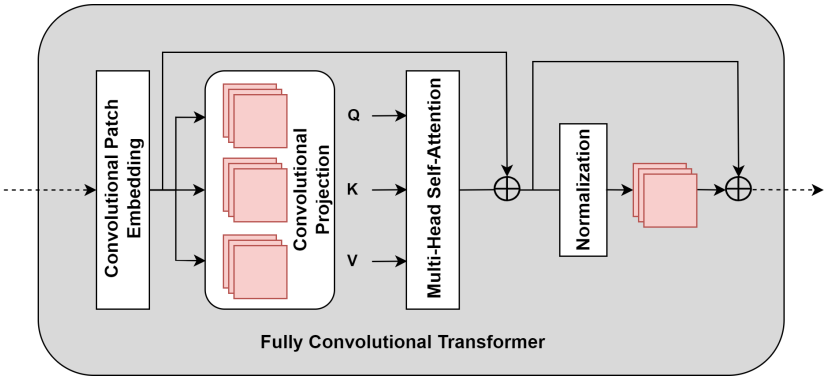


Figure 4.9: Details of the fully convolutional Transformer (FCT) module. It takes in feature maps from intermediate layers of the backbone network, creates patch embeddings, projects to Q, K, and V for the MHSA mechanism, and feeds to another convolution layer for classification.

$$\mathbf{z}_{l-1} = \textsc{PatchEmbed}(\hat{\mathbf{z}}_{l-1} + \boldsymbol{g}), \tag{4.4.9}$$

$$\mathbf{z}_{l} = \text{MHSA}(\textsc{ConvProj}(\mathbf{z}_{l-1})) + \mathbf{z}_{l-1}, \tag{4.4.10}$$

$$\mathbf{z}_{l+1} = \text{FL}(\text{N}(\mathbf{z}_{l})) + \mathbf{z}_{l}, \tag{4.4.11}$$

where $\hat{\mathbf{z}}_{l-1} \in \mathbb{R}^{h \times w \times c}$ is a feature map from an intermediate representation of the network and $\boldsymbol{g} \in \mathbb{R}^{h \times w \times c}$ is a global image descriptor. $\textsc{PatchEmbed}(\cdot)$ is a convolutional operator used to create patch embeddings, which involves extracting informative features from sub-regions of an image and representing them as feature vectors in a lower-dimensional space. The patch embeddings are used to generate Q, K, and V for MHSA (see Equation 4.4.5) using $\textsc{ConvProj}$ which is a depth-wise convolution. Before that, Q, K, and V are normalised using either batch normalisation or layer normalisation (see Equation 4.4.7) resulting in $\mathbf{z}_{l+1} \in \mathbb{R}^{h \times w \times c}$.

4.5 Experiments

Details of the dataset used for our experiments are given in Section 4.5.1. We present the pre-processing and augmentation techniques employed on the dataset in Section 4.5.2. Finally, we provide model implementation details and how the weights of the models are initialised in Sections 4.5.3 and 4.5.4, respectively.

4.5.1 Data

In this study, we use data from the Canadian Institute for Advanced Research with ten classes (CIFAR-10). There are 60,000 natural coloured images in the CIFAR-10 dataset, with 6,000 images per class, and an image size of $32 \times 32 \times 3$. CIFAR-10 has a training set and a testing set containing 50,000 and 10,000 images, respectively. The ten classes are airplane, automobile, bird, cat, deer, dog, frog, horse, ship, and truck (Figure 4.10). We use CIFAR-10 to measure the performance of the proposed model on natural images [Krizhevsky *et al.*, 2009]. Later, we use the proposed model to classify the classes in the APTOS dataset, and compare the results to those obtained in Chapter 2.

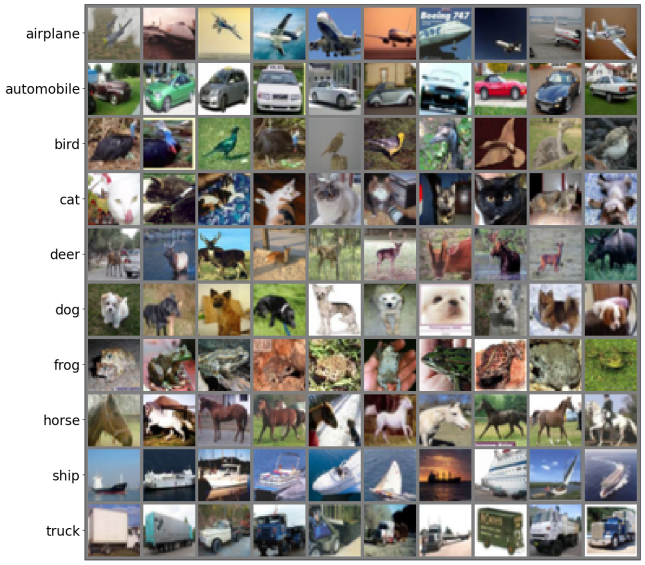


Figure 4.10: Showing randomly selected examples of the CIFAR-10 dataset.

4.5.2 Pre-processing and Augmentation

We do not apply any pre-processing techniques to the CIFAR-10 dataset. Instead, we augment the CIFAR-10 training dataset by random cropping and then padding the cropped regions, and random flipping vertically and horizontally. Moreover, we nor-

malise all the images to have pixel values between -1 and 1. We used the same settings in Section 2.5.2 to pre-process and augment the APTOS dataset.

4.5.3 Models and Implementation

In the next section, we compare the performance of different attention-based models in predicting the classes of CIFAR-10 and also the severity of diabetic retinopathy in the APTOS dataset. The models we use are in three major categories, namely convolution-based attention, vision Transformer, and hybrid models. Again, we compare the performance of the models when initialised with random weights and when initialised with pre-trained weights from ImageNet. In addition, we replace the final layers with a global average pooling layer, a dropout layer with 50% dropout rate and a linear layer with the number of units equal to the number of classes in the dataset used for the experiment.

4.5.4 Pre-training and Random Initialisation

We use similar settings from Sections 2.5.3.1 and 2.5.3.2 for both pre-training and random initialisation. That is, we use the data splitting settings and optimise the models with the Adam optimiser. For pre-trained initialisation, we initialise the weights of the model with pre-trained weights from the corresponding models. For random initialisation, we initialise the model weights with Kaiming He normal initialisation since the global extractor used for the study is a ResNet-50 model. For CIFAR-10 we use a batch size of 128 and a learning rate of 0.01, a weight decay of 10^{-4} and cyclical learning rates [Smith and Topin, 2018]. We also clip all gradients at global norm 1 [Dosovitskiy *et al.*, 2021]. When we train the models from scratch on CIFAR-10, we train for 200 epochs. For the APTOS dataset, we train for 100 epochs when training from scratch, and 20 epochs when we use pre-trained weights to initialise the model.

4.6 Results

In this section, we show the results obtained from the proposed model. First, we experiment with the CIFAR-10 dataset and then also the APTOS dataset. Also, we predict the classes of the APTOS dataset using the different attention-based models presented in Section 4.3.

4.6.1 Performance on CIFAR-10

We predict the classes of the CIFAR-10 image dataset using a modified ResNet-50 model. Specifically, we feed intermediate layers at different positions of the ResNet-50 model to the FCT module. We first train a vanilla ResNet-50 model and achieve 92.87% top-1 validation accuracy (see Table 4.1). Next, we experiment with two normalisation techniques in our proposed FCT model, namely batch normalisation (BN) and layer normalisation (LN). We observe a superior performance of BN over the LN version. Additionally, we initialise our model with pre-trained weights and observe a further increase in performance.

Table 4.1: Top-1 validation classification accuracy on the CIFAR-10 dataset.

Model	Top-1 Acc. (%)
vanilla (ResNet-50)	92.87
ours (with LN)	93.04
ours (with BN)	93.35
ours (pre-trained)	95.72

We visualise the attention maps of the various auxiliary layers used in our model. Since we add a global image descriptor to intermediate feature maps, we expect certain regions of the output to have high values if they contain similar parts of dominating regions of the global image descriptor. We observe that the earliest layer (that is the first FCT) produces localisations exclusive to the objects of interest. In particular, we

can observe some distinction between the localisation map generated and the background. We also observe that the localisation maps concentrate on specific regions in the images as the model gets deeper (see Figure 4.11).

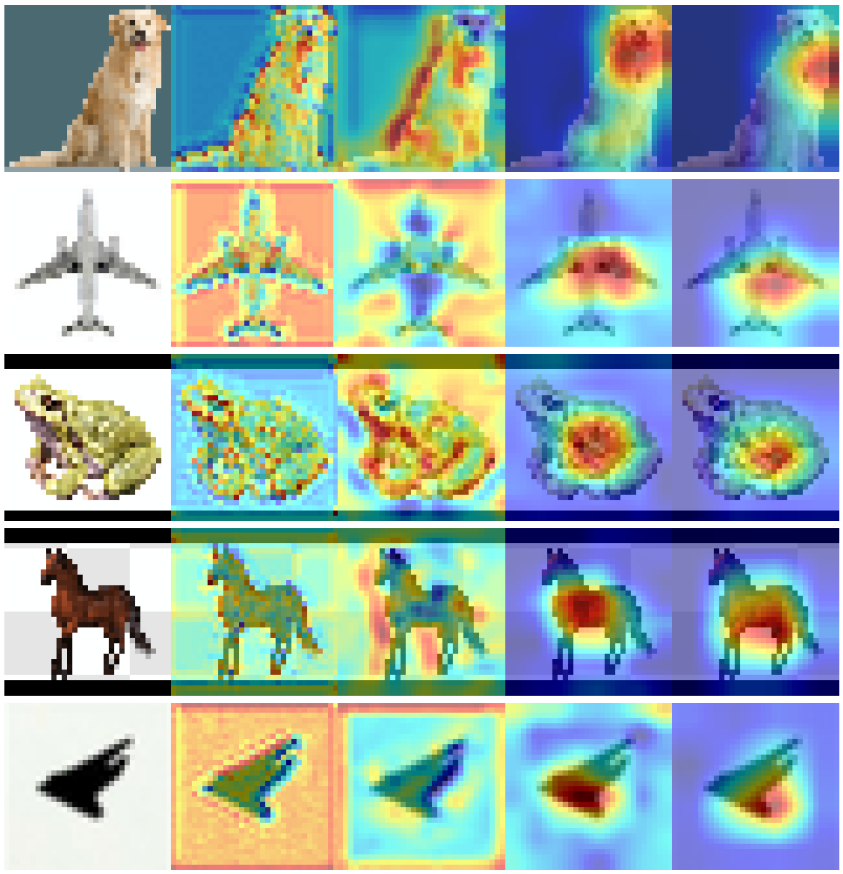


Figure 4.11: Discriminating regions of randomly selected images using Grad-CAM. The **FCT 1 – 3** columns in the figure represent the three FCT modules introduced at the intermediate layers of the ResNet-50 model. The **Overall** column represents the concatenated layer for classification.

4.6.2 Performance on APTOS

Next, we explore the performance of the various models discussed in Section 4.3 on the APTOS database along with our proposed approach. Again, we evaluate the per-

formance of the models with the AUC score. The models are grouped into three types, namely convolution-based attention, Transformers and hybrid. We note that the convolution-based attention model in general has fewer parameters compared to the other two types, while the Transformer models have more parameters (see Table 4.2).

Table 4.2: Performance of the convolution-based attention models, the Transformer models, and the hybrid models (including our proposed model) on the APTOS dataset.

Model	Type	Parameters	Training		Validation		Testing	
			Random Init.	Pre-trained	Random Init.	Pre-trained	Random Init.	Pre-trained
CBAM	Conv-based	26,034,789	96.81	98.25	94.82	97.00	95.52	97.45
SEResNet		26,049,269	97.25	97.83	94.90	97.10	95.72	96.53
ViT	Transformer	85,802,501	93.92	98.33	93.28	96.46	93.90	96.75
PiT		72,744,965	95.81	98.25	93.68	97.19	94.00	97.06
Swin		86,748,349	81.88	99.15	83.40	97.29	82.77	97.53
CVT	Hybrid	19,614,405	95.11	98.78	93.63	97.17	94.51	97.12
ConViT		85,774,885	92.20	96.70	91.83	96.68	92.86	96.29
FCT (ours)		24,101,509	95.17	97.58	95.50	97.04	95.66	97.46

Again, we observe a significant performance difference between randomly initialising the weights and using pre-trained weights. This difference is apparent in the Swin model which outperforms all the other models. The Swin model also produces acceptable results when randomly initialised. Generally, we observe that the convolution-based attention models perform better in the random initialisation experiments for the various sets of data. However, for the pre-trained experiments, the Transformer models outperform the other two types for all the sets of data.

Two of the hybrid models, namely the CVT model and our proposed model, explicitly introduce convolutions to the Transformer encoder, whereas the ConViT model does not. We observe that CVT and our proposed model outperform ConViT for all the sets of data. The ConViT model is a vision Transformer but with the addition of a position component, and we observe similar performance between the ViT and the ConViT models. They also have a similar number of parameters (approximately 85 million).

Overall, the testing set results are close to (and in some cases better than) that of the validation set, signalling that there is no real issue with overfitting.

With the second fewest number of parameters among the models, our proposed model outperforms most of the other models. It also performs better than the vanilla ResNet-50 model (see Table 2.1). In short, the proposed model outperforms many of the other models, uses fewer parameters, and provides new perspectives on designing a deep learning model. We also show plots of localised regions on randomly selected fundus images from each class generated by the different models used in this study (Figure 4.12).

4.7 Conclusion

In this chapter, we explored different attention-based models. The models were categorised into three groups, namely convolution-based attention models, Transformer models and hybrid models. Even though these models are different, they are similar in design and exhibit similar characteristics. In addition, we proposed a model that feeds intermediate feature maps of a ResNet-50 model to a fully convolutional Transformer (FCT) module.

We experimented with the proposed model on the CIFAR-10 dataset and observed improved performance over the baseline vanilla model (that is, the original ResNet-50 model). In addition, we found that batch normalisation slightly outperformed layer normalisation, and observed better results when we initialised the model with pre-trained weights. Furthermore, we visualised the attention maps of the various layers of the FCT module and noticed that in the earlier layers, there exists a clear distinction between the generated localisation maps and the background. Also, we noticed that as the model gets deeper, the attention maps focus on specific regions in the images.

Finally, we conducted experiments with the attention-based models using the APTOS dataset of retinal fundus images, to classify the severity of diabetic retinopathy. We

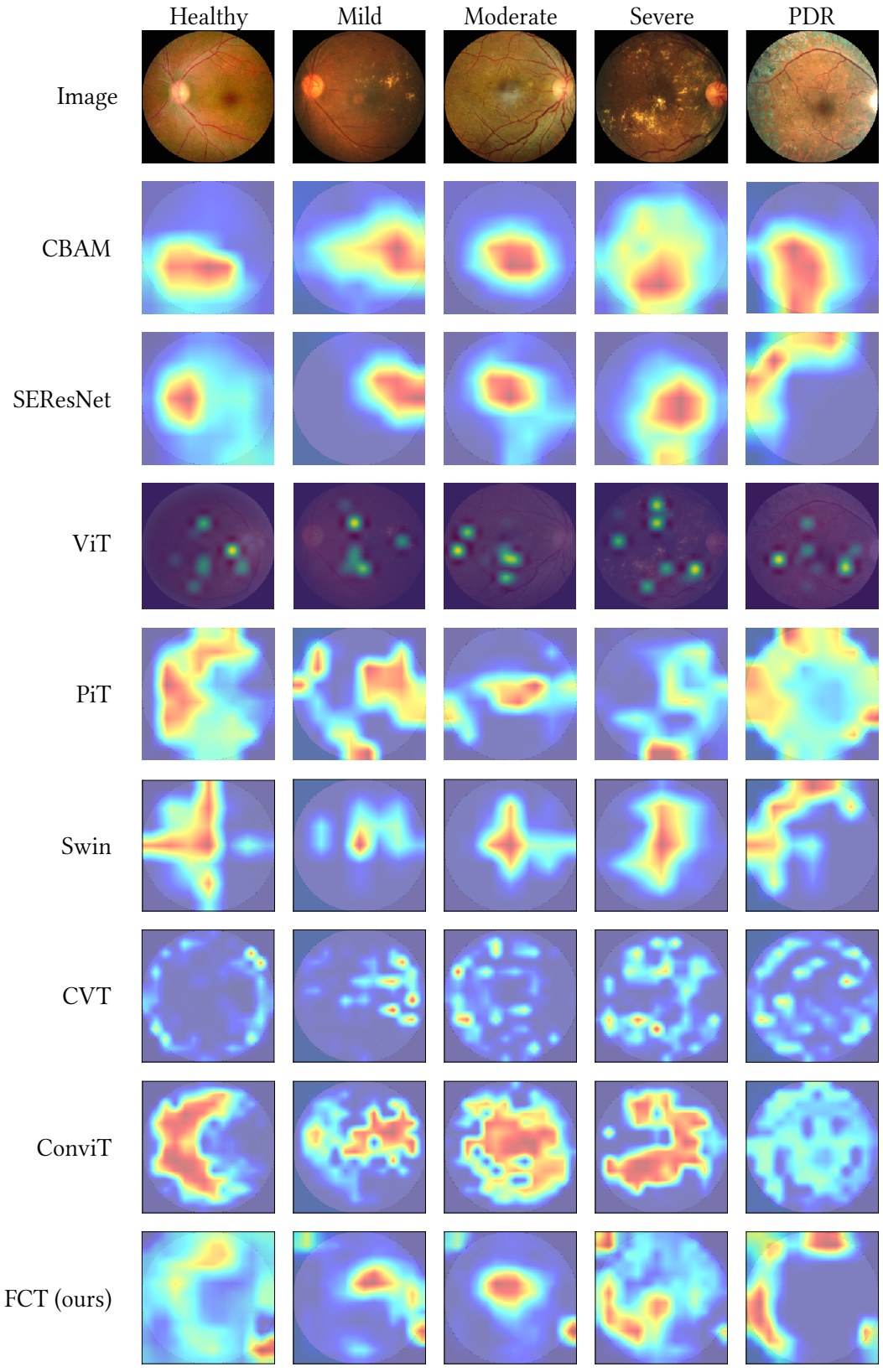


Figure 4.12: Showing localised regions from different models on randomly selected retinal fundus images from each class of diabetic retinopathy.

observed on-par performance between the models for all the sets of data. However, the convolution-based attention model obtained better results when weights are randomly initialised. Also, we observed that hybrid models with explicit convolutional operations, including CVT and the proposed model, outperform the hybrid model without explicit convolutions – that is the ConViT model.

In summary, the proposed model does not outperform all the other models but performs competitively while using fewer parameters. These results are promising and encourage a new perspective for future designs.

Chapter 5

Summary

This dissertation focused on building deep learning models to classify the severity of diabetic retinopathy, extract regions of interest, and predict cardiovascular risk factors from retinal fundus images. In detail, we built deep learning models first for single task problems and extended it to multi-task problems. We also improved performance by exploring hybrid models which combine convolutions and Transformer encoders. We did this in support of better diagnostic pipelines for retinopathy practitioners.

First, we addressed the low-quality issues associated with retinal fundus images with the help of contrast-limited adaptive histogram equalisation (CLAHE). We estimated the hyperparameters used for CLAHE for each retinal fundus image individually. The impact of using CLAHE was apparent. We observed improved results for the classification models after applying CLAHE. Moreover, we could train the models for fewer epochs (30 epochs fewer) than before CLAHE was applied.

In addition, we studied the use of transfer learning. We evaluated the performances of models initialised with random weights and models initialised with pre-trained weights from ImageNet. For all the experiments conducted, we observed that the models initialised with pre-trained weights outperformed their counterparts initialised with random weights. These are interesting results, as ImageNet only contains natural images and does not have any medical images. Thus, transfer learning can be imperative in

deep learning, especially in the medical imaging domain, as there is not a large amount of labelled data from which to train models from scratch.

Although the results obtained are impressive, they may not be useful for clinicians in determining which features to pay attention to. For this reason, we used gradient-weight class activation mapping (Grad-CAM) to generate localisation maps and identify discriminative regions on retinal fundus images, in the context of diabetic retinopathy classification. After generating the localisation maps, we observed that the model focuses attention mainly to the lesions on the retinal fundus images. The overall layout was to first pre-process the retinal fundus images, feed the output to a convolutional neural network (CNN) for classification purposes, and localise discriminative regions on the retinal fundus images (see Figure 2.12). We evaluated four state-of-the-art models for this task. In the end, ResNet-50 outperformed the other models by achieving an AUC score of 97.40% (see Table 2.2).

Furthermore, the dissertation focused on predicting cardiovascular risk factors (CVFs) using a multi-task learning (MTL) model. Particularly, we evaluated the performance of two approaches to creating MTL models. These were hard-parameter sharing (HPS) and soft-parameter sharing (SPS). We observed superior performance from the HPS approach. Moreover, we demonstrated the advantages of using multi-task learning over a single-task learning model, in that the MTL provides a unique approach to simultaneously predict several tasks (in this case, the CVFs). This approach saved time during training and attained on-par performance while training on fewer dataset instances than the single-task model (see Table 3.3).

Later, the dissertation focused on harnessing the advantages of convolutions and Transformer encoders. Thus, we created a hybrid model that fuses convolutions and Transformer encoders. Particularly, we fed intermediate feature maps from a ResNet-50 model to a fully convolutional Transformer (FCT) module. The FCT module replaces the linear operations of a Transformer encoder with convolutional operations. This was done to achieve locality even at earlier layers of the ResNet-50 model. First, we evalu-

ated the proposed model on the CIFAR-10 dataset and observed improved performance over the vanilla ResNet-50 model (see Table 4.1).

Finally, we compared the proposed hybrid model to other attention-based models. Specifically, we grouped the models into three groups, namely convolution-based attention models, Transformer models, and hybrid models. Generally, we observed that the convolution-based attention models obtained better results when initialised with random weights. Although the proposed model did not outperform all the other models, it attained impressive results. The hybrid model paradigm thus provides novel ideas for the future design of a deep learning model (see Table 4.2).

Most of the work in this dissertation focused on localisation maps generated by the trained models. Therefore, a future extension could be to perform a deeper analysis of the generated localisation maps, which could possibly lead to a better understanding of the underlying causes and effects of diabetic retinopathy. For example, how can one evaluate localisation maps quantitatively? Or how can we eliminate subjectivity in the topic of localisation maps in images (in our case retinal fundus images)? Localisation maps are generally evaluated by observation. Quantitative analyses of localisation maps can go a long way to gaining the trust of clinicians and assisting them to make better judgments. Quantifying the various localisation maps can also be a way to assess how trustworthy and interpretable the different models are [Teng *et al.*, 2022].

Another interesting line of future work can involve integrating multiple modalities, including optical coherence tomography (OCT) scans, retinal fundus images, and other patient records, to enhance the prediction of cardiovascular risk factors. This direction of work could be beneficial for improving the model's performance, provide an opportunity to estimate the risk of cardiovascular diseases (CVDs), and potentially mimic the diagnostic procedure performed by clinicians [Midena *et al.*, 2020; Munk *et al.*, 2021]. In addition, a compelling avenue for future work may emerge by exploring effective techniques of incorporating incomplete data records into the MTL model's training in the context of CVF predictions.

While our experiments with the proposed hybrid FCT model produced promising findings on CIFAR10 and retinal fundus images, it may be intriguing extending our investigations to include a wide range of benchmark datasets. For instance, since APTOS is considered as a fine-grained dataset [Dippel *et al.*, 2021], we could explore our hybrid model's performance on other fine-grained datasets such as CUB-200 [Wah *et al.*, 2011], Stanford Dogs [Khosla *et al.*, 2011], and many others.

We hope that the research presented in this dissertation provides a solid foundation for further advancement in deep learning applied to retinal fundus images and medical image analysis as a whole.

List of References

Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., Corrado, G.S., Davis, A., Dean, J., Devin, M., Ghemawat, S., Goodfellow, I., Harp, A., Irving, G., Isard, M., Jia, Y., Jozefowicz, R., Kaiser, L., Kudlur, M., Levenberg, J., Mané, D., Monga, R., Moore, S., Murray, D., Olah, C., Schuster, M., Shlens, J., Steiner, B., Sutskever, I., Talwar, K., Tucker, P., Vanhoucke, V., Vasudevan, V., Viégas, F., Vinyals, O., Warden, P., Wattenberg, M., Wicke, M., Yu, Y. and Zheng, X. (2015). TensorFlow: Large-scale machine learning on heterogeneous systems. Software available from tensorflow.org.
Available at: https://www.tensorflow.org/

Abràmoff, M.D., Garvin, M.K. and Sonka, M. (2010). Retinal imaging and image analysis. *IEEE reviews in biomedical engineering*, vol. 3, pp. 169–208.

Al-Bander, B. (2018). *Retinal Image Analysis Based on Deep Learning*. The University of Liverpool (United Kingdom).

Alom, M.Z., Yakopcic, C., Nasrin, M., Taha, T.M., Asari, V.K. *et al.* (2019). Breast cancer classification from histopathological images with inception recurrent residual convolutional neural network. *Journal of digital imaging*, vol. 32, no. 4, pp. 605–617.

Alzami, F., Megantara, R.A., Fanani, A.Z. *et al.* (2019). Diabetic retinopathy grade classification based on fractal analysis and random forest. In: *2019 International Seminar on Application for Technology of Information and Communication (iSemantic)*, pp. 272–276. IEEE.

Andreoli, J.-M. (2019). Convolution, attention and structure embedding. *arXiv preprint arXiv:1905.01289*.

Angeletti, G., Caputo, B. and Tommasi, T. (2018). Adaptive deep learning through visual domain localization. In: *2018 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 7135–7142. IEEE.

Antonetti, D.A., Silva, P.S. and Stitt, A.W. (2021). Current understanding of the molecular and cellular pathology of diabetic retinopathy. *Nature Reviews Endocrinology*, vol. 17, no. 4, pp. 195–206.

Aurangzeb, K., Aslam, S., Alhussein, M., Naqvi, R.A., Arsalan, M. and Haider, S.I. (2021). Contrast enhancement of fundus images by employing modified pso for improving the performance of deep learning models. *IEEE Access*, vol. 9, pp. 47930–47945.

Ayhan, M.S., Faber, H., Kühlewein, L., Inhoffen, W., Aliyeva, G., Ziemssen, F. and Berens, P. (2023). Multitask learning for activity detection in neovascular age-related macular degeneration. *Translational Vision Science & Technology*, vol. 12, no. 4, pp. 12–12.

Ba, J., Mnih, V. and Kavukcuoglu, K. (2014). Multiple object recognition with visual attention. *arXiv preprint arXiv:1412.7755*.

Ba, J.L., Kiros, J.R. and Hinton, G.E. (2016). Layer normalization. *arXiv preprint arXiv:1607.06450*.

Bahdanau, D., Cho, K. and Bengio, Y. (2014). Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.

Baker, N., Lu, H., Erlikhman, G. and Kellman, P.J. (2020). Local features and global shape information in object classification by deep convolutional neural networks. *Vision research*, vol. 172, pp. 46–61.

Barnett, A.J., Schwartz, F.R., Tao, C., Chen, C., Ren, Y., Lo, J.Y. and Rudin, C. (2021). A case-based interpretable deep learning model for classification of mass lesions in digital mammography. *Nature Machine Intelligence*, vol. 3, no. 12, pp. 1061–1070.

Bazzani, L., Bergamo, A., Anguelov, D. and Torresani, L. (2016). Self-taught object localization with deep networks. In: *2016 IEEE winter conference on applications of computer vision (WACV)*, pp. 1–9. IEEE.

Beede, E., Baylor, E., Hersch, F., Iurchenko, A., Wilcox, L., Ruamviboonsuk, P. and Vardoulakis, L.M. (2020). A human-centered evaluation of a deep learning system deployed in clinics for the detection of diabetic retinopathy. In: *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, pp. 1–12.

Bello, I., Fedus, W., Du, X., Cubuk, E.D., Srinivas, A., Lin, T.-Y., Shlens, J. and Zoph, B. (2021). Revisiting ResNets: Improved training and scaling strategies. *Advances in Neural Information Processing Systems*, vol. 34, pp. 22614–22627.

Bello, I., Zoph, B., Vaswani, A., Shlens, J. and Le, Q.V. (2019). Attention augmented convolutional networks. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 3286–3295.

Betzler, B.K., Yang, H.H.S., Thakur, S., Yu, M., Da Soh, Z., Lee, G., Tham, Y.-C., Wong, T.Y., Rim, T.H., Cheng, C.-Y. *et al.* (2021). Gender prediction for a multiethnic population via deep learning across different retinal fundus photograph fields: retrospective cross-sectional study. *JMIR medical informatics*, vol. 9, no. 8, p. e25165.

Bhagat, N., Grigorian, R.A., Tutela, A. and Zarbin, M.A. (2009). Diabetic macular edema: pathogenesis and treatment. *Survey of ophthalmology*, vol. 54, no. 1, pp. 1–32.

Bhusal, D., Panday, D. and Prasad, S. (2022). Multi-label classification of thoracic diseases using dense convolutional network on chest radiographs. *arXiv preprint arXiv:2202.03583*.

Burkardt, J. (2014). The truncated normal distribution. *Department of Scientific Computing Website, Florida State University*, vol. 1, p. 35.

Cai, L., Gao, J. and Zhao, D. (2020). A review of the application of deep learning in medical image classification and segmentation. *Annals of translational medicine*, vol. 8, no. 11.

Calì, C. and Longobardi, M. (2015). Some mathematical properties of the ROC curve and their applications. *Ricerche di Matematica*, vol. 64, no. 2, pp. 391–402.

Çallı, E., Sogancioglu, E., van Ginneken, B., van Leeuwen, K.G. and Murphy, K. (2021). Deep learning for chest X-ray analysis: A survey. *Medical Image Analysis*, vol. 72, p. 102125.

Campos, G.F.C., Mastelini, S.M., Aguiar, G.J., Mantovani, R.G., Melo, L.F.d. and Barbon, S. (2019). Machine learning hyperparameter selection for contrast limited adaptive histogram equalization. *EURASIP Journal on Image and Video Processing*, vol. 2019, no. 1, pp. 1–18.

Cao, L. and Li, H. (2020). Enhancement of blurry retinal image based on non-uniform contrast stretching and intensity transfer. *Medical & Biological Engineering & Computing*, vol. 58, no. 3, pp. 483–496.

Carrera, E.V., González, A. and Carrera, R. (2017). Automated detection of diabetic retinopathy using SVM. In: *2017 IEEE XXIV International Conference on Electronics, Electrical Engineering and Computing (INTERCON)*, pp. 1–4. IEEE.

Chalam, K., Chamchikh, J. and Gasparian, S. (2022). Optics and utility of low-cost smartphone-based portable digital fundus camera system for screening of retinal diseases. *Diagnostics*, vol. 12, no. 6, p. 1499.

Chen, S., Zhang, Y. and Yang, Q. (2021). Multi-task learning in natural language processing: An overview. *arXiv preprint arXiv:2109.09138*.

Cheng, J., Dong, L. and Lapata, M. (2016 Nov). Long short-term memory-networks for machine reading. In: *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pp. 551–561. Association for Computational Linguistics.

Chouhan, V., Singh, S.K., Khamparia, A., Gupta, D., Tiwari, P., Moreira, C., Damaševičius, R. and De Albuquerque, V.H.C. (2020). A novel transfer learning based approach for pneumonia detection in chest X-ray images. *Applied Sciences*, vol. 10, no. 2, p. 559.

Chu, X., Tian, Z., Wang, Y., Zhang, B., Ren, H., Wei, X., Xia, H. and Shen, C. (2021). Twins: Revisiting the design of spatial attention in vision transformers. *Advances in Neural Information Processing Systems*, vol. 34, pp. 9355–9366.

Collobert, R. and Weston, J. (2008). A unified architecture for natural language processing: Deep neural networks with multitask learning. In: *Proceedings of the 25th International Conference on Machine Learning*, pp. 160–167.

Cordonnier, J.-B., Loukas, A. and Jaggi, M. (2020). On the relationship between self-attention and convolutional layers. In: *International Conference on Learning Representations*.

Crawshaw, M. (2020). Multi-task learning with deep neural networks: A survey. *arXiv preprint arXiv:2009.09796*.

Dara, S. and Tumma, P. (2018). Feature extraction by using deep learning: A survey. In: *2018 Second International Conference on Electronics, Communication and Aerospace Technology (ICECA)*, pp. 1795–1801. IEEE.

Date, R.C., Jesudasen, S.J., Weng, C.Y. *et al.* (2019). Applications of deep learning and artificial intelligence in retina. *International Ophthalmology Clinics*, vol. 59, no. 1, pp. 39–57.

De Fauw, J., Ledsam, J.R., Romera-Paredes, B., Nikolov, S., Tomasev, N., Blackwell, S., Askham, H., Glorot, X., O'Donoghue, B., Visentin, D. *et al.* (2018). Clinically applicable deep learning for diagnosis and referral in retinal disease. *Nature medicine*, vol. 24, no. 9, pp. 1342–1350.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K. and Fei-Fei, L. (2009). Imagenet: A large-scale hierarchical image database. In: *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 248–255. IEEE.

Dey, S. (2018). *Hands-On Image Processing with Python: Expert techniques for advanced image analysis and effective interpretation of image data*. Packt Publishing Ltd.

Dieck, S., Ibarra, M., Moghul, I., Yeung, M.W., Pantel, J.T., Thiele, S., Pfau, M., Fleckenstein, M., Pontikos, N. and Krawitz, P.M. (2020). Factors in color fundus photographs that can be used by humans to determine sex of individuals. *Translational vision science & technology*, vol. 9, no. 7, pp. 8–8.

Dippel, J., Vogler, S. and Höhne, J. (2021). Towards fine-grained visual representations by combining contrastive learning with image reconstruction and attention-weighted pooling. *arXiv preprint arXiv:2104.04323*.

Dobrescu, A., Giuffrida, M.V. and Tsaftaris, S.A. (2020). Doing more with less: a multitask deep learning approach in plant phenotyping. *Frontiers in plant science*, vol. 11, p. 141.

Dodo, B.I. (2020). *Retinal layer segmentation in optical coherence tomography images.* Ph.D. thesis, Brunel University London.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S. *et al.* (2021). An image is worth 16x16 words: Transformers for image recognition at scale. In: *International Conference on Learning Representations*.

Duong, L., Cohn, T., Bird, S. and Cook, P. (2015). Low resource dependency parsing: Cross-lingual parameter sharing in a neural network parser. In: *Proceedings of the 53rd annual meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (volume 2: short papers)*, pp. 845–850.

D'Agostino Sr, R.B., Vasan, R.S., Pencina, M.J., Wolf, P.A., Cobain, M., Massaro, J.M. and Kannel, W.B. (2008). General cardiovascular risk profile for use in primary care: the framingham heart study. *Circulation*, vol. 117, no. 6, pp. 743–753.

d'Ascoli, S., Touvron, H., Leavitt, M.L., Morcos, A.S., Biroli, G. and Sagun, L. (2021). Convit: Improving vision transformers with soft convolutional inductive biases. In: *International Conference on Machine Learning*, pp. 2286–2296. PMLR.

Gahir, N.K. and Shah, M. (2020). Anatomy of the eye and the healthy fundus. *Diabetic Retinopathy: Screening to Treatment 2E (ODL)*, p. 13.

Gan, C., Wang, N., Yang, Y., Yeung, D.-Y. and Hauptmann, A.G. (2015). DevNet: A deep event network for multimedia event detection and evidence recounting. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2568–2577.

Gandhi, M. and Dhanasekaran, R. (2013). Diagnosis of diabetic retinopathy using morphological process and SVM classifier. In: *2013 International Conference on Communication and Signal Processing*, pp. 873–877. IEEE.

Gargeya, R. and Leng, T. (2017). Automated identification of diabetic retinopathy using deep learning. *Ophthalmology*, vol. 124, no. 7, pp. 962–969.

Glorot, X. and Bengio, Y. (2010). Understanding the difficulty of training deep feedforward neural networks. In: *Proceedings of the thirteenth International Conference on Artificial Intelligence and Statistics*, pp. 249–256. JMLR Workshop and Conference Proceedings.

Goldstein, B.A., Navar, A.M. and Carter, R.E. (2017). Moving beyond regression techniques in cardiovascular risk prediction: applying machine learning to address analytic challenges. *European heart journal*, vol. 38, no. 23, pp. 1805–1814.

Gondal, W.M., Köhler, J.M., Grzeszick, R., Fink, G.A. and Hirsch, M. (2017). Weakly-supervised localization of diabetic retinopathy lesions in retinal fundus images. In: *2017 IEEE International Conference on Image Processing (ICIP)*, pp. 2069–2073. IEEE.

Goodfellow, I., Bengio, Y. and Courville, A. (2016). *Deep Learning*. MIT Press. http://www.deeplearningbook.org.

Gramatikov, B.I. (2014). Modern technologies for retinal scanning and imaging: an introduction for the biomedical engineer. *Biomedical engineering online*, vol. 13, no. 1, pp. 1–35.

Graves, A., Wayne, G. and Danihelka, I. (2014). Neural turing machines. *arXiv preprint arXiv:1410.5401*.

Gu, J., Wang, Z., Kuen, J., Ma, L., Shahroudy, A., Shuai, B., Liu, T., Wang, X., Wang, G., Cai, J. *et al.* (2018). Recent advances in convolutional neural networks. *Pattern recognition*, vol. 77, pp. 354–377.

Gulshan, V., Peng, L., Coram, M., Stumpe, M.C., Wu, D., Narayanaswamy, A., Venugopalan, S., Widner, K., Madams, T., Cuadros, J. *et al.* (2016). Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. *JAMA*, vol. 316, no. 22, pp. 2402–2410.

Gupta, B. and Tiwari, M. (2019). Color retinal image enhancement using luminosity and quantile based contrast enhancement. *Multidimensional Systems and Signal Processing*, vol. 30, no. 4, pp. 1829–1837.

Hajian-Tilaki, K. (2013). Receiver operating characteristic ROC curve analysis for medical diagnostic test evaluation. *Caspian journal of internal medicine*, vol. 4, no. 2, p. 627.

Han, W., Zhang, Z., Zhang, Y., Yu, J., Chiu, C.-C., Qin, J., Gulati, A., Pang, R. and Wu, Y. (2020). ContextNet: Improving Convolutional Neural Networks for Automatic Speech Recognition with Global Context. In: *Proc. Interspeech 2020*, pp. 3610–3614.

He, K., Zhang, X., Ren, S. and Sun, J. (2015). Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In: *Proceedings of the IEEE international conference on computer vision*, pp. 1026–1034.

He, K., Zhang, X., Ren, S. and Sun, J. (2016). Deep residual learning for image recognition. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778.

Hendrycks, D. and Gimpel, K. (2016). Gaussian error linear units (gelus). *arXiv preprint arXiv:1606.08415*.

Heo, B., Yun, S., Han, D., Chun, S., Choe, J. and Oh, S.J. (2021). Rethinking spatial dimensions of vision transformers. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 11936–11945.

Hu, J., Shen, L. and Sun, G. (2018). Squeeze-and-excitation networks. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 7132–7141.

Hu, J., Wang, H., Wang, J., Wang, Y., He, F. and Zhang, J. (2021). SA-Net: A scale-attention network for medical image segmentation. *PloS one*, vol. 16, no. 4, p. e0247388.

Huang, W., Zhang, Y. and Li, L. (2019). Survey on multi-objective evolutionary algorithms. In: *Journal of Physics: Conference series*, vol. 1288, p. 012057. IOP Publishing.

Huang, X., Deng, Z., Li, D. and Yuan, X. (2021). MISSformer: An effective medical image segmentation transformer. *arXiv preprint arXiv:2109.07162*.

Hui, W., Tan, C., Gu, G. and Zhao, Y. (2022). Gradient-based refined class activation map for weakly supervised object localization. *Pattern Recognition*, vol. 128, p. 108664.

Hüllermeier, E., Fober, T. and Mernberger, M. (2013). *Inductive Bias*, pp. 1018–1018. Springer New York, New York, NY. ISBN 978-1-4419-9863-7.
Available at: https://doi.org/10.1007/978-1-4419-9863-7_927

Ioffe, S. and Szegedy, C. (2015). Batch normalization: Accelerating deep network training by reducing internal covariate shift. In: *International Conference on Machine Learning*, pp. 448–456. PMLR.

Jadhav, A. and Patil, P.B. (2015). Classification of diabetes retina images using blood vessel area. *International Journal on Cybernetics & Informatics (IJCI) Vol*, vol. 4.

Jahangiry, L., Farhangi, M.A. and Rezaei, F. (2017). Framingham risk score for estimation of 10-years of cardiovascular diseases risk in patients with metabolic syndrome. *Journal of Health, Population and Nutrition*, vol. 36, no. 1, pp. 1–6.

Jain, R., Nagrath, P., Kataria, G., Kaushik, V.S. and Hemanth, D.J. (2020). Pneumonia detection in chest X-ray images using convolutional neural networks and transfer learning. *Measurement*, vol. 165, p. 108046.

James, G., Witten, D., Hastie, T., Tibshirani, R. *et al.* (2013). *An introduction to statistical learning*, vol. 112. Springer.

Jetley, S., Lord, N.A., Lee, N. and Torr, P.H. (2018). Learn to pay attention. In: *International Conference on Learning Representations*.

Jiang, Y., Chen, L., Zhang, H. and Xiao, X. (2019). Breast cancer histopathological image classification using convolutional neural networks with small SE-ResNet module. *PloS one*, vol. 14, no. 3, p. e0214587.

Jin, Q., Meng, Z., Pham, T.D., Chen, Q., Wei, L. and Su, R. (2019). DUNet: A deformable network for retinal vessel segmentation. *Knowledge-Based Systems*, vol. 178, pp. 149–162.

Juneja, M. and Nagar, S. (2016). Particle swarm optimization algorithm and its parameters: A review. In: *2016 International Conference on Control, Computing, Communication and Materials (ICCCCM)*, pp. 1–5. IEEE.

Kakadiaris, I.A., Vrigkas, M., Yen, A.A., Kuznetsova, T., Budoff, M. and Naghavi, M. (2018). Machine learning outperforms acc/aha cvd risk calculator in mesa. *Journal of the American Heart Association*, vol. 7, no. 22, p. e009476.

Khan, A.F., Jalil, A., Haq, I.U. and Shah, S.I.H. (2021). Automatic localization of macula and identification of macular degeneration in retinal fundus images. In: *2021 International Conference on Electrical, Communication, and Computer Engineering (ICECCE)*, pp. 1–6. IEEE.

Khan, H.A., Jue, W., Mushtaq, M. and Mushtaq, M.U. (2020). Brain tumor classification in mri image using convolutional neural network. *Math. Biosci. Eng*, vol. 17, no. 5, pp. 6203–6216.

Khan, S., Rahmani, H., Shah, S.A.A. and Bennamoun, M. (2018). A guide to convolutional neural networks for computer vision. *Synthesis lectures on computer vision*, vol. 8, no. 1, pp. 1–207.

Khosla, A., Jayadevaprakash, N., Yao, B. and Li, F.-F. (2011). Novel dataset for fine-grained image categorization: Stanford dogs. In: *Proc. CVPR workshop on fine-grained visual categorization (FGVC)*, vol. 2. Citeseer, Colorado Springs, CO.

Kingma, D.P. and Ba, J. (2014). Adam: A method for stochastic optimization.

Kolb, H. (2011). Facts and figures concerning the human retina.

Korot, E., Pontikos, N., Liu, X., Wagner, S.K., Faes, L., Huemer, J., Balaskas, K., Denniston, A.K., Khawaja, A. and Keane, P.A. (2021). Predicting sex from retinal fundus photographs using automated deep learning. *Scientific reports*, vol. 11, no. 1, pp. 1–8.

Kowluru, R.A. and Chan, P.-S. (2008). Capillary dropout in diabetic retinopathy. *Diabetic retinopathy*, pp. 265–282.

Krishna, K., Toshniwal, S. and Livescu, K. (2018). Hierarchical multitask learning for CTC-based speech recognition. *arXiv preprint arXiv:1807.06234*.

Krizhevsky, A., Hinton, G. *et al.* (2009). Learning multiple layers of features from tiny images.

Krizhevsky, A., Sutskever, I. and Hinton, G.E. (2012). Imagenet classification with deep convolutional neural networks. In: *Advances in neural information processing systems*, pp. 1097–1105.

Kubanek, M., Bobulski, J. and Kulawik, J. (2019). A method of speech coding for speech recognition using a convolutional neural network. *Symmetry*, vol. 11, no. 9, p. 1185.

Kuran, U. and Kuran, E.C. (2021). Parameter selection for CLAHE using multi-objective cuckoo search algorithm for image contrast enhancement. *Intelligent Systems with Applications*, vol. 12, p. 200051.

Labhade, J.D., Chouthmol, L. and Deshmukh, S. (2016). Diabetic retinopathy detection using soft computing techniques. In: *2016 International Conference on Automatic Control and Dynamic Optimization Techniques (ICACDOT)*, pp. 175–178. IEEE.

LeCun, Y., Bottou, L., Bengio, Y. and Haffner, P. (1998). Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324.

Lee, D.-H. and Liu, J.-L. (2021). End-to-end multi-task deep learning and model based control algorithm for autonomous driving. *arXiv preprint arXiv:2112.08967*.

Lee, R., Wong, T.Y. and Sabanayagam, C. (2015). Epidemiology of diabetic retinopathy, diabetic macular edema and related vision loss. *Eye and vision*, vol. 2, no. 1, pp. 1–25.

Li, M., Du, W., Nian, F. *et al.* (2014). An adaptive particle swarm optimization algorithm based on directed weighted complex network. *Mathematical problems in engineering*, vol. 2014.

Li, P. and Mao, K. (2019). Knowledge-oriented convolutional neural network for causal relation extraction from natural language texts. *Expert Systems with Applications*, vol. 115, pp. 512–523.

Li, X., Hu, X., Yu, L., Zhu, L., Fu, C.-W. and Heng, P.-A. (2019). CANet: cross-disease attention network for joint diabetic retinopathy and diabetic macular edema grading. *IEEE transactions on medical imaging*, vol. 39, no. 5, pp. 1483–1493.

Liew, G. and Wang, J.J. (2011). Retinal vascular signs: a window to the heart? *Revista Española de Cardiología (English Edition)*, vol. 64, no. 6, pp. 515–521.

Lin, A., Xu, J., Li, J. and Lu, G. (2022). ConTrans: Improving transformer with convolutional attention for medical image segmentation. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 297–307. Springer.

Lindsay, G.W. (2020). Attention in psychology, neuroscience, and machine learning. *Frontiers in computational neuroscience*, vol. 14, p. 29.

Liu, J., Wang, D., Lu, L., Wei, Z., Kim, L., Turkbey, E.B., Sahiner, B., Petrick, N.A. and Summers, R.M. (2017). Detection and diagnosis of colitis on computed tomography using deep convolutional neural networks. *Medical physics*, vol. 44, no. 9, pp. 4630–4642.

Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S. and Guo, B. (2021). Swin transformer: Hierarchical vision transformer using shifted windows. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 10012–10022.

Lu, T., Zhang, X., Huang, Y., Guo, D., Huang, F., Xu, Q., Hu, Y., Ou-Yang, L., Lin, J., Yan, Z. *et al.* (2020). pFISTA-SENSE-resnet for parallel MRI reconstruction. *Journal of Magnetic Resonance*, vol. 318, p. 106790.

Luong, M.-T., Pham, H. and Manning, C.D. (2015). Effective approaches to attention-based neural machine translation. *arXiv preprint arXiv:1508.04025*.

Ma, J., Fan, X., Yang, S.X., Zhang, X. and Zhu, X. (2018). Contrast limited adaptive histogram equalization-based fusion in YIQ and HSI color spaces for underwater image enhancement. *International Journal of Pattern Recognition and Artificial Intelligence*, vol. 32, no. 07, p. 1854018.

MacGillivray, T., Trucco, E., Cameron, J., Dhillon, B., Houston, J. and Van Beek, E. (2014). Retinal imaging as a source of biomarkers for diagnosis, characterization and prognosis of chronic illness or long-term conditions. *The British journal of radiology*, vol. 87, no. 1040, p. 20130832.

Mach, F., Baigent, C., Catapano, A.L., Koskinas, K.C., Casula, M., Badimon, L., Chapman, M.J., De Backer, G.G., Delgado, V., Ference, B.A. *et al.* (2019). 2019 ESC/EAS guidelines for the management of dyslipidaemias: lipid modification to reduce cardiovascular risk. *Atherosclerosis*, vol. 290, pp. 140–205.

MacKay, D.J. (2003). *Information theory, inference and learning algorithms*. Cambridge university press.

Majib, M.S., Rahman, M.M., Sazzad, T.S., Khan, N.I. and Dey, S.K. (2021). VGG-SCNet: A vgg net-based deep learning framework for brain tumor detection on MRI images. *IEEE Access*, vol. 9, pp. 116942–116952.

Marcos, L., Alirezaie, J. and Babyn, P. (2022). Low dose CT denoising by ResNet with fused attention modules and integrated loss functions. *Frontiers in Signal Processing*, vol. 1, p. 812193.

Marcus, G. (2018). Deep learning: A critical appraisal. *arXiv preprint arXiv:1801.00631*.

Mateen, M., Wen, J., Song, S. and Huang, Z. (2018). Fundus image classification using VGG-19 architecture with PCA and SVD. *Symmetry*, vol. 11, no. 1, p. 1.

Mensah, S.O., Bah, B. and Brink, W. (2021). Towards the localisation of lesions in diabetic retinopathy. In: *Intelligent Computing*, pp. 100–107. Springer.

Midena, E., Frizziero, L., Torresin, T., Boscolo Todaro, P., Miglionico, G. and Pilotto, E. (2020). Optical coherence tomography and color fundus photography in the screening of age-related macular degeneration: A comparative, population-based study. *Plos one*, vol. 15, no. 8, p. e0237352.

Min, B.S., Lim, D.K., Kim, S.J. and Lee, J.H. (2013). A novel method of determining parameters of clahe based on image entropy. *International Journal of Software Engineering and Its Applications*, vol. 7, no. 5, pp. 113–120.

Mittal, K. and Rajam, V. (2020). Computerized retinal image analysis-a survey. *Multimedia Tools and Applications*, vol. 79, no. 31, pp. 22389–22421.

Mnih, V., Heess, N., Graves, A. *et al.* (2014). Recurrent models of visual attention. In: *Advances in neural information processing systems*, pp. 2204–2212.

Mo, J., Zhang, L. and Feng, Y. (2018). Exudate-based diabetic macular edema recognition in retinal images using cascaded deep residual networks. *Neurocomputing*, vol. 290, pp. 161–171.

Moeskops, P., Wolterink, J.M., van der Velden, B.H., Gilhuijs, K.G., Leiner, T., Viergever, M.A. and Išgum, I. (2016). Deep learning for multi-task medical image segmentation in multiple modalities. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 478–486. Springer.

Mohan, S. and Mahesh, T. (2013). Particle swarm optimization based contrast limited enhancement for mammogram images. In: *2013 7th International Conference on Intelligent Systems and Control (ISCO)*, pp. 384–388. IEEE.

Mookiah, M.R.K., Acharya, U.R., Chua, C.K., Lim, C.M., Ng, E. and Laude, A. (2013). Computer-aided diagnosis of diabetic retinopathy: A review. *Computers in biology and medicine*, vol. 43, no. 12, pp. 2136–2155.

More, L.G., Brizuela, M.A., Ayala, H.L., Pinto-Roa, D.P. and Noguera, J.L.V. (2015). Parameter tuning of CLAHE based on multi-objective optimization to achieve different contrast levels in medical images. In: *2015 IEEE International Conference on Image Processing (ICIP)*, pp. 4644–4648. IEEE.

Munk, M.R., Kurmann, T., Marquez-Neila, P., Zinkernagel, M.S., Wolf, S. and Sznitman, R. (2021). Assessment of patient specific information in the wild on fundus photography and optical coherence tomography. *Scientific reports*, vol. 11, no. 1, p. 8621.

Nair, V. and Hinton, G.E. (2010). Rectified linear units improve restricted boltzmann machines. In: *ICML*.

Najafabadi, M.M., Villanustre, F., Khoshgoftaar, T.M., Seliya, N., Wald, R. and Muharemagic, E. (2015). Deep learning applications and challenges in big data analytics. *Journal of big data*, vol. 2, no. 1, pp. 1–21.

Nayak, C., Kaur, L. and Kumar, S. (2013). Retinal blood vessel segmentation algorithm for diabetic retinopathy using wavelet: a survey. *International Journal on Recent and Innovation Trends in Comp and Comm*, vol. 3, no. 3, pp. 927–930.

Nguyen, T.T. and Wong, T.Y. (2009). Retinal vascular changes and diabetic retinopathy. *Current diabetes reports*, vol. 9, no. 4, pp. 277–283.

Nie, D., Gao, Y., Wang, L. and Shen, D. (2018). ASDNet: attention based semi-supervised deep networks for medical image segmentation. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 370–378. Springer.

Nusinovici, S., Rim, T.H., Yu, M., Lee, G., Tham, Y.-C., Cheung, N., Chong, C.C.Y., Da Soh, Z., Thakur, S., Lee, C.J. *et al.* (2022). Retinal photograph-based deep learning predicts biological age, and stratifies morbidity and mortality risk. *Age and ageing*, vol. 51, no. 4, p. afac065.

Oliveira, A., Pereira, S. and Silva, C.A. (2018). Retinal vessel segmentation based on fully convolutional neural networks. *Expert Systems with Applications*, vol. 112, pp. 229–242.

Oquab, M., Bottou, L., Laptev, I. and Sivic, J. (2014). Learning and transferring mid-level image representations using convolutional neural networks. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1717–1724.

Ormazabal, V., Nair, S., Elfeky, O., Aguayo, C., Salomon, C. and Zuñiga, F.A. (2018). Association between insulin resistance and the development of cardiovascular disease. *Cardiovascular diabetology*, vol. 17, no. 1, pp. 1–14.

O'Shea, K. and Nash, R. (2015). An introduction to convolutional neural networks. *arXiv preprint arXiv:1511.08458*.

O'Mahony, N., Campbell, S., Carvalho, A., Harapanahalli, S., Hernandez, G.V., Krpalkova, L., Riordan, D. and Walsh, J. (2019). Deep learning vs. traditional computer vision. In: *Science and information conference*, pp. 128–144. Springer.

Palermo, B.J., D'Amico, S.L., Kim, B.Y. and Brady, C.J. (2022). Sensitivity and specificity of handheld fundus cameras for eye disease: a systematic review and pooled analysis. *Survey of Ophthalmology*, vol. 67, no. 5, pp. 1531–1539.

Pang, Y., Sun, M., Jiang, X. and Li, X. (2017). Convolution in convolution for network in network. *IEEE transactions on neural networks and learning systems*, vol. 29, no. 5, pp. 1587–1597.

Panwar, N., Huang, P., Lee, J., Keane, P.A., Chuan, T.S., Richhariya, A., Teoh, S., Lim, T.H. and Agrawal, R. (2016). Fundus photography in the 21st century—a review of recent technological

advances and their implications for worldwide healthcare. *Telemedicine and e-Health*, vol. 22, no. 3, pp. 198–208.

Park, N. and Kim, S. (2022). How do vision transformers work? In: *International Conference on Learning Representations*.

Pascal, L., Perdomo, O.J., Bost, X., Huet, B., Otálora, S. and Zuluaga, M.A. (2022). Multi-task deep learning for glaucoma detection from color fundus images. *Scientific Reports*, vol. 12, no. 1, pp. 1–10.

Passricha, V. and Aggarwal, R.K. (2018). *Convolutional neural networks for raw speech recognition*. IntechOpen.

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Kopf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J. and Chintala, S. (2019). Pytorch: An imperative style, high-performance deep learning library. In: Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E. and Garnett, R. (eds.), *Advances in Neural Information Processing Systems*, vol. 32. Curran Associates, Inc.
Available at: https://proceedings.neurips.cc/paper_files/paper/2019/file/bdbca288fee7f92f2bfa9f7012727740-Paper.pdf

Pereira, S., Meier, R., Alves, V., Reyes, M. and Silva, C.A. (2018). Automatic brain tumor grading from mri data using convolutional neural networks and quality assessment. In: *Understanding and interpreting machine learning in medical image computing applications*, pp. 106–114. Springer.

Pironkov, G., Dupont, S. and Dutoit, T. (2016). Multi-task learning for speech recognition: an overview. In: *ESANN*.

Pisano, E.D., Zong, S., Hemminger, B.M., DeLuca, M., Johnston, R.E., Muller, K., Braeuning, M.P. and Pizer, S.M. (1998). Contrast limited adaptive histogram equalization image processing to improve the detection of simulated spiculations in dense mammograms. *Journal of Digital imaging*, vol. 11, no. 4, pp. 193–200.

Poplin, R., Varadarajan, A.V., Blumer, K., Liu, Y., McConnell, M.V., Corrado, G.S., Peng, L. and Webster, D.R. (2018). Prediction of cardiovascular risk factors from retinal fundus photographs via deep learning. *Nature Biomedical Engineering*, vol. 2, no. 3, pp. 158–164.

Power, S.P., Moloney, F., Twomey, M., James, K., O'Connor, O.J. and Maher, M.M. (2016). Computed tomography and patient risk: facts, perceptions and uncertainties. *World journal of radiology*, vol. 8, no. 12, p. 902.

Purves, D., Augustine, G.J., Fitzpatrick, D., Katz, L.C., LaMantia, A.-S., McNamara, J.O. and Williams, S.M. (2001). Anatomical distribution of rods and cones. *Neuroscience*.

Qummar, S., Khan, F.G., Shah, S., Khan, A., Shamshirband, S., Rehman, Z.U., Khan, I.A. and Jadoon, W. (2019). A deep learning ensemble approach for diabetic retinopathy detection. *IEEE Access*, vol. 7, pp. 150530–150539.

Raghu, M., Zhang, C., Kleinberg, J. and Bengio, S. (2019). Transfusion: Understanding transfer learning for medical imaging. *Advances in neural information processing systems*, vol. 32.

Rahman, I., Vasant, P.M., Singh, B.S.M. and Abdullah-Al-Wadud, M. (2016). On the performance of accelerated particle swarm optimization for charging plug-in hybrid electric vehicles. *Alexandria Engineering Journal*, vol. 55, no. 1, pp. 419–426.

Rajalakshmi, R., Prathiba, V., Arulmalar, S. and Usha, M. (2021). Review of retinal cameras for global coverage of diabetic retinopathy screening. *Eye*, vol. 35, no. 1, pp. 162–172.

Ramachandran, P., Parmar, N., Vaswani, A., Bello, I., Levskaya, A. and Shlens, J. (2019). *Stand-alone self-attention in vision models*. Curran Associates Inc., Red Hook, NY, USA.

Raman, R., Srinivasan, S., Virmani, S., Sivaprasad, S., Rao, C. and Rajalakshmi, R. (2019). Fundus photograph-based deep learning algorithms in detecting diabetic retinopathy. *Eye*, vol. 33, no. 1, pp. 97–109.

Rao, A., Park, J., Woo, S., Lee, J.-Y. and Aalami, O. (2021). Studying the effects of self-attention for medical image analysis. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 3416–3425.

Reddy, G.T., Bhattacharya, S., Ramakrishnan, S.S., Chowdhary, C.L., Hakak, S., Kaluri, R. and Reddy, M.P.K. (2020). An ensemble based machine learning model for diabetic retinopathy classification. In: *2020 international conference on emerging trends in information technology and engineering (ic-ETITE)*, pp. 1–6. IEEE.

Ridker, P.M., Buring, J.E., Rifai, N. and Cook, N.R. (2007). Development and validation of improved algorithms for the assessment of global cardiovascular risk in women: the reynolds risk score. *JAMA*, vol. 297, no. 6, pp. 611–619.

Ridker, P.M., Paynter, N.P., Rifai, N., Gaziano, J.M. and Cook, N.R. (2008). C-reactive protein and parental history improve global cardiovascular risk prediction: the reynolds risk score for men. *Circulation*, vol. 118, no. 22, pp. 2243–2251.

Rim, T.H., Lee, G., Kim, Y., Tham, Y.-C., Lee, C.J., Baik, S.J., Kim, Y.A., Yu, M., Deshmukh, M., Lee, B.K. *et al.* (2020). Prediction of systemic biomarkers from retinal photographs: development and validation of deep-learning algorithms. *The Lancet Digital Health*, vol. 2, no. 10, pp. e526–e536.

Rodriguez, M., AlMarzouqi, H. and Liatsis, P. (2022). Multi-label retinal disease classification using transformers. *arXiv preprint arXiv:2207.02335*.

Rogers, T.W., Gonzalez-Bueno, J., Garcia Franco, R., Lopez Star, E., Méndez Marín, D., Vassallo, J., Lansingh, V., Trikha, S. and Jaccard, N. (2021). Evaluation of an ai system for the detection of diabetic retinopathy from images captured with a handheld portable fundus camera: the mailor ai study. *Eye*, vol. 35, no. 2, pp. 632–638.

Rosenfeld, A. and Tsotsos, J.K. (2019). Intriguing properties of randomly weighted networks: Generalizing while learning next to nothing. In: *2019 16th Conference on Computer and Robot Vision (CRV)*, pp. 9–16. IEEE.

Roth, G.A., Mensah, G.A., Johnson, C.O., Addolorato, G., Ammirati, E., Baddour, L.M., Barengo, N.C., Beaton, A.Z., Benjamin, E.J., Benziger, C.P. *et al.* (2020). Global burden of cardiovascular diseases and risk factors, 1990–2019: update from the gbd 2019 study. *Journal of the American College of Cardiology*, vol. 76, no. 25, pp. 2982–3021.

Roychowdhury, A. and Banerjee, S. (2018). Random forests in the classification of diabetic retinopathy retinal images. In: *Advanced computational and communication paradigms*, pp. 168–176. Springer.

Ruder, S. (2017*a*). An overview of multi-task learning in deep neural networks. *arXiv preprint arXiv:1706.05098*.

Ruder, S. (2017*b*). Transfer Learning - Machine Learning's Next Frontier. http://ruder.io/transfer-learning/.

Sahu, S., Singh, A.K., Ghrera, S., Elhoseny, M. *et al.* (2019). An approach for de-noising and contrast enhancement of retinal fundus image using clahe. *Optics & Laser Technology*, vol. 110, pp. 87–98.

Salem, N., Malik, H. and Shams, A. (2019). Medical image enhancement based on histogram algorithms. *Procedia Computer Science*, vol. 163, pp. 300–311.

Salkind, N.J. (2005). *Encyclopedia of human development*. Sage Publications.

Sarvamangala, D. and Kulkarni, R.V. (2022). Convolutional neural networks in medical image understanding: a survey. *Evolutionary intelligence*, vol. 15, no. 1, pp. 1–22.

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

Setiawan, A.W., Mengko, T.R., Santoso, O.S. and Suksmono, A.B. (2013). Color retinal image enhancement using CLAHE. In: *International Conference on ICT for Smart Society*, pp. 1–3. IEEE.

Shinohara, Y. (2016). Adversarial multi-task learning of deep neural networks for robust speech recognition. In: *Interspeech*, pp. 2369–2372. San Francisco, CA, USA.

Simonyan, K., Vedaldi, A. and Zisserman, A. (2014). Deep inside convolutional networks: Visualising image classification models and saliency maps. In: *Workshop at International Conference on Learning Representations*.

Simonyan, K. and Zisserman, A. (2015). Very deep convolutional networks for large-scale image recognition. In: *International Conference on Learning Representations*.

Singh, A., Sengupta, S. and Lakshminarayanan, V. (2020). Explainable deep learning models in medical image analysis. *Journal of Imaging*, vol. 6, no. 6, p. 52.

Sinha, A. and Dolz, J. (2020). Multi-scale self-guided attention for medical image segmentation. *IEEE journal of biomedical and health informatics*, vol. 25, no. 1, pp. 121–130.

Smith, L.N. and Topin, N. (2018). Super-convergence: Very fast training of residual networks using large learning rates.

Springenberg, J.T., Dosovitskiy, A., Brox, T. and Riedmiller, M. (2015). Striving for simplicity: The all convolutional net. In: *Workshop at International Conference for Learning Representation*.

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I. and Salakhutdinov, R. (2014). Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research*, vol. 15, no. 1, pp. 1929–1958.

Szegedy, C., Ioffe, S., Vanhoucke, V. and Alemi, A.A. (2017). Inception-v4, inception-resnet and the impact of residual connections on learning. In: *Thirty-first AAAI Conference on Artificial Intelligence*.

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V. and Rabinovich, A. (2015). Going deeper with convolutions. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1–9.

Tan, C., Sun, F., Kong, T., Zhang, W., Yang, C. and Liu, C. (2018). A survey on deep transfer learning. In: *International conference on artificial neural networks*, pp. 270–279. Springer.

Teng, Q., Liu, Z., Song, Y., Han, K. and Lu, Y. (2022). A survey on the interpretability of deep learning in medical diagnosis. *Multimedia Systems*, vol. 28, no. 6, pp. 2335–2355.

Teo, Z.L., Tham, Y.-C., Yu, M., Chee, M.L., Rim, T.H., Cheung, N., Bikbov, M.M., Wang, Y.X., Tang, Y., Lu, Y. *et al.* (2021). Global prevalence of diabetic retinopathy and projection of burden through 2045: systematic review and meta-analysis. *Ophthalmology*, vol. 128, no. 11, pp. 1580–1591.

Thung, K.-H. and Wee, C.-Y. (2018). A brief review on multi-task learning. *Multimedia Tools and Applications*, vol. 77, no. 22, pp. 29705–29725.

Tjandrasa, H., Putra, R.E., Wijaya, A.Y. and Arieshanti, I. (2013). Classification of non-proliferative diabetic retinopathy based on hard exudates using soft margin SVM. In: *2013 IEEE International Conference on Control System, Computing and Engineering*, pp. 376–380. IEEE.

Toğaçar, M., Özkurt, K.B., Ergen, B. and Cömert, Z. (2020). BreastNet: a novel convolutional neural network model through histopathological images for the diagnosis of breast cancer. *Physica A: Statistical Mechanics and its Applications*, vol. 545, p. 123592.

Tolstikhin, I.O., Houlsby, N., Kolesnikov, A., Beyer, L., Zhai, X., Unterthiner, T., Yung, J., Steiner, A., Keysers, D., Uszkoreit, J. *et al.* (2021). Mlp-mixer: An all-mlp architecture for vision. *Advances in Neural Information Processing Systems*, vol. 34, pp. 24261–24272.

Touvron, H., Cord, M. and Jégou, H. (2022). DeIT III: Revenge of the ViT. In: *Computer Vision – ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXIV*, p. 516–533. Springer-Verlag, Berlin, Heidelberg. ISBN 978-3-031-20052-6.

Tripuraneni, N., Jordan, M. and Jin, C. (2020). On the theory of transfer learning: The importance of task diversity. *Advances in Neural Information Processing Systems*, vol. 33, pp. 7852–7862.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł. and Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*, vol. 30.

Verma, K., Deep, P. and Ramakrishnan, A. (2011). Detection and classification of diabetic retinopathy using retinal images. In: *2011 Annual IEEE India Conference*, pp. 1–6. IEEE.

Vu, D.-Q., Wang, C.-Y., Wang, J.-C. *et al.* (2019). Age and gender recognition using multi-task cnn. In: *2019 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, pp. 1937–1941. IEEE.

Wah, C., Branson, S., Welinder, P., Perona, P. and Belongie, S. (2011). The caltech-ucsd birds-200-2011 dataset. Tech. Rep. CNS-TR-2011-001.

Wang, F., Jiang, M., Qian, C., Yang, S., Li, C., Zhang, H., Wang, X. and Tang, X. (2017). Residual attention network for image classification. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3156–3164.

Wang, W. and Gang, J. (2018). Application of convolutional neural network in natural language processing. In: *2018 International Conference on Information Systems and Computer Aided Education (ICISCAE)*, pp. 64–70. IEEE.

Weng, S.F., Reps, J., Kai, J., Garibaldi, J.M. and Qureshi, N. (2017). Can machine-learning improve cardiovascular risk prediction using routine clinical data? *PloS one*, vol. 12, no. 4, p. e0174944.

WHO (2021). Cardiovascular diseases (CVDs).
Available at: https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds)

WHO (2022). Diabetes.
Available at: https://www.who.int/news-room/fact-sheets/detail/diabetes

Woo, S., Park, J., Lee, J.-Y. and Kweon, I.S. (2018). CBAM: Convolutional block attention module. In: *Proceedings of the European conference on computer vision (ECCV)*, pp. 3–19. Springer-Verlag. ISBN 978-3-030-01233-5.

Wu, H., Xiao, B., Codella, N., Liu, M., Dai, X., Yuan, L. and Zhang, L. (2021). CvT: Introducing convolutions to vision transformers. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 22–31.

Wu, J. (2017). Introduction to convolutional neural networks. *National Key Lab for Novel Software Technology. Nanjing University. China*, vol. 5, no. 23, p. 495.

Xie, X., Fan, H., Yu, Z., Bai, H. and Tang, Y. (2022). Weakly-supervised medical image segmentation based on multi-task learning. In: *International Conference on Intelligent Robotics and Applications*, pp. 395–404. Springer.

Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A., Salakhudinov, R., Zemel, R. and Bengio, Y. (2015). Show, attend and tell: Neural image caption generation with visual attention. In: *International conference on machine learning*, pp. 2048–2057. PMLR.

Yadav, S.S. and Jadhav, S.M. (2019). Deep convolutional neural network based medical image classification for disease diagnosis. *Journal of Big Data*, vol. 6, no. 1, pp. 1–18.

Yamashita, T., Asaoka, R., Terasaki, H., Murata, H., Tanaka, M., Nakao, K. and Sakamoto, T. (2020). Factors in color fundus photographs that can be used by humans to determine sex of individuals. *Translational Vision Science & Technology*, vol. 9, no. 2, pp. 4–4.

Yin, W., Kann, K., Yu, M. and Schütze, H. (2017). Comparative study of CNN and RNN for natural language processing. *arXiv preprint arXiv:1702.01923*.

Young, K., Booth, G., Simpson, B., Dutton, R. and Shrapnel, S. (2019). Deep neural network or dermatologist? In: *Interpretability of machine intelligence in medical image computing and multimodal learning for clinical decision support*, pp. 48–55. Springer.

Zeiler, M.D. and Fergus, R. (2014). Visualizing and understanding convolutional networks. In: *European conference on computer vision*, pp. 818–833. Springer.

Zhang, A., Lipton, Z.C., Li, M. and Smola, A.J. (2021). Dive into deep learning. *arXiv preprint arXiv:2106.11342*.

Zhang, L., Yuan, M., An, Z., Zhao, X., Wu, H., Li, H., Wang, Y., Sun, B., Li, H., Ding, S. *et al.* (2020). Prediction of hypertension, hyperglycemia and dyslipidemia from retinal fundus photographs via deep learning: A cross-sectional study of chronic diseases in central china. *PloS one*, vol. 15, no. 5, p. e0233166.

Zhang, S., Fu, H., Yan, Y., Zhang, Y., Wu, Q., Yang, M., Tan, M. and Xu, Y. (2019). Attention guided network for retinal image segmentation. In: *International conference on medical image computing and computer-assisted intervention*, pp. 797–805. Springer.

Zhang, X. and Chutatape, O. (2005). A SVM approach for detection of hemorrhages in background diabetic retinopathy. In: *Proceedings. 2005 IEEE International Joint Conference on Neural Networks, 2005.*, vol. 4, pp. 2435–2440. IEEE.

Zhang, Y. and Yang, Q. (2018). An overview of multi-task learning. *National Science Review*, vol. 5, no. 1, pp. 30–43.

Zhao, H., Jia, J. and Koltun, V. (2020). Exploring self-attention for image recognition. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10076–10085.

Zhao, W. (2017). Research on the deep learning of the small sample data based on transfer learning. In: *AIP Conference Proceedings*, vol. 1864, p. 020018. AIP Publishing LLC.

Zhao, Y., Wang, X., Che, T., Bao, G. and Li, S. (2022). Multi-task deep learning for medical image computing and analysis: A review. *Computers in Biology and Medicine*, p. 106496.

Zhou, B., Khosla, A., Lapedriza, A., Oliva, A. and Torralba, A. (2016). Learning deep features for discriminative localization. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2921–2929.

Zhou, M., Jin, K., Wang, S., Ye, J. and Qian, D. (2017). Color retinal image enhancement based on luminosity and contrast adjustment. *IEEE Transactions on Biomedical engineering*, vol. 65, no. 3, pp. 521–527.

Zhou, Y., Chen, H., Li, Y., Liu, Q., Xu, X., Wang, S., Yap, P.-T. and Shen, D. (2021). Multi-task learning for segmentation and classification of tumors in 3D automated breast ultrasound images. *Medical Image Analysis*, vol. 70, p. 101918.

Zhu, J., Zhang, E. and Del Rio-Tsonis, K. (2012). Eye anatomy. *eLS*.

Zhu, Y., Zhou, Y., Xu, H., Ye, Q., Doermann, D. and Jiao, J. (2019). Learning instance activation maps for weakly supervised instance segmentation. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 3116–3125.

Zhuang, F., Qi, Z., Duan, K., Xi, D., Zhu, Y., Zhu, H., Xiong, H. and He, Q. (2020). A comprehensive survey on transfer learning. *Proceedings of the IEEE*, vol. 109, no. 1, pp. 43–76.

Zou, K.H., O'Malley, A.J. and Mauri, L. (2007). Receiver-operating characteristic analysis for evaluating diagnostic tests and predictive models. *Circulation*, vol. 115, no. 5, pp. 654–657.

Zuiderveld, K. (1994). Contrast limited adaptive histogram equalization. *Graphics gems*, pp. 474–485.