

University of Dundee

DOCTOR OF PHILOSOPHY

Investigating the retina as a source of biomarkers for systemic conditions using artificial intelligence

Syed, Mohammad Ghouse

Award date:
2023

Licence:
Copyright of the Author. All Rights Reserved

[Link to publication](#)

General rights

Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

- Users may download and print one copy of any publication from the public portal for the purpose of private study or research.
- You may not further distribute the material or use it for any profit-making activity or commercial gain
- You may freely distribute the URL identifying the publication in the public portal

Take down policy

If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.

University of Dundee

DOCTOR OF PHILOSOPHY

Investigating the retina as a source of biomarkers for systemic conditions using artificial intelligence

Syed, Mohammad Ghouse

Award date:
2023

[Link to publication](#)

General rights

Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

- Users may download and print one copy of any publication from the public portal for the purpose of private study or research.
- You may not further distribute the material or use it for any profit-making activity or commercial gain
- You may freely distribute the URL identifying the publication in the public portal

Take down policy

If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.

Investigating the retina as a source of biomarkers for systemic conditions using artificial intelligence



**University
of Dundee**

Mohammad Ghouse Syed

Supervisors: Prof. Emanuele Trucco, Computing, School of Science and Engineering

Dr. Alexander Doney, School of Medicine

Prof. Stephen McKenna, Computing, School of Science and Engineering

This dissertation is submitted in fulfilment of the requirements for the degree

of

Doctor of Philosophy

School of Medicine

May 2023

I would like to dedicate this thesis to my parents, wife, kids, siblings . . .

Declaration of Authorship

Candidate's Declaration

I, Mohammad Ghouse Syed, hereby declare that I am the author of this thesis; that all references cited have been consulted by me; that the work of which this thesis is a record has been done by me, and that it has not been previously accepted for a higher degree.

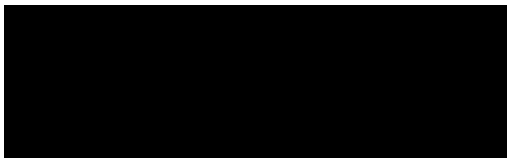
Signature

8th March 2023

Date

Supervisor's Declaration

I, Emanuele Trucco, hereby declare that I am the supervisor of the candidate, and that the conditions of the relevant Ordinance and Regulations have been fulfilled.



Signature

8th March 2023

Date

Acknowledgements

First and foremost, I would like to thank my wonderful supervisor Prof. Emanuele Trucco for the opportunity of doing my PhD under his esteemed guidance, patience, continuous support (professionally and personally), encouraging learning technical skills that helped me to carry my research. I learned many things from him and the ones I admire most are his efficient and productive ways of handling multiple projects, post-docs and students; and his swift ability to understand the content of technical presentations. It is my absolute privilege to work with him and I will definitely try to implement the things learnt into my life and career.

I thank my second supervisor, Dr. Alexander Doney of the NHS Ninewells Hospital and Medical School, for his indispensable support in providing the GoDARTS data that was crucial for my thesis work and in providing much knowledge and guidance on the clinical context. Without that, I would have been lost in the ocean of the medical world. Dr Doney motivated and encouraged me in the direction of contemporary research related to retina and cardiovascular diseases.

I thank my other mentor, Dr. Ify Mordi, also at Ninewells, for his immense help in providing the clean data needed for my analysis as well as excellent research ideas which led to a publication being prepared. He was very generous with his time and introduced me to various topics related to the medical domain and statistical analysis.

I thank also Prof. Stephen McKenna of the Computer Vision and Image Processing (CVIP) group in Computing for his discussions and critical inputs related to deep learning, and Dr. V. Prathiba at the Madras Diabetes Research Foundation, Chennai, India for discussions and inputs related to retinas.

I am grateful to National Institute for Health Research (NIHR) for funding my PhD study within the INSPIRED Global Health Scotland-India Unit program. I thank Prof. Colin Palmer, INSPIRED PI, for his continuous support in approving the required training needs and conference travels. I thank Dr. Shona Mathews, Isobel Ovens, Dr. Fred Comerford for their guidance and for providing excellent administrative support.

I thank Mahamadou Niakate, Computing Officer in Computing, for his tremendous help with the technical support, especially during the COVID-19 restrictions. I thank Dr. Joseph Ward, Senior DevOps Engineer at Health Informatics Centre, University of Dundee, for his great help with the initial set-up of the deep learning environment in the GPU node of safe haven.

It was a pleasure to be a part of the CVIP Group at Dundee. I am grateful to all CVIP members for their useful technical discussions, especially during journal meetings. I thank also the whole VAMPIRE team (Dundee and Edinburgh) for the technical discussions at the VAMPIRE plenaries. I thank all VAMPIRE Dundee team members, especially Stephen Hogg and Dr Muthu R. K. Mookiah for many useful technical discussions and feedback. Special thanks to Muthu for his overall guidance and brotherly support all the time.

I thank my friends and fellow PhD students from the INSPIRED project, Mehul, Sushrima, Jyothsna and Gittu, for their support and help whenever needed; special thanks to Anand, Aravind and Charvi for excellent discussions and teachings on several topics and for hosting me when needed. Thank you to Aditya Nar for providing the data for my analysis.

Finally, and most importantly, a very special thank you to my family, my parents, Sulthan Bee and Khader Basha: whatever I am today is because of you. My wonderful and beautiful wife, Shaik Afrin Sultana: thank you for allowing me to do this PhD, taking care of our kids, and for your overall tremendous support. I could not have achieved this without you. To my brothers, Ali and Ghazi: thank you for supporting me all the time. Finally, to my lovely kids, Zoya and Ziyar: thank you for bearing with my absence during my studies.

List of Publications and Presentations

1 Publications from Project

- (a) **Syed, M. G.**, Wang, H., Trucco, E., Huang, Y., Mordi, I., and Doney, A. Biological Vascular Age from Retinal Photographs Predicts All-Cause Death and Cardiovascular Events: a GoDARTS study (under preparation for Journal submission).
- (b) **Syed, M. G.**, Trucco, E., Doney, A., and Mordi, I. Integrating a Deep-Learning Cardiovascular Risk Score Derived from Retinal Images and a Coronary Heart Disease Polygenic Risk Score to Predict Clinical Outcomes (under preparation for Journal submission).
- (c) Mordi, I. R., Trucco, E., **Syed, M. G.**, MacGillivray, T., Nar, A., Huang, Y., George, G., Hogg, S., Radha, V., Prathiba, V., et al. (2022). Prediction of major adverse cardiovascular events from retinal, clinical, and genomic data in individuals with type 2 diabetes: A population cohort study. *Diabetes Care*, 45(3):710–716.
- (d) **Syed, M. G.**, Doney, A., George, G., Mordi, I., and Trucco, E. (2021). Are cardiovascular risk scores from genome and retinal image complementary? A deep learning investigation in a diabetic cohort. In *International Workshop on Ophthalmic Medical Image Analysis - OMIA (MICCAI workshops)*, pages 109–118. Springer.

2 Other publications

- (a) Huang, Y., **Syed, M. G.**, et al., Genomic Determinants of Biological Age Estimated By Deep Learning Applied to Retinal Images (under preparation for Journal submission).
- (b) Soca, A., **Syed, M. G.**, Trucco, E., Harvey, J., and Doney, A. Prediction of dementia outcome from retinal and genomic data in the GoDARTS cohort (accepted at AD/PD 2023, International Conference on Alzheimer's and Parkinson's Diseases and related neurological disorders).

3 Unrefereed Presentations

Poster presentations

- **Syed, M. G.**, and Trucco, E. (2019). A framework to generate synthetic test image sets parameterized by difficulty level. Medical Image Understanding and Analysis: 23rd Annual Conference, MIUA 2019 (invited talk).
- **Syed, M. G.**, and Trucco, E. (2019). A framework to generate synthetic test image sets parameterized by difficulty level. Scottish Imaging Network, A Platform for Scientific Excellence: SINAPSE 2019.
- **Syed, M. G.** (2018). Convolutional neural network architectures for image classification and detection, GoDARTS 20th Anniversary celebration seminar in Pitlochry.

Others

- *Blitz talk*: Synthetic data generation and fine-tuning of convolutional neural networks. **SINAPSE Annual Scientific Meeting, Dundee**, June 2019.
- *Oral presentation*: PhD research work progress presented at annual **PhD Symposium, Computing, School of Science and Engineering, University of Dundee** (2018 - 2021).

- *Oral presentation:* PhD research work progress presented at annual **PhD Symposium, School of Medicine, University of Dundee** (2018 - 2021).
- Presented work progress regularly during my PhD studies (2018 - 2021) at:
 - Bi-annual CVIP technical workshops, Dundee/online.
 - Quarterly VAMPIRE Plenary, Dundee/Edinburgh/online.
 - Bi-annual INSPIRED PhD Symposium, Dundee/online.
- Presented PhD research work to general public who attended **Doors Open Event**, Ninewells Hospital, Dundee, 2019.

List of Training Events

The following training events have been attended during my PhD studies:

- Full day workshop in NRS Good Clinical Practice, Ninewells hospital, Dundee, 2018.
- Medical Image Computing Summer School (MedICSS), 9th July - 13th July 2018, University College London (UCL), London, UK.
- Medical Image Understanding and Analysis (MIUA), virtual conference, 2020 and 2021.
- International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI), virtual conference, 2020 and 2021.
- Bi-weekly journal meetings, Computing, University of Dundee/online (2018 - 2021).
- Monthly GoDARTS group and journal meetings, Ninewells hospital, Dundee/online (2018 - 2021).
- Several Organisational and Professional Development (OPD) courses conducted by University of Dundee.
- Internal meetings and discussions with fellow PhD students and faculty at Computing/Ninewells hospital, University of Dundee.

Abstract

According to World Health Organization (WHO), Cardiovascular Disease (CVD), diabetes, and Chronic Kidney Disease (CKD) are among the top 10 causes of death worldwide. CVDs are the leading cause of death globally with an estimated 17.9 million deaths which constitute 32% of global mortality. The retina is the only organ in the human body that provides direct observation of a risk portion of the microvasculature, hence a unique opportunity for the non-invasive study of several systemic diseases including CVD, stroke, diabetes, hypertension, Diabetic Retinopathy (DR), CKD, and dementia. Retinal investigation can enable early diagnosis, preventive measures, and planning better treatment. There has been increasing interest in applying deep learning (DL) to retinal images to identify the risk of CVDs, its risk factors, and their interplay with cardiovascular risk scores.

This thesis investigates retinal images as a source of biomarkers for identifying associations with cardiovascular diseases, cardiovascular risk factors, cardiovascular risk scores, cardiovascular death, mortality, microvasculature diseases (CKD, Diabetic Peripheral Neuropathy (DPN), DR) in a large diabetic cohort, GoDARTS (Genetics of Diabetes Audit and Research in Tayside Scotland) using DL. A breakdown of the main items of work follows.

A framework is proposed for generating synthetic datasets, parameterized by difficulty level to test classifiers for medical image analysis. This framework was used for hyperparameter tuning of the DL model and identifying a robust model for subsequent work with real data. To our best knowledge, this had not been addressed in the literature on synthetic medical data at the time the work was carried out. EfficientNet-B2 was chosen as the best-performing DL architecture to perform experiments on real retinal images from GoDARTS, following a systematic and comparative performance analysis using synthetic data generated

from MESSIDOR images. Several architectures including VGG16, ResNet50, InceptionV3, DenseNet201, and EfficientNet-B2 were evaluated during the analysis.

DL methods were employed to investigate biomarkers in retinal images for predicting cardiovascular (CV) risk factors, such as age, sex, Systolic Blood Pressure (SBP), Diastolic Blood Pressure (DBP), High-Density Lipoprotein (HDL), Total Cholesterol (TC), Glycated Haemoglobin (GH), Body Mass Index (BMI), and Triglycerides (Trig) using only retinal images from GoDARTS. The findings indicated that for age prediction on the test dataset, the Mean Absolute Error (MAE) was 3.951 (95% CI 3.908, 3.995) years and R^2 was 0.809 (0.804, 0.814). The model achieved an Area Under Curve (AUC) of 0.899 (0.895, 0.903) for sex prediction, with an accuracy of 0.811 (0.806, 0.817), a sensitivity of 0.886 (0.88, 0.891), and specificity of 0.717 (0.708, 0.727). However, the DL models did not yield significant results for predicting other CV risk factors, with an MAE of 5.88 (5.71, 6.07) and R^2 of 0.14 (0.1, 0.17) for DBP prediction. Additionally, the model's performance on test data for stratifying systemic disease outcomes within 12 years from the date of retinal imaging in terms of Area Under Curve (AUC) was 0.74, 0.71, 0.642, 0.633, and 0.57 for CKD, ACD, MACE, DR, and DPN, respectively.

Furthermore, a study was conducted to investigate the Predicted Age Difference (PAD), which is the difference between chronological age and the age predicted by DL using retinal images, in individuals with type 2 diabetes. The aim was to find any association between PAD and Major Adverse Cardiovascular Event (MACE) and All-Cause Death (ACD) over a period of 10 years. The results showed a strong statistically significant association. According to Coxph regression analysis that was adjusted for age at imaging and sex, a 1-year increase in retinal PAD score raised the risk of MACE and ACD by 5.8% (hazard ratio (HR) = 1.0587, 95% CI = 1.028 - 1.089, $P = 1.06e-4$) and 5.9% (HR = 1.0597, 95% CI = 1.034 - 1.085, $P = 1.62e-06$) respectively. Even after adjusting the coxph model for Pooled Cohort Equations (PCE) Atherosclerotic Cardiovascular Disease (ASCVD) risk score, the associations still remained significant for MACE and ACD events with PAD score. These findings were similar to the results obtained by using the average of predictions from both left and right

retinal images (individual-level predictions) when assessing the age predictions from only left-eye and only right-eye retinas.

We investigated the potential of a DL approach applied to retinal images for predicting PCE ASCVD clinical risk score and genetic factors, as represented by Genome-Wide Polygenic Risk Scores (GWPRS). The model achieved an R^2 of 0.554 (0.528, 0.579) and MAE of 0.107 (0.104, 0.11) for PCE ASCVD risk score, but showed no indication that retinal images contain information related to GWPRS. This investigation represents the first time that the complementarity of retinal and genetic information for CVD risk has been studied using DL. Statistically significant associations were observed between the clinical risk score predicted by the DL model from retinal images and 10-year MACE and Cardiovascular Death (CV death). The Coxph regression analysis showed that a 1% increase in the retinal predicted risk score increases the risk of developing MACE (HR = 1.029, 95% CI = 1.015 - 1.042, P = 3.4e-5) and CV death (HR = 1.019, 95% CI = 1.004 - 1.034, P = 0.009). Again, these associations are similar to the results obtained using individual-level predictions from only left-eye and only right-eye retinal images.

We propose a method for utilizing DL to perform multi-modal image analysis, combining images and tabular (spreadsheet) data. Our method involves converting tabular data to images using a tabular data converted to image (T2I) algorithm. We applied this method to the available CV risk factors in the form of spreadsheet data and combined it with retinal images for stratifying 5-year MACE from the date of retinal imaging. When using only retinal images as input, the DL model achieved an AUC of 66.78% on the test dataset for 5-year MACE stratification. However, when using multi-modal image data (retina + T2I), the DL model achieved a higher AUC of 72.54%. These results suggest that utilizing multi-modal image data has more predictive power for predicting 5-year MACE compared to using retinal images alone.

To summarize, the use of DL algorithms on diabetic retinopathy screening images enables accurate predictions of age, sex, DBP, and CVD risk score. Additionally, the algorithms can moderately stratify certain systemic conditions such as MACE, CKD, and ACD. However,

further research is needed to validate these results on a larger population and to explore the implementation of these approaches in real-time clinical practices.

Table of contents

Acknowledgements	iv
List of Publications and Presentations	vi
List of Training Events	ix
Abstract	x
List of figures	xix
List of tables	xxiv
Nomenclature	xxix
1 Introduction	1
1.1 Background and motivation	1
1.1.1 Motivation for this work	1
1.2 Systemic diseases	3
1.2.1 Cardiovascular diseases	3
1.2.2 Diabetes	4
1.2.3 Chronic kidney disease	4
1.2.4 Eye examination	5
1.3 Retinal imaging	5
1.4 Deep Learning	8

1.4.1	Basic components of a convolutional neural network	11
1.4.2	Early deep learning architectures	13
1.4.3	Visualizing convolutional neural networks	14
1.5	List of aims	15
1.6	Structure of the thesis	16
2	Related work	18
2.1	About this chapter	18
2.2	Retinal vasculature	18
2.3	ML and DL in retinal image analysis	21
2.4	DL and retinal images biomarkers	23
2.5	Conclusions	30
3	Material	31
3.1	About this chapter	31
3.2	Public datasets	31
3.2.1	MESSIDOR	32
3.2.2	IDRID	32
3.3	Clinical dataset: GoDARTS	35
3.3.1	GoDARTS history and structure	35
3.3.2	Pre-processing	40
3.4	Conclusions	42
4	Identifying robust CNN architecture	46
4.1	Introduction	46
4.1.1	Context and motivation	46
4.1.2	Generating synthetic medical images	47
4.1.3	Parameterizing synthetic datasets by difficulty level	47
4.1.4	Fine-tuning and validation	48
4.2	Proposed framework for synthetic data generation	48

4.3	Materials and methods	49
4.3.1	Pre-processing	50
4.3.2	Dictionary of lesion patches	50
4.3.3	Parameter space	51
4.3.4	Classifier	53
4.3.5	CNN Training	55
4.4	Results	57
4.5	Validating CNN models on IDRID	58
4.5.1	Pre-processing	60
4.5.2	CNN Training	60
4.5.3	Results	61
4.6	Discussions	63
4.7	Conclusions	64
5	Predicting demographic and clinical features	66
5.1	About this chapter	66
5.2	Demographic features	67
5.2.1	Materials	67
5.2.2	Methods	67
5.2.3	Results	73
5.3	Retinal predicted age and chronological age	78
5.3.1	Introduction	78
5.3.2	Materials	78
5.3.3	Methods (survival analysis)	81
5.3.4	Results	85
5.4	Clinical measurements	95
5.4.1	Materials	95
5.4.2	Methods	98
5.4.3	Results	98
5.5	Disease outcomes	100

5.5.1	Materials	100
5.5.2	Methods	102
5.5.3	Results	102
5.6	Discussions	104
5.7	Conclusions	107
6	Predicting cardiovascular risk scores	110
6.1	About this chapter	110
6.2	Introduction	110
6.3	Materials	112
6.3.1	Outcome variables	112
6.4	Methods	114
6.4.1	Pre-processing	114
6.4.2	Deep learning architecture and training	114
6.5	Prediction of CV risk scores	115
6.5.1	Results	115
6.6	Prediction of MACE and CV death	116
6.6.1	Materials	117
6.6.2	Methods	118
6.6.3	Results on PCE Risk prediction at baseline	119
6.6.4	Results on MACE and CV death in high PRS group	120
6.7	Discussions	123
6.8	Conclusions	126
7	Converting tabular data to images for deep learning	127
7.1	About this chapter	127
7.2	Introduction and motivation	127
7.3	Materials	129
7.3.1	Outcome variables	129
7.4	Methods	132

7.4.1	Tabular data to image conversion	132
7.4.2	IGTD implementation	134
7.4.3	Deep learning architecture and training	135
7.4.4	Machine learning methods and training	137
7.5	Results	137
7.5.1	Tabular data to image conversion	139
7.5.2	Optimal feature ordering	140
7.5.3	ML methods for MACE	141
7.5.4	Retina and T2I for MACE	142
7.6	Discussions	145
7.7	Conclusions	146
8	Conclusions and future work	148
8.1	Summary of work	148
8.2	Contributions	154
8.3	Limitations and future work	156
8.3.1	Synthetic data generation	156
8.3.2	DL architecture	157
8.3.3	Explainability of DL	157
8.3.4	GoDARTS	158
8.3.5	Generalizability of results	160
	References	161
	Appendix A Additional figures and results	184

List of figures

1.1	(a) Connection of the eye to the brain. (b) The main parts of the eye for the purpose of this thesis.	6
1.2	A Topcon TRC 50 EX fundus camera [1].	8
1.3	A sample monochrome retinal image illustrating the different fields of views [2].	9
1.4	A cartoon drawing comparing biological neuron (left) and a common mathematical neuron model (right) [3].	10
1.5	A sample 3-layered ANN with 3 inputs, 2 hidden layers, and 1 output layer [3].	11
1.6	A typical CNN architecture [4].	11
1.7	An input image of 5x5 dimensions convolved with a 3x3 filter.	12
1.8	Plots of common activation functions [5].	13
1.9	An example of feature extraction. An input image (or a feature map) is passed through a filter bank, followed by non-linearity and spatial pooling/sub-sampling [6].	13
1.10	The CAMs of four classes from ILSVRC [7].	15
2.1	A sample fundus image illustrating important anatomical landmarks and lesions [8].	19
2.2	Illustration of retinal vasculature parameters in fundus image [9].	20
2.3	Retinal image zones used by VAMPIRE [10].	21
2.4	Process involved in building of analytical model [11].	22
2.5	Recent trend and distribution of papers [8].	23

3.1	Sample MESSIDOR fundus images [12].	33
3.2	Sample IDRID fundus images [13].	43
3.3	Sample GoDARTS fundus images of left- and right-eyes of different individuals (I).	44
3.4	Block diagram for image pre-processing in GoDARTS.	45
3.5	Example of image pre-processing in GoDARTS.	45
4.1	Measurements for maximum square inscribed in a circular retina.	50
4.2	Image pre-processing in MESSIDOR: (a) Original MESSIDOR image; (b) largest inscribed square (retinal region considered); (c) image resized to 256×256.	51
4.3	Sample patches for each lesion type, cropped from retinal images graded 1 to 3; Rows (top to bottom): EX, HE, MA; Columns (left to right): patch dimensions - 9×9, 17×17, 25×25.	52
4.4	Examples of augmented images from datasets at various points in parameter space, with coordinates [patch size, number of patches, transparency].	54
4.5	Visualization of the performance of <i>VGG16_partially_trained</i> in 3-D parameter space, using color coding and two viewpoints for clarity (180° rotation around a vertical axis through the center of the ground plane). The region of the parameter space challenging the classifier is clearly visible.	57
4.6	Visualization of the performance of <i>VGG16_transfer_learning</i> in 3-D parameter space. Color coding and viewpoints as per the previous figure. The improvement with respect to Figure 4.5 is immediately visible.	58
4.7	Comparative performance visualization of all modified CNNs in 3-D parameter space.	59
4.8	Example of image pre-processing in IDRID.	61
5.1	Scatter plot for actual age at imaging and predicted age at imaging in the complete test data and its subset.	74
5.2	Sample grad-CAM heatmaps from model trained for age prediction.	75

5.3	Sample grad-CAM heatmaps from model trained for predicting sex (correct predictions).	76
5.4	Pixel wise mean and standard deviation heatmaps with a common color bar.	79
5.5	Pixel wise mean and standard deviation heatmaps with individual color bars.	80
5.6	Scatter plots for actual age at imaging and predicted age at imaging in the whole test data and its subset in T2D individuals. BL = baseline.	87
5.7	Sample grad-CAM heatmaps of two individuals from a model trained for age prediction in T2D individuals. I1 = Individual 1, I2 = Individual 2. . . .	89
5.8	KM curves of right-censored survival data for upper (high PAD) and lower (low PAD) tertiles of retinal PAD groups computed from the average of the left eye and right eye image predictions for age.	90
5.9	KM curves of right-censored survival data for groups of predicted age computed from the average of the left eye and right eye image predictions for age.	93
5.10	KM curves for mortality from right censored survival data for upper (high τ_{rate}) and lower (low τ_{rate}) tertiles of τ_{rate} groups computed from the average of the left eye and right eye image predictions for age at last and first available retinal images. τ_{rate} = PALFD rate.	94
5.11	Scatter plot for actual and predicted labels of clinical measurements in the whole test data. The green line is a diagonal line.	101
6.1	Scatter plot depicting the actual and predicted labels in the test data for CV risk score and GWPRS. The blue line is a diagonal line.	116
6.2	Sample grad-CAM heatmaps for PCE risk prediction. The top row is right-eye images. The bottom row is left-eye images.	117
6.3	Scatter plot depicting the actual and predicted labels in the test data for clinical CV risk score. The blue line is a diagonal line.	120
6.4	KM curves of right-censored survival data for upper (high PRS), middle and lower (low PRS) tertiles of retinal PRS groups derived from individual-level prediction at baseline.	122

6.5	KM curves of right-censored survival data for top 20% and bottom 80% of the retinal Ω_{rate} groups computed from individual-level predictions for risk at last and first available retinal images.	124
7.1	Block diagram of T2I conversion.	133
7.2	An example showing conversion of a sample tabular data with 6 features to a 2D gray image. Left: Sample tabular data. Middle: Dedicated feature space in the image area. Right: Tabular data to image conversion for row no. 2	134
7.3	DL model 1: A modified EfficientNet-B2 with image as input and sigmoid output. The GAP layer gives a vector of length 1,408.	136
7.4	DL model 2: A multi-modal DL architecture with retinal image and T2I image as input and sigmoid output. The GAP layer results in a vector of length 1,408 and the concatenation layer outputs a vector of length 2,816.	136
7.5	DL model 3: Modified version of DL model 2 with an additional dropout layer before the output layer.	136
7.6	DL model 4: Modified version of DL model 3 by replacing dropout layer with FC layer of 100 nodes.	137
7.7	DL model 5: Modified version of DL model 4 by introducing dropout layer before and after FC layer.	137
7.8	An example of T2I conversion with a sample subset data.	139
7.9	An illustration of IGTD algorithm on the GoDARTS tabular data with square error function.	141
7.10	An example of T2I with feature ordering.	142
7.11	A T2I example representation of a single row data with randomly shuffling the features in 10 different ways.	144
A.1	Sample grad-CAM heatmaps of three individuals from model trained for age prediction in T2D individuals. I1 = Individual 1, I2 = Individual 2, I3 = Individual 3.	188

A.2	KM curves of right-censored survival data for upper (high PAD) and lower (low PAD) tertiles of retinal PAD groups using only left eye image predictions for age.	189
A.3	KM curves of right-censored survival data for upper (high PAD) and lower (low PAD) tertiles of retinal PAD groups using only right eye image predictions for age.	190
A.4	KM curves of right-censored survival data for groups of predicted age computed from only left eye retinal images	192
A.5	KM curves of right-censored survival data for groups of predicted age computed from only right eye retinal images	193
A.6	KM curves for mortality from right censored survival data for upper (high τ_{rate}) and lower (low τ_{rate}) tertiles of τ_{rate} groups computed from only left eye predictions. τ_{rate} = PALFD rate.	194
A.7	KM curves for mortality from right censored survival data for upper (high τ_{rate}) and lower (low τ_{rate}) tertiles of τ_{rate} groups computed from only right eye predictions. τ_{rate} = PALFD rate.	195
A.8	More examples of sample grad-CAM heatmaps for PCE risk prediction. . .	196
A.9	KM curves of right censored survival data for upper (high PRS), middle and lower (low PRS) tertiles of retinal PRS groups derived from only left retina. . .	197
A.10	KM curves of right-censored survival data for upper (high PRS), middle and lower (low PRS) tertiles of retinal PRS groups derived from only right retina. . .	198
A.11	KM curves of right censored survival data for top 20% and bottom 80% of the retinal Ω_{rate} groups computed from only left eye image predictions. . .	199
A.12	KM curves of right censored survival data for top 20% and bottom 80% of the retinal Ω_{rate} groups computed from only right eye image predictions. . .	200

List of tables

1.1	Different types of activation functions.	12
3.1	MESSIDOR criteria for grading for DR and DME [12]; n_MA: number of MA, n_HE: number of HE, NV = 1: NV, NV = 0: no NV.	34
3.2	IDRID criteria for grading for DR and DME [13]; NPDR, PDR.	35
3.3	Descriptive characteristics of continuous features in GoDARTS; M = Male; F = Female; n = individuals used for feature	40
3.4	Descriptive characteristics of binary features in GoDARTS; M = Male; F = Female; n = individuals used for feature.	41
4.1	Summary of modified CNN models.	55
4.2	Summary of training specifications on MESSIDOR data.	56
4.3	DR class distribution in the IDRID data splits.	62
4.4	Summary of training specifications on IDRID data.	62
4.5	Performance of modified CNNs on IDRID test data.	63
5.1	Dataset characteristics of whole cohort and data splits used for predicting demographic features.	68
5.2	Data distribution for age subgroups in the whole cohort and data splits. In parenthesis, the value includes the proportion of data in the respective column.	69
5.3	Summary of training specifications for predicting age and sex using all retinal images in GoDARTS.	70
5.4	Confusion matrix terminology.	72

5.5	Model’s performance for predicting age on complete test data. 95% CI values are computed using bootstrap samples.	74
5.6	Age sub-group-wise model performance with mean actual age, mean predicted age, MAE, and its 95% CI for the age sub-groups with an interval of 10 years on the complete test data.	75
5.7	Model’s performance for predicting sex using all test data images. 95% CI values are computed using bootstrap samples.	76
5.8	Number of images in sub-groups.	77
5.9	Dataset characteristics of whole cohort and data split in T2D participants.	81
5.10	Data distribution for age subgroups in the whole cohort and data splits of T2D individuals. In parenthesis, the value includes the proportion of data in the respective column.	82
5.11	Summary of training specifications for predicting age using all retinal images in T2D individuals of GoDARTS.	83
5.12	Model’s performance for predicting age in T2D individuals. 95% CI values are computed using bootstrap samples.	86
5.13	Age sub-group-wise model performance with mean actual age, mean predicted age, MAE, and its 95% CI for the age sub-groups with an interval of 10 years on the complete test data of T2D individuals.	88
5.14	Retinal Image distribution of whole cohort and data splits for predicting continuous clinical measurements.	95
5.15	Baseline characteristics of clinical measurements in GoDARTS; n = individuals used for the feature.	97
5.16	Summary of training specifications for predicting continuous clinical features using retinal images.	99
5.17	Model’s performance for predicting continuous clinical features. Bootstrap samples are utilized to estimate the 95% confidence interval (CI) values.	100
5.18	Retinal Image distribution of whole cohort and data splits for predicting microvascular complications.	101

5.19	Baseline characteristics of Disease outcomes in GoDARTS; n = individuals used for the feature. prop. = proportion.	102
5.20	Summary of training specifications for predicting disease outcomes using retinal images.	103
5.21	Model's performance for predicting disease outcomes. 95% CI values are computed using bootstrap samples.	104
6.1	Baseline characteristics of CV risk scores in GoDARTS; n = individuals used for the feature.	113
6.2	Summary of training specifications for predicting CV risk scores using retinal images.	114
6.3	Model performance on estimating PCE ASCVD and genetic risk scores in the test data. 95% CI values are computed using 2,000 bootstrap samples. . .	115
6.4	The performance of the models in estimating ASCVD risk scores using the test data, along with 95% CI computed using 2,000 bootstrap samples. . . .	120
7.1	Retinal Image distribution of whole cohort and data splits for predicting 5-year MACE.	130
7.2	Baseline characteristics of the dataset for 5-years MACE prediction in GoDARTS; n = individuals used for the feature.	131
7.3	Summary of training specifications for predicting 5-year MACE using retinal images.	138
7.4	Training specifications for IGTD.	140
7.5	Optimal feature ordering with IGTD on 9 features and 10 features.	142
7.6	ML models performance for classification of 5-years MACE in the test dataset. 9 features = age, sex_male, DBP, SBP, BMI, GH, HDL, TC, eye. 10 features = 9 features + Genome-Wide Polygenic Risk Scores (GWPRS). . .	143
7.7	DL model 1 performance for classification of 5-years MACE in the test dataset. 9f = age, sex_male, DBP, SBP, BMI, GH, HDL, TC, eye. 10f = 9f + GWPRS.	143

7.8	DL model 2, 3, 4, and 5 performance for classification of 5-years MACE in the test dataset. 9f = age, sex_male, DBP, SBP, BMI, GH, HDL, TC, eye. 10f = 9f + GWPRS.	145
A.1	Age sub-group-wise model performance on the complete test data using only left eye images.	184
A.2	Age sub-group-wise model performance on the complete test data using only right eye images.	185
A.3	Age sub-group-wise model performance on the test data of T2D individuals using both left and right eye image predictions at baseline.	185
A.4	Age sub-group-wise model performance on the test data of T2D individuals using only left eye image predictions at baseline.	186
A.5	Age sub-group-wise model performance on the test data of T2D individuals using only right eye image predictions at baseline.	186
A.6	Age sub-group-wise model performance on the test data of T2D individuals using the average of left and right eye image predictions at baseline.	187
A.7	Coxph regression results adjusted for age, and sex for predicting MACE and ACD using PAD from only left- and right-eye retinal images. PAD score is a continuous variable. The results of PAD Middle and high tertile are with respect to PAD low tertile group.	187
A.8	Coxph regression results adjusted for CV risk score for predicting MACE and ACD using PAD from only left- and right-eye retinal images. PAD score is a continuous variable. The results of PAD Middle and high tertile are with respect to PAD low tertile group.	191
A.9	Coxph regression results adjusted for sex for predicting MACE and ACD using DLPA and CA from only left- and right-eye retinal images. DLPA = DL predicted age, CA = chronological age. DLPA and CA are continuous variables.	191

-
- A.10 Coxph regression results adjusted for predicted age and sex for predicting ACD using τ_{rate} from only left- and right-eye retinal images. $\tau_{rate_{lt}} = \tau_{rate}$ low tertile, $\tau_{rate_{mt}} = \tau_{rate}$ middle tertile, $\tau_{rate_{ht}} = \tau_{rate}$ high tertile. The results of $\tau_{rate_{mt}}$ group and $\tau_{rate_{ht}}$ group are with respect to $\tau_{rate_{lt}}$ group. 191
- A.11 Coxph regression results adjusted for age, sex and GWPRS for predicting MACE and CV Death using PRS from only left- and right-eye retinal images. PRS score is a continuous variable. The results of PRS middle and high tertile are with respect to PRS low tertile group. 194
- A.12 Coxph regression results adjusted for predicted age, sex and GWPRS for predicting MACE and CV Death using Ω_{rate} from only left- and right-eye retinal images. $\Omega_{rate_{t20}} = \Omega_{rate}$ Top 20%, $\Omega_{rate_{b80}} = \Omega_{rate}$ Bottom 20%. The results of $\Omega_{rate_{t20}}$ are with respect to $\Omega_{rate_{b80}}$ group. 195

Nomenclature

Acronyms / Abbreviations

1-D One-dimensional

2-D Two-dimensional

3-D Three-dimensional

ACD All Cause Death

ACR Albumin Creatinine Ratio

AD Alzheimer Disease

AMD Age-related Macular Degeneration

ANN Artificial Neural Network

ASCVD Atherosclerotic Cardiovascular Disease

AUC Area Under Receiver Operating Characteristic (ROC) Curve

AUPRC Area Under the Precision-Recall Curve

AVR Arterio-Venous Ratio

BES Beijing Eye Study

BMI Body Mass Index

BP	Blood Pressure
CAM	Class Activation Mapping
CCL	Cancer Cell Lines
CHI	Community Health Index
Chol	Cholesterol
CI	Confidence Interval
CKD	Chronic Kidney Disease
CLAHE	Contrast Limited Adaptive Histogram Equalization
CNN	Convolutional Neural Network
CNS	Central Nervous System
Conv	Convolutional
Coxph	Cox proportional hazard
CRAE	Central Retinal Arteriolar Equivalents
CR	Chronological Rate
CRVE	Central Retinal Venular Equivalents
CT	Computerized Tomography
CV	Cardiovascular
CVD	Cardiovascular Disease
CV death	Cardiovascular death
DARTS	Diabetes Audit and Research in Tayside Scotland

DBN	Deep Belief Network
DBP	Diastolic Blood Pressure
DL	Deep Learning
DL-FAS	Deep-Learning Funduscopy Atherosclerosis Score
DM	Diabetes Mellitus
DME	Diabetic Macular Edema
DNN	Deep Neural Network
DPN	Diabetic Peripheral Neuropathy
DR	Diabetic Retinopathy
DRS	Diabetic Retinopathy Screening
eGFR	estimated Glomerular Filtration Rate
EMR	Electronic Medical Record
EX	Hard Exudates
FC	Fully Connected
F	Female
FoV	Field of View
FPR	False Positive Rate
GAN	Generative Adversarial Network
GAP	Global Average Pooling
GH	Glycated Haemoglobin

GIMP GNU Image Manipulation Program

GoDARTS Genetics of Diabetes Audit and Research in Tayside Scotland

Grad-CAM Gradient-based Class Activation Mapping

GWPRS Genome-Wide Polygenic Risk Scores

HbA1c Haemoglobin A1c

HDL High-Density Lipoprotein

HE Hemorrhages

HIC Health Informatics Center

HPCSNUH Health Promotion Center of Seoul National University Hospital

HR Hazard Ratio

ICD International Classification of Diseases

IDRID Indian Diabetic Retinopathy Image Dataset

IGTD Image Generator from Tabular Data

ILSVRC ImageNet Large Scale Visual Recognition Challenge

INSPIRED INdian-Scotland PartnershIp for pREcision mEDicine in Diabetes

IQR Interquartile Range

KM Kaplan-Meier

KNN K-Nearest Neighbors

LDL low-density lipoprotein

MACE Major Adverse Cardiovascular Event

MAE Mean Absolute Error

MA Microaneurysms

MD Macular Degeneration

MDRF Madras Diabetes Research Foundation

MESSIDOR Methods to Evaluate Segmentation and Indexing Techniques in the field of
Retinal Ophthalmology

MI Myocardial Infraction

ML Machine Learning

M Male

MRI Magnetic Resonance Imaging

MSE Mean Squared Error

MS Multiple Sclerosis

NIHR National Institute of Health Research

NLP Natural Language Processing

NPDR Non-Proliferative Diabetic Retinopathy

NRS National Records of Scotland

NV Neo Vascularization

OC Optic Cup

OCT-A Optical Coherence Tomography Angiography

OCT Optical Coherence Tomography

ODD Optic Disc Diameters

OD Optic Disc

PAD Predicted Age Difference

PALFD Predicted Age Last First Difference

PCE Pooled Cohort Equations

PD Parkinson Disease

PDR Proliferative Diabetic Retinopathy

PRLFD Predicted Risk Last First Difference

PRS Predicted Risk Score

RDR Referable Diabetic Retinopathy

ReLU Rectified Linear Unit

RNN Recurrent Neural Network

ROC Receiver Operating Characteristic

SBP Systolic Blood Pressure

SBRIA Seoul National University Bundang Hospital Retina Image Archive

SEED Singapore Epidemiology of Eye Diseases

SE Soft Exudates

SH Safe Haven

SLO Scanning Laser Ophthalmoscopy

SNDREAMS Sankara Nethralaya Diabetic Retinopathy Epidemiology and Molecular Genetics Study

SP2 Singapore Prospective Study Program

SVM Support Vector Machine

T1D Type 1 Diabetes

T2D Type 2 Diabetes

T2I Tabular data to Image

TC Total Cholesterol

TML Tabular data Machine Learning

TNR True Negative Rate

TPR True Positive Rate

Trig Tryglicerides

UoD University of Dundee

VAMPIRE Vessel Assessment and Measurement Platform for Images of the REtina

VPT Vibration Perception Threshold

WHO World Health Organization

Chapter 1

Introduction

1.1 Background and motivation

This chapter covers the motivation for this work, the background for systemic diseases, retinal imaging, and deep learning. The subsequent summary of contributions is followed by the overview of the thesis organization.

1.1.1 Motivation for this work

According to a 2019 World Health Organization (WHO) report, Cardiovascular Disease (CVD) (ischaemic heart disease, stroke), diabetes and Chronic Kidney Disease (CKD) are among the top 10 causes of death worldwide [14]. The number of deaths due to ischaemic heart disease has increased enormously from 2 million in 2000 to 8.9 million in 2019 and there is a 70% increase in the number of deaths due to diabetes from 2000 to 2019 [14]. CVDs are the leading cause of death globally with an estimated 17.9 million deaths in 2019 which constitutes 32% of global mortality [15]. Deaths reported worldwide in 2019 due to diabetes were 1.5 million and CKD were 1.3 million [14]. Every year CVD causes over 3.9 million deaths in Europe and the health care cost associated with CVD in EU is almost €210 billion every year [16]. Reliable early diagnosis would contribute to containing, delaying, or and control the development of disease, reducing costs on healthcare systems,

and improving the patients' quality of life[17, 18]. Therefore there is a need for early diagnosis techniques that can be shown to be reliable, accurate, and ultimately suitable for use in clinical pathways. There is mounting evidence that retinal biomarkers can help in the early diagnosis of systemic diseases like CVD, diabetes, CKD and dementias, among others [19, 20, 9]. Retinal images are easy to acquire, cheap, and non-invasive compared to other instruments visualizing the body vasculature, e.g., Magnetic Resonance Imaging (MRI) or Computerized Tomography (CT). The retina can be observed directly through a natural hole, the pupil, with relatively inexpensive cameras (again in comparison to other instruments). For more details, the reader is referred to Section 2.4.

Deep Learning (DL) has shown promising results in various applications including disease diagnosis, disease prediction, image segmentations etc [21]. DL has emerged as a powerful tool for retinal fundus image analysis [8]. It can automatically learn and extract complex features from large datasets. Retinal biomarkers including changes in the thickness and volume of retinal layers, variations in the optic nerve head and retinal blood vessels, and the presence of lesions or abnormalities in the retina can be indicative of various disease conditions such as Age-related Macular Degeneration (AMD), Diabetic Retinopathy (DR) and glaucoma [19, 22]. Some retinal biomarkers have been found to be associated with an increased risk of developing systemic conditions like CVD [23]. Current challenges include the generalizability of DL applications of retinal biomarkers over diverse cohorts (demographics, age, health conditions, lifestyle, existing morbidities, ethnicity, social background and others). This thesis investigates the role of retinal images as a source of biomarkers for identifying associations with CVD, Cardiovascular (CV) risk factors, CV risk scores, Cardiovascular death (CV death), All Cause Death (ACD), microvasculature diseases (CKD, Diabetic Peripheral Neuropathy (DPN), DR) in a diabetic cohort, Genetics of Diabetes Audit and Research in Tayside Scotland (GoDARTS) (Section 3.3), using DL. Motivated by the above, this research was mainly conducted in cohorts of prevalently elderly-aged individuals with a mean age of 66 years.

The work reported in this thesis was part of the INdian-Scotland PartnershIp for pRecision mEdicine in Diabetes (INSPIRED)-National Institute of Health Research (NIHR) program

[24], a partnership between the University of Dundee (UoD), UK and the Madras Diabetes Research Foundation (MDRF), India. The participating academic groups in Dundee were the Discipline of Computing, School of Science and Engineering, and the Division of Clinical and Molecular Medicine, School of Medicine. The DL research work was carried out in Computing, UoD, using the computing resources provided by the Health Informatics Center (HIC) [25], including the ISO-certified safe haven providing an indispensable interface between clinical data generation and data use by research groups.

1.2 Systemic diseases

A systemic condition or systemic disease is defined as a disorder that affects a number of organs and tissues or the whole body [26, 27]. There are several systemic diseases; for the purpose of this thesis, this section mainly discusses those related to CVD, diabetes, CKD.

1.2.1 Cardiovascular diseases

CVD is a broad term referring to a group of conditions affecting the heart or blood vessels. It is mainly associated with fat deposits inside arteries (atherosclerosis) and an increased risk of blood clots [28]. Damage to arteries in organs such as the brain, heart, kidneys, and eyes are typically associated with CVD. There are four main types of CVD.

1. Coronary heart disease occurs when the blood flow to the heart is blocked or reduced. This can lead to angina, heart attacks, and heart failure.
2. Stroke occurs when the blood supply to some parts of the brain is stopped.
3. Peripheral arterial disease occurs when the blood supply in the arteries of limbs (usually legs) is blocked.
4. Aortic diseases are conditions that affect the aorta. The most common one is the aortic aneurysm, where the wall aorta swells locally outwards.

The exact reason for CVD is not completely clear; several risk factors can increase the chances of getting CVD. These include age, sex, high blood pressure, high cholesterol, diabetes, smoking, alcohol, inactivity, excess weight, ethnicity, and family history of CVD. Guidelines for assessing CVD risk exist [29, 30]; for instance, the American College of Cardiology recommends calculating 10-years risk (using several risk factors) in patients ages between 20 and 79 years [30].

1.2.2 Diabetes

Diabetes (diabetes mellitus) is a long-lasting disease that occurs when the sugar (glucose) level in the blood remains elevated [31, 32]. Diabetes is caused by insufficient insulin secretion, insulin action, or both [33]. Insulin is a hormone produced by the pancreas to control the amount of sugar in the blood. Among the symptoms of diabetes are increased thirst, frequent urination, tiredness, increased hunger, weight loss, slow healing of wounds, and blur in vision. There are two main types of diabetes.

1. Type 1 Diabetes (T1D) occurs when the body cells that produce insulin are attacked and destroyed by the body's immune system.
2. Type 2 Diabetes (T2D) occurs when the body cells stop reacting to insulin or when the body cells do not produce enough insulin.

There are no specific preventive measures for T1D; on the other hand, T2D can be prevented or delayed by a healthy diet and physical activity, among others [34]. The several complications of diabetes include CVD, CKD, foot ulcers, nerve damage, diabetic retinopathy, and cognitive impairment [35].

1.2.3 Chronic kidney disease

CKD is a condition in which the functioning of the kidneys is compromised [36]. Kidney failure occurs often due to diabetes and high blood pressure [37] and also high cholesterol, kidney infections, kidney inflammations, and polycystic kidney disease [36]. Usually, there

are no symptoms for CKD in the early stages; at an advanced stage, symptoms can include tiredness, swollen feet, ankles, or hands, shortness of breath, and blood in the urine. CKD is usually diagnosed with blood and urine tests. The blood test measures the amount of waste filtered by kidneys in a minute, or estimated Glomerular Filtration Rate (eGFR). Healthy kidneys filter more than 90ml/min. The urine test quantifies the levels of albumin and creatinine to compute Albumin Creatinine Ratio (ACR). The eGFR and ACR values are used to determine the level of kidney damage and help clinicians decide the appropriate treatment.

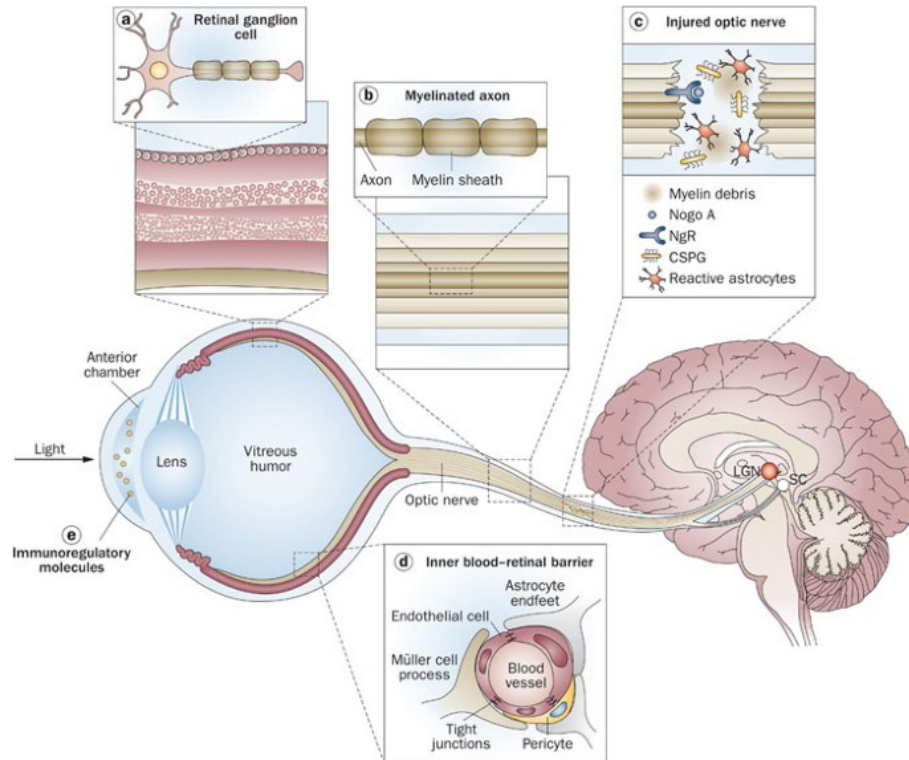
1.2.4 Eye examination

Regular examination of the eyes can play an important role in the identification of systemic diseases [38]. The American Academy of Ophthalmology recommends that all adults older than 40 years should have their eyes examined regularly, and even earlier and frequently in the case of the insurgence of risk factors associated with systemic conditions like diabetes, high blood pressure, or family history [38]. The eye examination includes checks of medical history, visual acuity, pupil, side vision, eye movement/pressure, the front part of the eye, retina, and optic nerve. Many parts of the eye such as its outer surface (eyelids, conjunctiva, and cornea) and the retina (back of the eye, or fundus), might provide important clues for the diagnosis of systemic conditions [39].

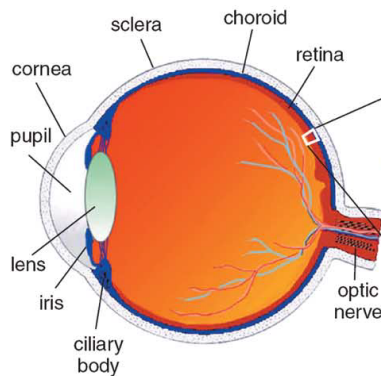
1.3 Retinal imaging

The retina is a layered tissue lining the interior of the back of the eye. It enables the conversion of incoming light into a neural signal that is further processed in pathways reaching the visual cortex of the brain [19]. During embryonic development, the retina and optic nerve extend from the diencephalon and are thus considered part of the Central Nervous System (CNS) [40–42, 20]. Studies of the optic nerve have revealed much about the axonal response, including at the time of CNS injury [20, 43, 44]. The retina displays similarities to the brain and spinal cord in terms of anatomy, functionality, response to injury, and immunology

[20, 45, 46]. Figure 1.1 illustrates the connection of the eye to the brain and the main parts of the eye relevant to the work of this thesis.



(a) Illustration of eye as an extension of CNS [20].



(b) cross-sectional view of eye and its anatomy [19].

Fig. 1.1 (a) Connection of the eye to the brain. (b) The main parts of the eye for the purpose of this thesis.

Researchers often refer to the retina as a window to the brain and vascular system [20, 47]. As the eye is an extension of the brain, various ocular and specifically retinal changes have been detected and characterized in patients with CNS disorders [19] such

as CVD complications (stroke, myocardial infarction), Multiple Sclerosis (MS), Parkinson Disease (PD), Alzheimer Disease (AD), CKD [48, 49]; neurodegenerative disorders [20] such as age-related macular degeneration (AMD), glaucoma; and diabetic retinopathy (DR). The retina is imaged with noninvasive and comparatively inexpensive procedures and instruments compared to other methods of imaging the human vascular system.

Three major imaging technologies for capturing the retina are fundus imaging, Scanning Laser Ophthalmoscopy (SLO) and Optical Coherence Tomography (OCT) [9]. Fundus imaging is easy to capture, cost-effective, suitable for mass screening programs, and routinely used for retina examinations. The fundus of the eye is the internal back surface of the eye bulb (posterior pole), opposite the lens. The main anatomical landmarks are the optic disc, macula, fovea, and blood vessel network [50]. In fundus imaging, a 2-D image of the 3-D retinal semi-transparent tissues is obtained using reflected light [19]. The retina can be photographed directly through the pupil using a fundus camera; an example (Topcon fundus camera) is shown in Figure 1.2. Patients sit at the device with their chin on a chin rest and their forehead against a bar. An ophthalmic photographer focuses and aligns the camera and presses the shutter release to trigger a flash of light and register a fundus photograph. Typical fields of view are 30°, 45°(most common), or 60°with a magnification of 2.5 times [51]. Figure 1.3 shows a sample monochrome retinal image with different fields of view.

Fundus imaging is widely used for large-scale screening of diabetic retinopathy, and also glaucoma, and age-related macular degeneration [19]. Retinal imaging instruments, once only manual ophthalmoscopes and low-resolution fundus cameras, have been evolving and include nowadays cost-effective fundus imaging [52, 53], functional imaging [54–56], adaptive optics [57], OCT [58], Optical Coherence Tomography Angiography (OCT-A) [59], autofluorescence [60], ultra-widefield angiography [61, 62], hyperspectral retinal imaging [63, 64]. Accordingly, progress has been accomplished in the automatic analysis of fundus images, although challenges remain, for instance (depending on imaging modality, level of accuracy, and others) for lesion detection, abnormality detection and assessment, and, data analysis, associating clinical outcomes with fundus images either via statistical techniques or machine learning [19].



Fig. 1.2 A Topcon TRC 50 EX fundus camera [1].

In recent years, DL has emerged as a powerful tool for computer vision tasks, including image classification, object detection, and segmentation. DL algorithms are increasingly being used in fundus image analysis due to their ability to learn relevant features automatically from the images and make accurate predictions. In particular, DL has shown great promise in the analysis of DR, a leading cause of adult blindness. By analyzing fundus images of the eye, DL models can detect and grade the severity of DR with high accuracy, enabling early intervention and treatment. Additionally, DL has been used to predict various cardiovascular risk factors, such as age, sex, and blood pressure, from fundus images. More details are provided in Chapter 2.

1.4 Deep Learning

Artificial Neural Network (ANN)s are statistical models directly inspired by biological neural networks [65]. The basic computational unit of the brain is the neuron. Figure 1.4 shows

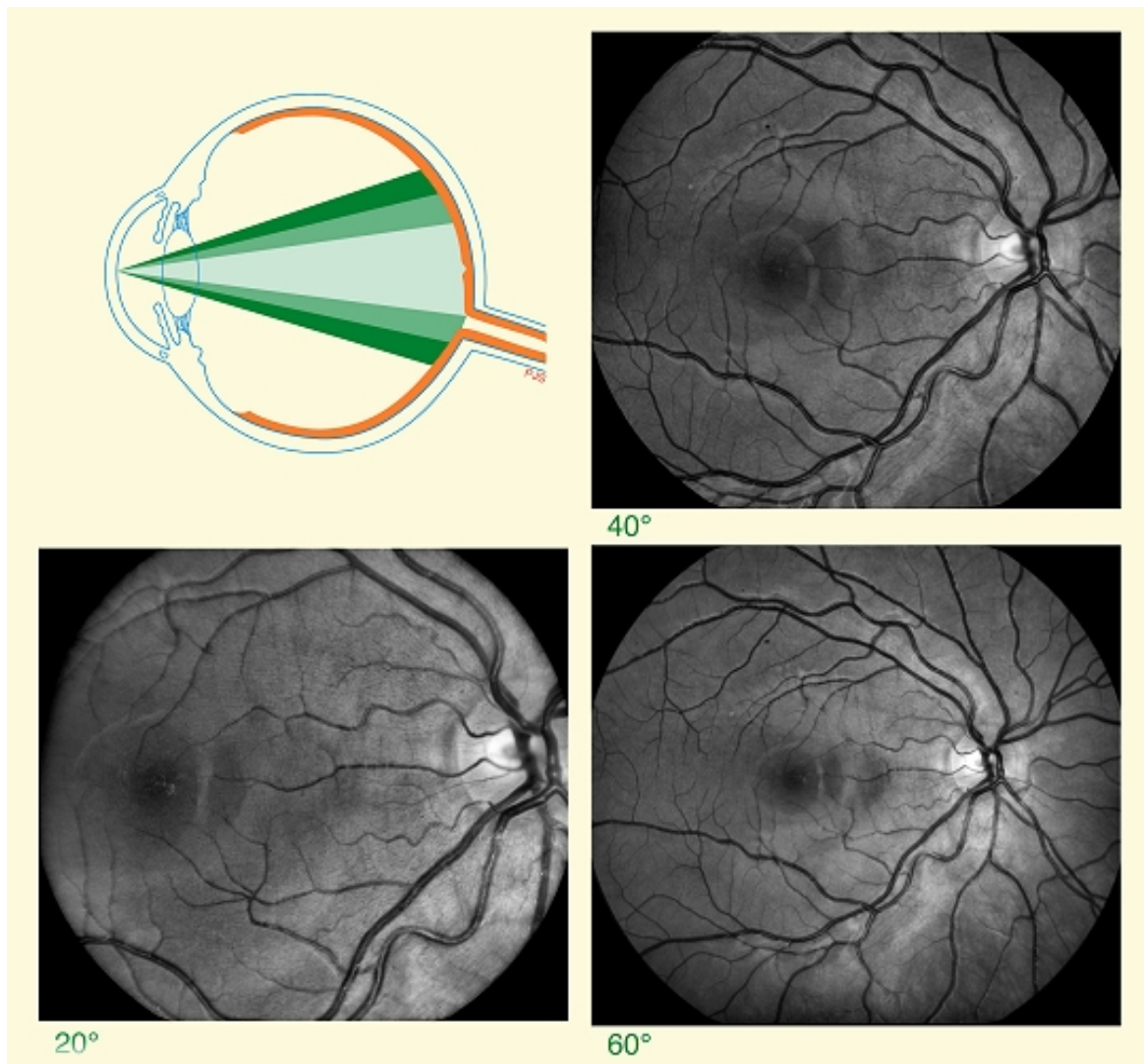


Fig. 1.3 A sample monochrome retinal image illustrating the different fields of views [2].

a cartoon drawing of a biological neuron and the common mathematical model, a linear combination of weighted inputs followed by a nonlinearity [3]. This is the neuron model adopted in conventional neural networks, i.e., before the appearance of deep learning ones [66]. ANNs are modeled as acyclic graphs of neurons. In *fully connected layered neural network*, neurons between two adjacent layers are fully connected pairwise. In ANNs, the input layer is not counted as a layer; it simply carries values for the input variables (nodes). A sample 3-layered neural network is shown in Figure 1.5.

A Deep Neural Network (DNN) is an ANN with multiple layers between the input and output layers [67, 68], like conventional ones. Importantly, the layers of a DNN can be made of different processing components, whereas all nodes in a ANN perform the same operation (described above).

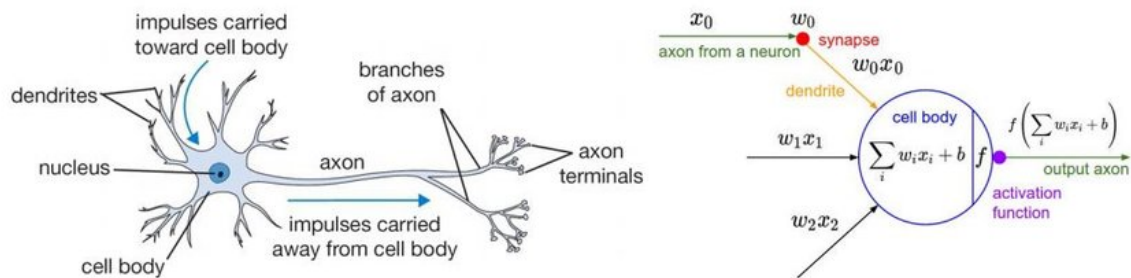


Fig. 1.4 A cartoon drawing comparing biological neuron (left) and a common mathematical neuron model (right) [3].

DL methods are representation-learning methods based on DNN. Hence presentations, encoded in the weights of the convolutional layers of the network, are learned at multiple levels. They transform the representation at one level (starting with the raw input) into a representation at a higher, slightly more abstract level, progressively expanding the data segment considered (for instance, an image patch). Given enough layers, very complex, nonlinear functions can be learned [66].

ANN and DNN solve optimization problems, finding the value of the weights that minimizes an objective function, or *loss function* during training. This is achieved by the *back-propagation algorithm* [69–73]. In backpropagation, the gradients of the objective function with respect to the network weights are computed and the weights are adjusted iteratively to minimize an error measure between the actual and the predicted output vector [74]. This assumes that the correct output vector values are known, i.e., the network is trained within the *supervised learning* paradigm.

Some common DL architectures are Deep Belief Network (DBN) [75], autoencoders [76], Generative Adversarial Network (GAN) [77], Convolutional Neural Network (CNN) [78], Recurrent Neural Network (RNN) [79].

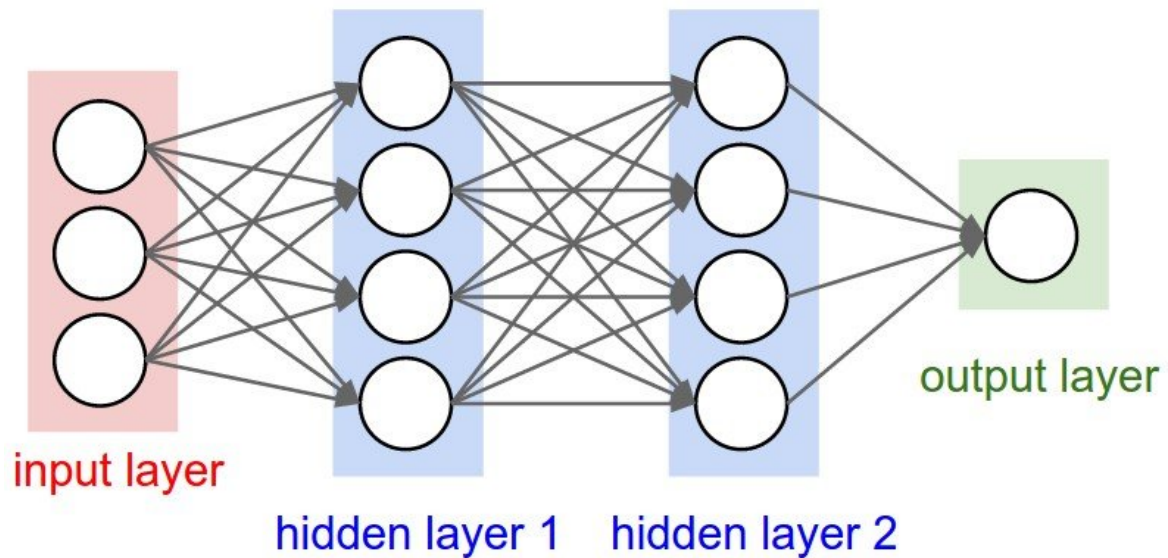


Fig. 1.5 A sample 3-layered ANN with 3 inputs, 2 hidden layers, and 1 output layer [3].

1.4.1 Basic components of a convolutional neural network

The Convolution Neural Network (CNN) is a specific type of deep learning architecture [80, 81] inspired by the visual cortex of mammals [82, 83]. A typical architecture is shown in Figure 1.6 and consists of a series of *convolutional layers*, *pooling layers* (also known as a subsampling layer), *nonlinear layers*, typically followed by a fully connected neural network and an output layer [81].

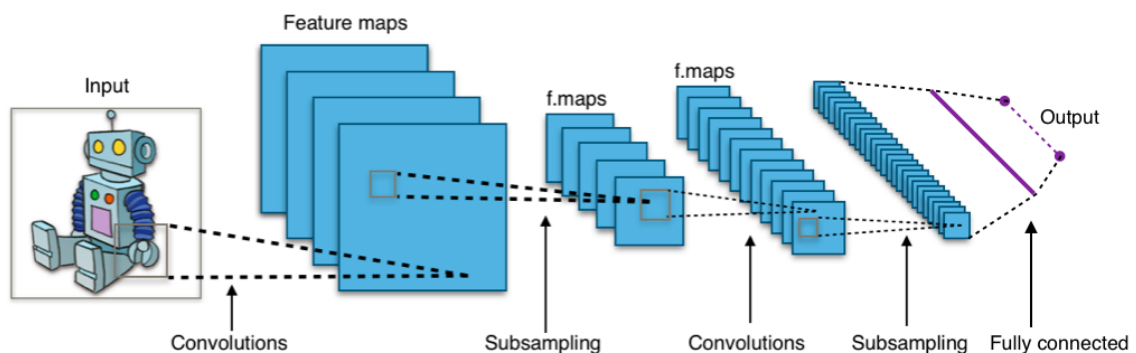


Fig. 1.6 A typical CNN architecture [4].

Convolutional layers contain *feature maps*. A feature map is the output of a single filter applied to the previous layer. Layers can be regarded as a *battery of filters*. Each unit of a feature map is connected to local patches in the previous layer's feature maps through a set

of weights called a filter bank. In an image analysis network, the first layer receives the pixel intensities as input; hence each filter acts on an image patch (*receptive field*). All units in a feature map share the same filter bank. Different feature maps in a layer use different filter banks [66]. A mathematical formula for the convolution operation is provided in Equation (1.1), where S is the output feature map, I is the input image, K is filter and m, n are the pixel indices of S . An example of convolution operation is shown in Figure 1.7.

The result of the convolution, basically a locally weighted sum, or dot product, of the filter bank and local patch is then passed through a *non-linearity activation unit* such as a Rectified Linear Unit (ReLU). The ReLU is sometimes followed by a subtractive and divisive local normalization [6]. Commonly used activation functions, along with their mathematical definition, are listed in Table 1.1, and their plots are shown in Figure 1.8.

$$S(m,n) = (I * K)(m,n) = \sum_i \sum_j I(i,j)K(m-i,n-j) \tag{1.1}$$

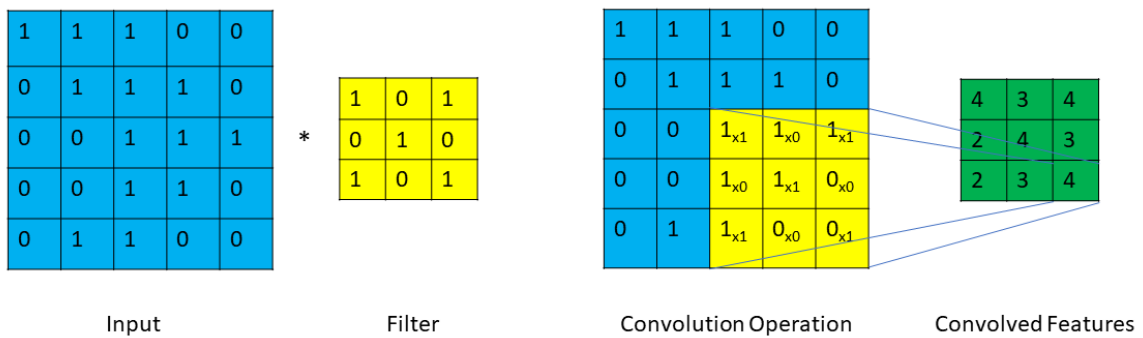


Fig. 1.7 An input image of 5x5 dimensions convolved with a 3x3 filter.

Table 1.1 Different types of activation functions.

Activation function	Mathematical formula
Sigmoid	$\sigma(x) = \frac{1}{1+e^{-x}}$
Tanh	$\tanh(x) = 2\sigma(2x) - 1$
ReLU	$f(x) = \max(0,x)$

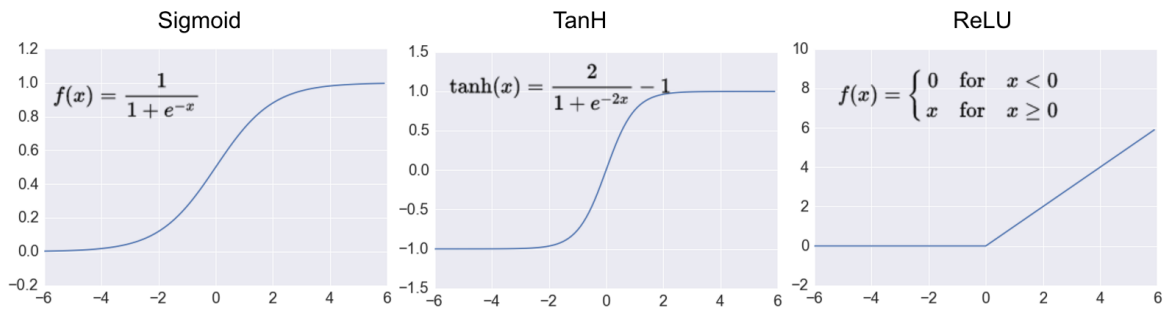


Fig. 1.8 Plots of common activation functions [5].

A typical pooling unit computes the maximum of a local patch of units in one feature map. This layer treats each feature map separately. The average pooling layer computed the average values over a neighborhood of pixels in each feature map. The output of this layer results in a reduced size of the feature map. If the average operation is replaced by a maximum then this is called maximum pooling [6]. Figure 1.9 shows an example of feature extraction in a CNN.

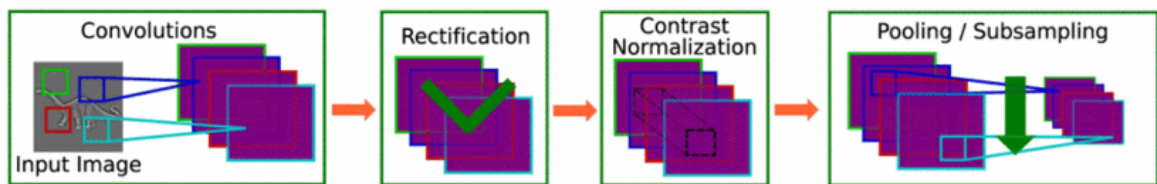


Fig. 1.9 An example of feature extraction. An input image (or a feature map) is passed through a filter bank, followed by non-linearity and spatial pooling/sub-sampling [6].

1.4.2 Early deep learning architectures

LeNet, which was first introduced by LeCun et al. in 1990 [80] and later improved [81], was a pioneering work in the field of CNNs. It was specifically designed for classifying handwritten digits and was successful in recognizing visual patterns directly from the input image without any pre-processing. However, due to a lack of sufficient training data and computing power at the time, this architecture failed to perform well in complex problems [84].

The open computer vision challenges have created increasingly large image databases, allowing users across the world to develop or improve deep learning image recognition architectures and models. The ImageNet Large Scale Visual Recognition Challenge (ILSVRC) is a challenge that started in 2010 [85]. It has become the standard benchmark for large-scale object recognition and presents various challenges. The competition is accompanied by a workshop where the progress and lessons learned from the most successful and innovative entries each year are discussed. The publicly available dataset used for the annual competition allows for the development and comparison of categorical object recognition algorithms. The ILSVRC 2010 challenge consisted of a subset of the ImageNet database containing 1,461,406 images with 1,000 object classes. The ImageNet database itself contains 15 million labeled high-resolution images belonging to roughly 22,000 categories [86].

In ILSVRC [85], CNN architectures like AlexNet (2012) [86], ZFNet (2013) [87, 88], GoogLeNet/Inception (2014) [89, 90], VGGNet (2014) [91, 92], ResNet (2015) [93], DenseNet (2016) [94], ResNext (2016) [95], SENet (2017) [96], PNASNet (2018) [97], EfficientNet (2019) [98] have proven the top performers. Among these, AlexNet was a breakthrough CNN that almost halved the error rate for object recognition at the time and precipitated the rapid adoption of deep learning by the computer vision community [66].

Several of the above DL architectures were tried in the initial experiments, described in Chapter 4. The EfficientNet-B2 was eventually selected for further experiments with real clinical data (GoDARTS), described in Chapters 5, 6, 7.

1.4.3 Visualizing convolutional neural networks

Zhou et al. [7] proposed the Class Activation Mapping (CAM) technique to enable a CNN to both classify an image *and* localize class-specific image regions. Figure 1.10 shows sample images that highlight the image regions found by the network to be the ones contributing the most to the classification task, e.g., the head of the animal for briard (dog breed) and hen, the plates in barbell (weightlifting equipment), and the bell in bell-cote (a small roofed structure for bells).

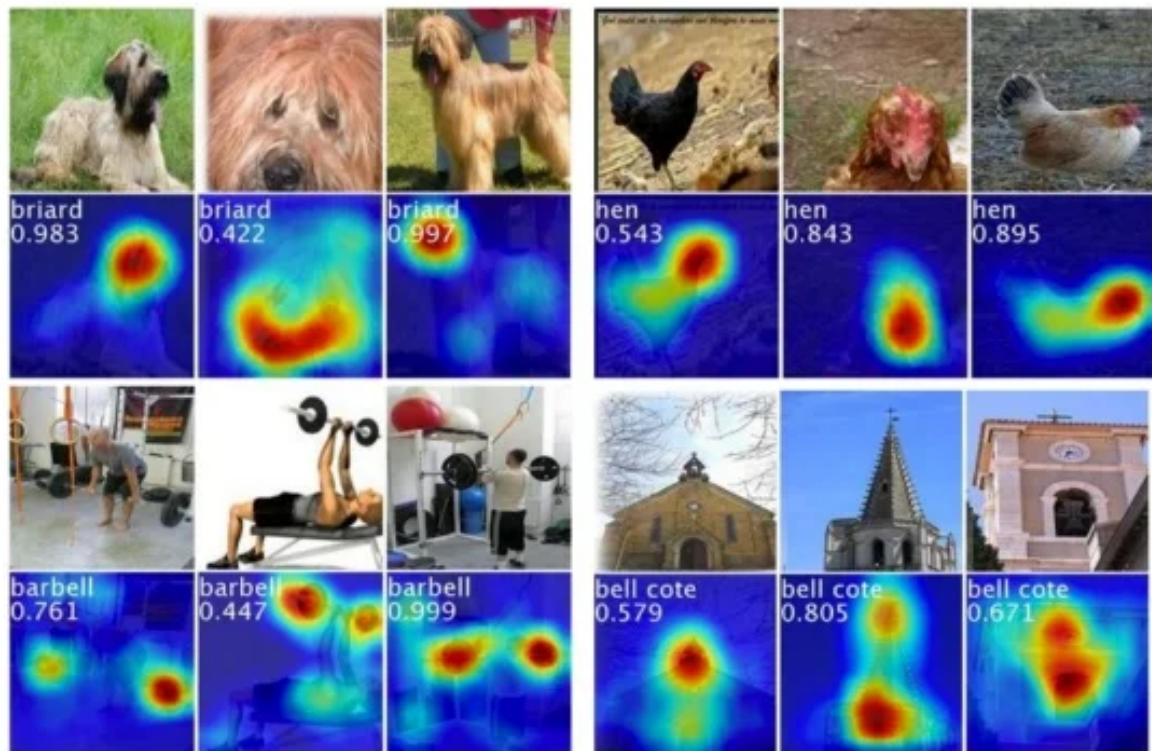


Fig. 1.10 The CAMs of four classes from ILSVRC [7].

Selvaraju et al. [99] have proposed Gradient-based Class Activation Mapping (Grad-CAM) which uses the gradients of any target class, flowing into the final convolutional layer to produce a localization map highlighting the important regions in the image for predicting the class. This technique is a generalization of CAM and overcomes the limitations of CAM. The Grad-CAM algorithm was implemented for visualizing the CNN results in the experiments, described in Chapter 5 and Chapter 6.

1.5 List of aims

This thesis set out to address several aims, provided below, which will be discussed in the following chapter. Subsequently, Chapter 8 will summarize the contributions made towards achieving these aims. The motivation for setting these aims is explained in Section 2.4, from which we drew our inspiration.

1. A framework for generating systematically synthetic datasets *parameterized by difficulty level* to test classifiers for medical image analysis.
2. An analysis of DL for predicting CV risk factors, namely age, sex, Systolic Blood Pressure (SBP), Diastolic Blood Pressure (DBP), High-Density Lipoprotein (HDL), Total Cholesterol (TC), Glycated Haemoglobin (GH), Body Mass Index (BMI) Tryglicerides (Trig); Major Adverse Cardiovascular Event (MACE), ACD and microvasculature complications (CKD, DPN, DR) using only retinal images in a diabetic cohort.
3. An investigation of the difference between age predicted from a retinal image using DL and chronological age as a retinal biomarker for MACE and ACD.
4. An investigation of the complementary role of retinal images, a commonly used clinical score, and a Genome-Wide Polygenic Risk Scores (GWPRS) as predictors of CVD risk.
5. An investigation of the DL predicted clinical CV risk score from retinal images to find associations with MACE and CV death.
6. A method to convert spreadsheet data to images and perform multi-modal image analysis, concatenating retinal images data and spreadsheet to image converted data, for the risk stratification of MACE using DL.

1.6 Structure of the thesis

- **Chapter 2 (Related work):** Reviews and compares the traditional Machine Learning (ML) algorithms used in retinal image analysis from the retinal features extracted using retinal vasculature measuring software with DL algorithms applied on retinal images. It also reviews the recent literature on retinal image biomarkers using DL algorithms for predicting systemic diseases, and their associated risk factors.
- **Chapter 3 (Material):** Describes the datasets, Methods to Evaluate Segmentation and Indexing Techniques in the field of Retinal Ophthalmology (MESSIDOR) and

Indian Diabetic Retinopathy Image Dataset (IDRID), used in the initial experiments and GoDARTS, used for the main analysis with real retinal images in this thesis. This chapter also includes the pre-processing methods applied to GoDARTS retinal images.

- **Chapter 4 (Identifying robust CNN architecture):** Proposes a framework for generating synthetic datasets parameterized by difficulty level to test classifiers in medical image analysis. It also compares the performance of five different DL models on the synthetic dataset and validated these models on an independent dataset to choose the best performing DL model.
- **Chapter 5 (Predicting demographic and clinical features):** Investigates the retinal images for predicting demographic features, CV risk factors and systemic disease outcomes in GoDARTS using EfficientNet-B2; and also experiments on association of difference between the DL predicted age and chronological age with 10 years MACE and ACD.
- **Chapter 6 (Predicting cardiovascular risk scores):** Contains the DL investigation results on predicting the clinical CV risk score from retinal images and investigated the complementarity of retinal and genetic information for CVD risk. It presents the association of DL predicted risk score with 10 years MACE and CV death.
- **Chapter 7 (Converting tabular data to images for deep learning):** This chapter introduces to a novel approach for converting tabular data into grayscale image data. It discusses the performance results for 5 years MACE stratification from the retinal images, Tabular data to Image (T2I) data as well as a combination of retinal images and T2I data using DL.
- **Chapter 8 (Conclusions and future work):** The final chapter summarises the main conclusions of this work and shares thoughts and suggestions for future directions to be explored based on the experiments and results presented.

Chapter 2

Related work

2.1 About this chapter

This Chapter introduces several software tools available to measure retinal vasculature and the associations of these vascular measurements to vascular risk factors and systemic diseases. A comparison of performance for disease classification using the traditional Machine Learning (ML) and Deep Learning (DL) algorithms was presented. The recent literature on applying DL methods to retinal images for predicting Cardiovascular (CV) risk factors and systemic conditions were discussed in detail.

2.2 Retinal vasculature

The retina is a highly vascularized tissue. Properties and changes in the vascular morphology have been associated with several disease outcomes [19, 20]. Figure 2.1 represents a sample fundus image showing some of the relevant anatomical landmarks, biomarkers, and lesions. Retinal vasculature parameters like Central Retinal Arteriolar Equivalent (CRAE), Central Retinal Venular Equivalent (CRVE) and their ratio, the Arterio-Venous Ratio (AVR), vessel tortuosity, vessel width, bifurcation angles, branching angles, and vessel caliber have been used for early stratification of Diabetic Retinopathy (DR), Macular Degeneration (MD), hypertension, stroke, neovascular glaucoma, and other cardiovascular diseases [100–108].

Figure 2.2 illustrates some of the most important retinal vascular parameters in current biomarker studies in a fundus image. A more recent candidate biomarker is the fractal dimension of the retinal vasculature regarded as a digital pattern [109–112].

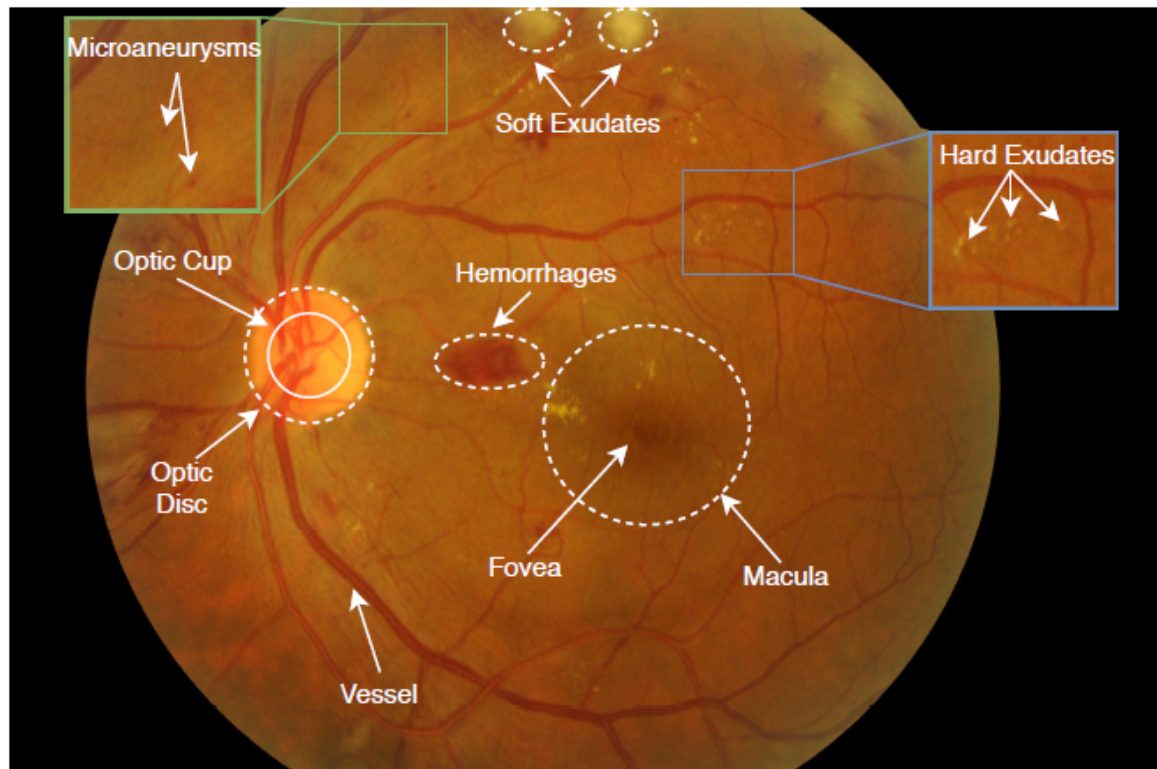


Fig. 2.1 A sample fundus image illustrating important anatomical landmarks and lesions [8].

Several software tools have been developed to quantitate the retinal vasculature effectively, such as IVAN [113], SIVA [114], VAMPIRE [115] and QUARTZ [116]. Traditionally width-related parameters have been calculated from the major arteries and veins in the Zone B region of the retina, an annulus surrounding the Optic Disc (OD) between 1.0 and 1.5 Optic Disc Diameters (ODD) from OD center, much used in retinal biomarker studies [117, 118, 113]. Figure 2.3 shows a sample retinal image marked with different zone regions used by Vessel Assessment and Measurement Platform for Images of the RETina (VAMPIRE). It is challenging to accurately segment the blood vessels and classify artery-vein in the retinal images, a fundamental step before the retinal vasculature parameters can be measured [119]. Several studies extend measurements like tortuosity, bifurcation angles, and fractal dimension to Zone C, the annulus between 1.0 and 2.5 ODD from OD center [120–124].

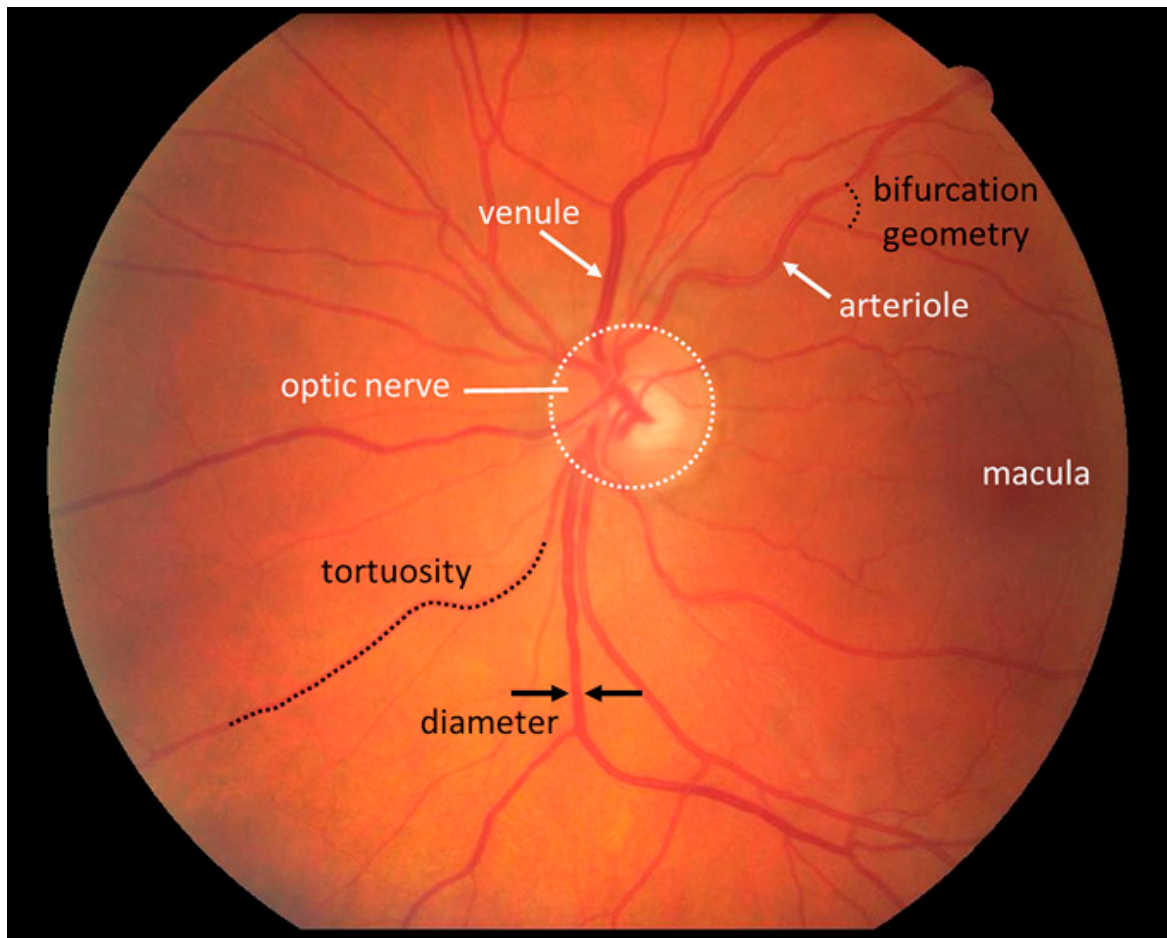


Fig. 2.2 Illustration of retinal vasculature parameters in fundus image [9].

Several studies have investigated the relationship between retinal vessel caliber and vascular risk factors, Cardiovascular Disease (CVD)s (like stroke, heart failure, coronary artery disease), mortality [125, 126, 104, 127]. Findings suggest that retinal vessel caliber is a biomarker for risk stratification of CVD and Chronic Kidney Disease (CKD) [128, 129]. Seidelmann et al. [104] found that long-term mortality in both males and females is associated with the narrowing of the retinal arterioles and widening of venules. Sun et al. [125] in their analysis of an Asian population, they found the smaller retinal arterioles associated with hypertension and larger retinal venules with smoking, dyslipidemia, hyperglycemia and higher Body Mass Index (BMI). These associations were replicated in a white population from USA, Canada, and Australia.



Fig. 2.3 Retinal image zones used by VAMPIRE [10].

Along with the CVDs, the retinal vasculature has been associated with brain and ocular diseases like glaucoma [130], Age-related Macular Degeneration (AMD) [131], DR [132, 133], Alzheimer’s disease [134, 10]. For glaucoma, several studies have associated glaucoma with the narrowing of retinal arterioles and decreased fractal dimension. Widening of retinal venular caliber is associated with the increasing severity of DR [133]. In small vessel disease, reduced fractal dimension has been associated with aging [10].

2.3 ML and DL in retinal image analysis

ML is a wide class of algorithms that learn to perform a specific task without explicit programming, either from examples (supervised learning) or without supervision (unsupervised learning). DL is a subset of ML that has introduced novel neural network architectures compared to the classic ones [66]. For building an analytical model [11] (see Figure 2.4), with the explicit program, feature extraction and model building are performed manually; ML

relies on features that are well defined and extracted manually; DL networks have the capability of automated feature extraction with minimal human effort due to their well-designed architectures.

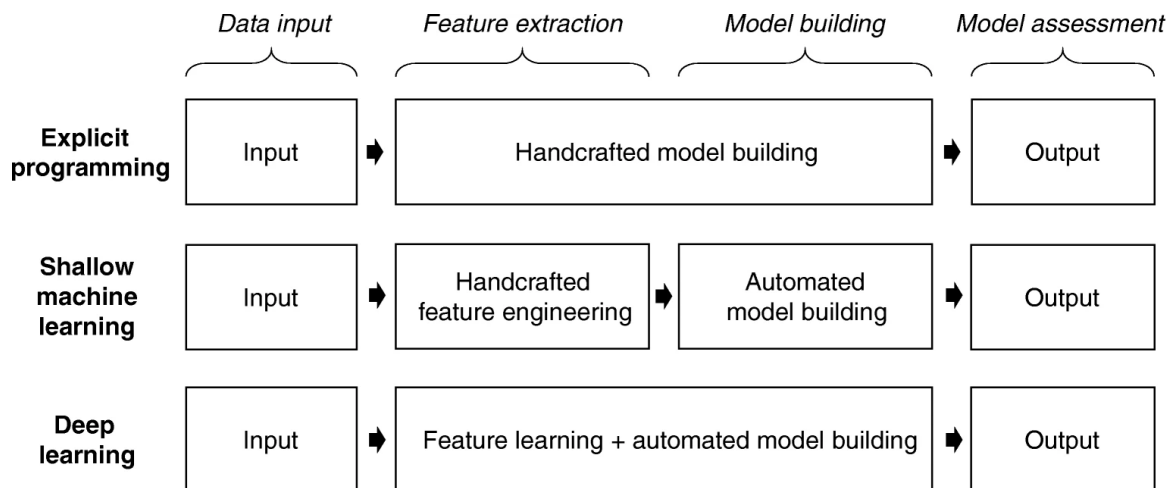


Fig. 2.4 Process involved in building of analytical model [11].

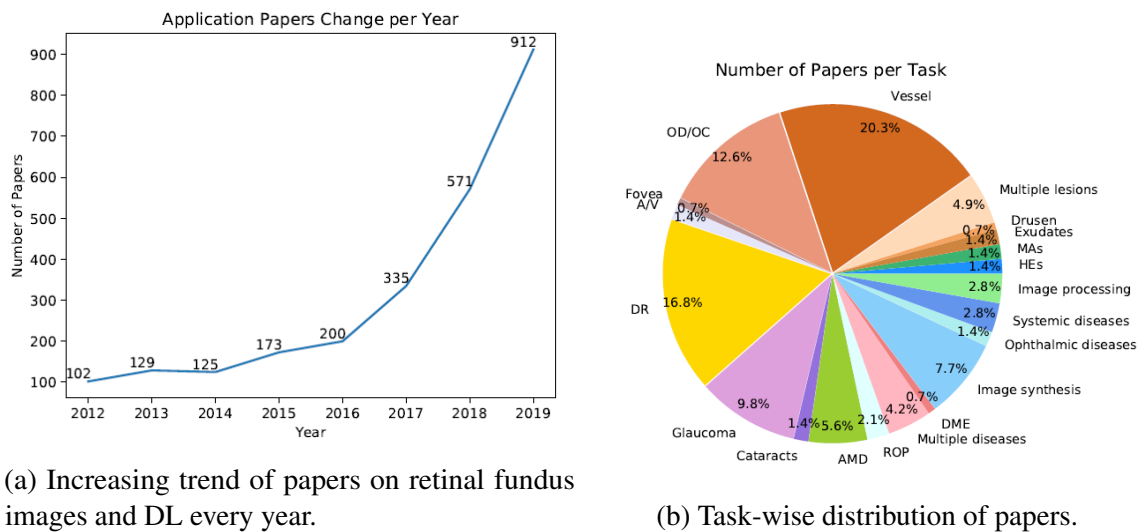
In recent years, traditional ML algorithms like linear regression, logistic regression, decision trees, and support vector machines have been outperformed by DL in many applications. For instance, for computer-aided diagnosis [135]:

1. *AMD classification* On AREDS dataset [136], deep feature with Support Vector Machine (SVM) [137] (DL algorithm) achieved an accuracy of 95% and SURF feature with random forest [138] (ML algorithm) obtained 91.8%.
2. *Glaucoma classification* On RIM-ONE dataset [139], Deep Convolutional Neural Network (CNN) [140] (DL algorithm) obtained an accuracy of 89.4% and K-Nearest Neighbors (KNN) [141] (ML algorithm) got 89%.
3. *DR classification* On MESSIDOR dataset [142], CNN Deep Residual Learning [143] (DL algorithm) obtained Sensitivity of 94% and Specificity of 98% and KNN [144] (ML algorithm) achieved 98.88% and 48.72% respectively.
4. *Segmentation* For OD segmentation, retinal vessel segmentation and lesion detection DL models outperformed ML models [145–150].

2.4 DL and retinal images biomarkers

Much studied biomarkers from fundus images are measurements related to the OD, macula, fovea, Optic Cup (OC), vasculature, and DR related lesions such as Microaneurysms (MA), Hemorrhages (HE), Soft Exudates (SE), Hard Exudates (EX). Identifying and measuring these biomarkers can be used in a variety of tasks.

The top performing DL architectures on the ImageNet Large Scale Visual Recognition Challenge (ILSVRC) dataset, described in Section 1.4.2, are generally used as backbone models for various tasks like classification and segmentation problems. Figure 2.5 shows the increasing trend of deep learning research papers on fundus images in recent years and the distribution of papers in respective tasks [8]. There are several studies of DL in fundus image analysis for various tasks as shown in Figure 2.5b but this section focuses on literature related to systemic conditions from which the thesis work was inspired.



(a) Increasing trend of papers on retinal fundus images and DL every year.

(b) Task-wise distribution of papers.

Fig. 2.5 Recent trend and distribution of papers [8].

Predicting CV risk factors

Poplin et al. [23] was the first to apply DL methods to quantify the CV risk factors from the fundus images. For developing the DL model, they used fundus images from 48,101 patients from UK Biobank [151] and 236,234 patients from EyePACS [152] and tested these

models using fundus images from 12,026 patients and 999 patients from UK Biobank and EyePACS respectively. The mean age in the test dataset was 56.9 ± 8.2 years (UK Biobank) and 54.9 ± 10.9 years (EyePACS). More details on the dataset characteristics are available in [23]. An ensemble model of ten Inception-v3 [90] architecture was used for training and testing on the data with an input image size of 587×587 pixels. The test results reported were the average of the results from these ten ensemble models trained on the same data. The results reported for the UK Biobank test dataset for CV risk factors included estimates of age (Mean Absolute Error (MAE) = 3.26 years), sex (Area Under Receiver Operating Characteristic (ROC) Curve (AUC) = 0.97), smoking status (AUC = 0.71), Systolic Blood Pressure (SBP) (MAE = 11.23 mmHg). For HbA1c, Diastolic Blood Pressure (DBP) and BMI the DL model performance did not improve much compared to the baseline method. The authors applied a DL technique, soft attention [153, 154], to visualize the saliency map in the input fundus image that is critical for the predictions made. They reported that OD and blood vessels are the important regions in the input fundus images to generate the predictions.

Kim et al. [155] trained DL models for age and sex prediction using 219,302 retinal fundus images from normal participants (without hypertension, DM, and smoking history) and these trained models were tested in four test sets using fundus images from normal participants (24,366 images) and participants with hypertension (40,659 images), Diabetes Mellitus (DM) (14,189 images), and any smoking history (113,510 images). In total, they used 412,026 fundus images from 155,449 participants from Seoul National University Bundang Hospital Retina Image Archive (SBRIA) [156, 157] for the DL model training and testing. The mean age was 46.70 ± 16.67 years in the normal training set, and 46.64 ± 15.83 in the normal test set, and it was higher by 10 years in other test sets with underlying vascular conditions. ResNet-152 [93] DL architecture was used for their work. For age prediction in the test data, the model performance in normal participants was MAE = 3.06 years and $R^2 = 0.92$, MAE and R^2 in another test dataset with hypertension (3.46 years, 0.74), DM (3.55 years, 0.75), and smoking (2.65 years, 0.86). For sex prediction, the AUC is greater than 0.96 in all the test datasets. They also reported that for participants with age >60 years, the MAE is above 4 years and accuracies declined in all the test datasets. Their Class Activation

Mapping (CAM) heatmaps indicate that fovea, OD, and retinal vessels are highly activated during the DL model predictions.

Rim et al. [158] developed DL models to predict 47 systemic biomarkers using fundus images with seven diverse Asian and European cohorts. Their outcome variables include demographic features, Blood Pressure (BP), body composition measurements (muscle mass, height, and bodyweight), several pathological measurements, biomarkers related to liver, thyroid, and kidney function, inflammation, and diabetes. To train and test the DL models they used 236,257 fundus images from 72,890 participants from different cohorts including the Severance Main Hospital and the Severance Gangnam Hospital, Seoul, South Korea, the Beijing Eye Study cohort [159], the Singapore Epidemiology of Eye Diseases (SEED) [160, 161], and the UK Biobank [151]. The data was divided into a training set (86,994 images from 27,516 participants), an internal test set (21,698 images from 6,879 participants) from the Severance Main Hospital. Four test sets were further used for testing the trained DL model referred to as external test sets (9,324 images from 4,343 participants from the Severance Gangnam Hospital; 4,234 images from 1,060 participants from the Beijing Eye Study; 63,275 images from 7,726 participants from the SEED study; 50,732 images from 25,366 participants from the UK Biobank). The mean age was 52.92 ± 7.51 years in the training set, and 53.0 ± 7.67 in the internal test set. The VGG16 [92] DL architecture was adopted for training and validation on the data with input image size 300x300 pixels. 37 out of 47 systemic biomarkers were not predicted well from fundus images using DL ($R^2 \leq 0.14$) across all external test sets. Prediction of age in the internal test set gave MAE of 2.43 years and $R^2 = 0.83$. In one of the external test sets with UK Biobank, the MAE of 4.5 years and $R^2 = 0.51$. For sex prediction in the internal test set AUC was 0.96 and in the external test set with UK Biobank AUC was 0.8. For body muscle mass R^2 was 0.52 in the internal test data and R^2 was 0.33 in the external test dataset (Severance Gangnam Hospital data). R^2 values in the internal test data for height was 0.42, bodyweight was 0.36 and creatinine was 0.38. But the performance was poor for height, body weight, and creatinine in the external test data using UK Biobank. The heatmap showed that in both internal and external test datasets retinal vessels and OD were highlighted as the important features for the predictions.

Gerrits et al. [162] investigated DL models for the prediction of cardiometabolic risk factors including age, sex, BP, smoking status, glycaemic status, total lipid panel, sex steroid hormones, and bioimpedance measurements in the Qatar Biobank study [163]. They used 12,000 retinal images from 3,000 participants from the Qatar Biobank initiative and the mean age of participants was 46.07 ± 13.0 years. The MobileNet-V2 [164, 165] DL architecture was used for training and testing with input image size 400×400 pixels. Results were MAE = 3.21 years and $R^2 = 0.85$ for age prediction and for sex AUC = 0.96. The authors reported person-level predictions by averaging the results from the four retinal images available per person (two images per eye). Person-level predictions were often more accurate than those from only one retinal image of either the left or right eye and specifically, for age MAE = 2.78 years and $R^2 = 0.89$, for sex AUC = 0.97, for SBP (MAE = 8.96 mmHg and $R^2 = 0.4$), for DBP MAE = 6.84 mmHg and $R^2 = 0.24$, for Haemoglobin A1c (HbA1c) MAE = 0.61% and $R^2 = 0.34$, for relative fat mass MAE = 5.68 units and $R^2 = 0.43$, for smoking status AUC = 0.78 and for testosterone MAE = 3.76 nmol/L and $R^2 = 0.54$. For other features, including the total lipid panel, the authors reported poor performance from the model with $R^2 \leq 0.05$.

Further, recent studies on DL for the prediction of CVD risk factors and systemic conditions from the retinal fundus images exist [166–169]. The results reported are in line with the literature discussed above.

Retinal age gap for mortality

Very recently, Zhu et al. [170] developed a DL model to predict age from fundus images and then investigated the association between the difference of predicted age and chronological age (retinal age gap) with mortality risk. This work is similar to parts of our own investigation. They used a total of 80,169 fundus images from 46,969 participants from UK Biobank [151], of which 19,200 images from 11,052 participants without previous medical history at baseline were used for training the DL model and 35,913 participants to find the associations between retinal age and mortality. The UK Biobank is a prospective, population-based cohort study comprising UK residents aged between 37 and 73 years. The study has collected a vast amount of phenotypic, genetic, and imaging data, including color fundus photographs and

OCT scans that were obtained from a subset of participants between 2009 and 2010 [171]. The mean age of the participants in the test data was 56.8 ± 8.04 years. The Xception [172] DL architecture was used for training and testing on the data with input image size 299×299 pixels. The DL model performance results in the test data for age prediction was MAE = 3.55 years and $R^2 = 0.81$. Using Cox regression they reported that a one-year increase in the retinal age gap increased the risk of all-cause mortality by 2% as well as the risk of cause-specific mortality attributable to non-cardiovascular and non-cancer disease by 3%. The attention maps generated using Grad-CAM [99] in the test data indicated that the region around the retinal vessels is important for age prediction.

Predicting disease outcomes

This subsection focuses on recent studies of DL applications related to four systemic disease outcomes namely Major Adverse Cardiovascular Event (MACE), CKD, Diabetic Peripheral Neuropathy (DPN), DR using retinal fundus images.

1. *MACE*: Poplin et al. [23] reported an AUC of 0.70 (95% CI: 0.648 to 0.740) for the prediction of MACE within 5 years of retinal imaging from retinal fundus images alone using ensemble DL models. More details on the retinal image data and DL architecture were provided in the above sub-section 'Predicting CV risk factors'. This analysis used data from UK Biobank. In total there were 631 MACE occurring with 5 years of retinal imaging, of which 150 events were present in the test dataset.
2. *CKD*: Sabanayagam et al. [173] developed three DL models to detect CKD from 1) fundus images alone; 2) risk factors which include age, sex, ethnicity, diabetes, and hypertension; 3) hybrid mode, combining images and risk factors. They used data from three cross-sectional studies, the SEED study [160, 161, 174] for developing and validating the DL models, and two independent external datasets, Singapore Prospective Study Program (SP2) [48] and Beijing Eye Study (BES) [175]. They used 12,970 fundus images from 6,485 SEED participants for training and testing the DL models and 7,470 images from 3,735 SP2 participants and 3,076 images from 1,538

BES participants for external testing. The mean age in both the training and testing datasets was 58.4 ± 9.9 years. cCondenseNet [176] DL was used for training and testing on the data with input image size 512×512 pixels. For CKD prediction in the SEED test set they reported an AUC of 0.911 for retinal images, 0.916 for risk factors, and 0.938 for hybrid mode; in SP2 test set the AUC was 0.733 for retinal images, 0.829 for risk factors and 0.810 for hybrid mode; in BES test set the AUC was 0.835 for retinal images, 0.887 for risk factors and 0.858 for hybrid mode. They reported similar performance (slightly lower values) in people with diabetes and hypertension. Another recent study reported the stratification of CKD and Type 2 Diabetes (T2D) from fundus images in combination with clinical dataset using DL obtained an AUC of 0.85-0.93 and MAE for predicting estimated Glomerular Filtration Rate (eGFR) was $11.1\text{--}13.4 \text{ ml min}^{-1} \text{ per } 1.73 \text{ m}^2$ [177]. They used 115,344 fundus images from 57,672 participants.

3. *DPN*: There is little research on predicting DPN directly from fundus images using DL. Recently Cervera et al. [178] developed a DL model to detect DPN from fundus images. Using Sankara Nethralaya Diabetic Retinopathy Epidemiology and Molecular Genetics Study (SNDREAMS) [179], a population study based in South India, the authors obtained 23,784 retinal fundus images from 1,561 diabetic participants, of which 17,028 images from 1,081 participants for training and 1,892 retinal images from 121 individuals for testing the DL model. Of 1,516 participants, 276 had DPN and the mean age of the population with DPN was 57.55 ± 10.059 and without DPN was 55.71 ± 10.214 . As a part of their hyper-parameter tuning (via grid search) they tried three different DL models: Inception-v3 [90], Squeezenet v1.0 [180] and Densenet121 [94], all with input image size 720×576 pixels on one of the 5-fold cross-validations data split. Finally, Squeezenet was selected based on the comparative evaluation on the validation split of the specific 5-fold cross-validations data and used for training and testing on all 5 folds. The AUC reported in the whole test dataset was 0.7097, 0.715 in the test data without DR participants and 0.8673 in the test data which includes only participants having DR.

4. *DR*: several studies applied DL to detect DR signs [181]. Gulshan et al. [182] used 128,175 fundus images from 69,572 individuals from EyePACS [152] (United States) and three eye hospitals in India (Aravind Eye Hospital, Sankara Nethralaya, and Narayana Nethralaya) for DL algorithm development. The trained DL models were tested on two datasets: EyePACS, with 9,963 images from 4,997 individuals, and Methods to Evaluate Segmentation and Indexing Techniques in the field of Retinal Ophthalmology (MESSIDOR)-2 [142], with 1,748 images from 874 individuals. The mean age was 55.1 ± 11.2 years in the training dataset, 54.4 ± 11.3 years for the test data from EyePACS, and 57.6 ± 15.9 years for test data from MESSIDOR. Inception-v3 [90] DL was used for training and testing. For predicting Referable Diabetic Retinopathy (RDR), defined as moderate and worse DR, referable Diabetic Macular Edema (DME), or both, they reported an AUC of 0.991 (95%CI, 0.988 - 0.993) for EyePACS and 0.990 (95%CI, 0.986 - 0.995) for Messidor-2 in the respective test datasets.

Predicting CV risk score

Chang et al. [169] used DL model to predict carotid artery atherosclerosis from fundus images named the Deep-Learning Fundusoscopic Atherosclerosis Score (DL-FAS). They used 15,408 fundus images from 6,597 participants from Health Promotion Center of Seoul National University Hospital (HPCSNUH) [183] for training the DL model and tested it on 32,227 participants. The mean age of the participants in the test data was 52.6 ± 10.6 years. Xception [172] DL was used for training and testing. For predicting carotid artery atherosclerosis (as a binary classification problem) the authors reported AUC in the test data as 0.713 and Area Under the Precision-Recall Curve (AUPRC), accuracy, sensitivity, specificity, positive and negative predictive values as 0.569, 0.583, 0.891, 0.404, 0.465, and 0.865 respectively. They also found that the individuals in the test data with DL-FAS > 0.66 had an increased risk of CVD deaths compared to the individuals with DL-FAS < 0.33 with a hazard ratio of 8.33 (95% CI 3.16-24.7). Their saliency maps, generated using guided propagation, suggest that retinal vessels contribute to the prediction of atherosclerosis.

2.5 Conclusions

Several software tools are available to measure morphometric properties of the retinal vasculature in Zone B and Zone C of fundus camera retinal images. Commonly used parameters are the summative measure of width, CRAE and CRVE; their ratio, the AVR; and vessel tortuosity, vessel width, bifurcation angles, branching angles, and vessel caliber. Subsets of these have been implicated with the early stratification of DR, MD, hypertension, stroke, neovascular glaucoma, and other cardiovascular diseases. The accuracy of the estimates of these parameters depends on accurate blood vessel segmentation and artery-vein classification in the retinal image which is a challenging task. Accuracy and uncertainty estimation for retinal parameters, and more generally in medical image analysis, is in itself an active field of investigation [184].

DL algorithms have outperformed other ML algorithms in the classification of disease conditions associated with retinas as described in Section 2.3. The recent literature was reviewed for predicting CV risk factors, CV risk score, systemic diseases like MACE, CKD, DPN, DR using DL from only retinal image as input to the model. The literature was also reviewed on the retinal age gap for predicting All Cause Death (ACD).

The literature that was identify using DL mainly focused on middle-aged individuals with mean age of less than 55 years old. It considers participants from different world regions including UK, USA, India, Singapore, Qatar, China, and Korea. Some of the datasets were collected from screening programs, where a large population voluntarily participated in studies such as UK Biobank or EyePACS. Most of the retinal images considered in this analysis are from healthy participants. There are built data repositories supporting the same analysis in the wider population of different age groups and health conditions from retinal images using DL.

The next chapter (Chapter 3) introduces the MESSIDOR and Indian Diabetic Retinopathy Image Dataset (IDRID) datasets, to be used in the framework based on synthetic, controlled data to identify the best performing DL classification architecture to adopt for the main analysis work reported in this thesis.

Chapter 3

Material

3.1 About this chapter

Three datasets were used for the experiments presented in this thesis; two of them are public datasets and the other one is a clinical dataset, Genetics of Diabetes Audit and Research in Tayside Scotland (GoDARTS), collected as part of a large, longitudinal cohort study. This data is not available publicly. The public datasets were used for synthetic data generation (Chapter 4) and for fine-tuning the hyper-parameters of Deep Learning (DL) models (Chapter 4) subsequently used in the analysis of the clinical dataset. This chapter describes the specifications of the three datasets used in this thesis.

3.2 Public datasets

There are several public repositories of fundus images available, for example, STARE [185], DRIVE [186], DIARETDB1 [187], ROC [188], HEI-MED [189], e-optha [190], MESSIDOR [12], Kaggle DR [191], IDRID [13]. The MESSIDOR [12] and IDRID [13] datasets were selected for the initial experiments based on several criteria, including large volume, high-quality images, and availability of ground truths for DR grading and pixel-level annotations. MESSIDOR was used for synthetic data generation and fine-tuning the hyper-parameters of DL models (such as number of Fully Connected (FC) layers, number of FC

nodes, optimizer functions, loss functions, batch size, etc) and IDRID as additional test data for validating the fine-tuned model.

3.2.1 MESSIDOR

MESSIDOR (Méthodes d'Évaluation de Systèmes de Segmentation et d'Indexation Dédiées à l'Ophthalmologie Rétinienne, i.e., Methods to Evaluate Segmentation and Indexing Techniques in the field of Retinal Ophthalmology) [12] consists of 1,200 retinal fundus color images, macula centered, acquired in three French ophthalmology departments using a color video 3CCD camera mounted on a Topcon TRC NW6 non-mydratic retinograph. All the images were captured with a 45° field of view and at three different image resolution levels: 1440 × 960 (width × height), 2240 × 1488 and 2304 × 1536 pixels. 800 images were acquired with pupil dilation and 400 without, all during routine clinical examinations. Figure 3.1 shows six sample MESSIDOR images representing variations in image quality. The MESSIDOR dataset offers high-quality retinal fundus images with good visualization of the retinal structure. For this experiment, all MESSIDOR retinal images were used without implementing any specific image quality checks.

Each image was graded on a 4-point scale for Diabetic Retinopathy (DR) and a 3-point scale for risk of Diabetic Macular Edema (DME) by medical experts based on the number of Microaneurysms (MA)/Hemorrhages (HE)/Neo Vascularization (NV)/Hard Exudates (EX) visible in the retinal image. The criteria for grading DR and DME are provided in Table 3.1.

Of the 1,200 images provided, 547 are graded 0 (no DR sign/lesion) 153 1, 247 2, and 253 3. Similarly, for DME, 974 images are graded 0 (no DME), 75 1, and 151 2.

The MESSIDOR image set (in *tif* format) can be downloaded from [192] by filling out a simple online form with basic contact details.

3.2.2 IDRID

IDRID (Indian Diabetic Retinopathy Image Dataset) [13] contains 516 macula-centered, color retinal fundus images of diabetic patients acquired during routine clinical examinations

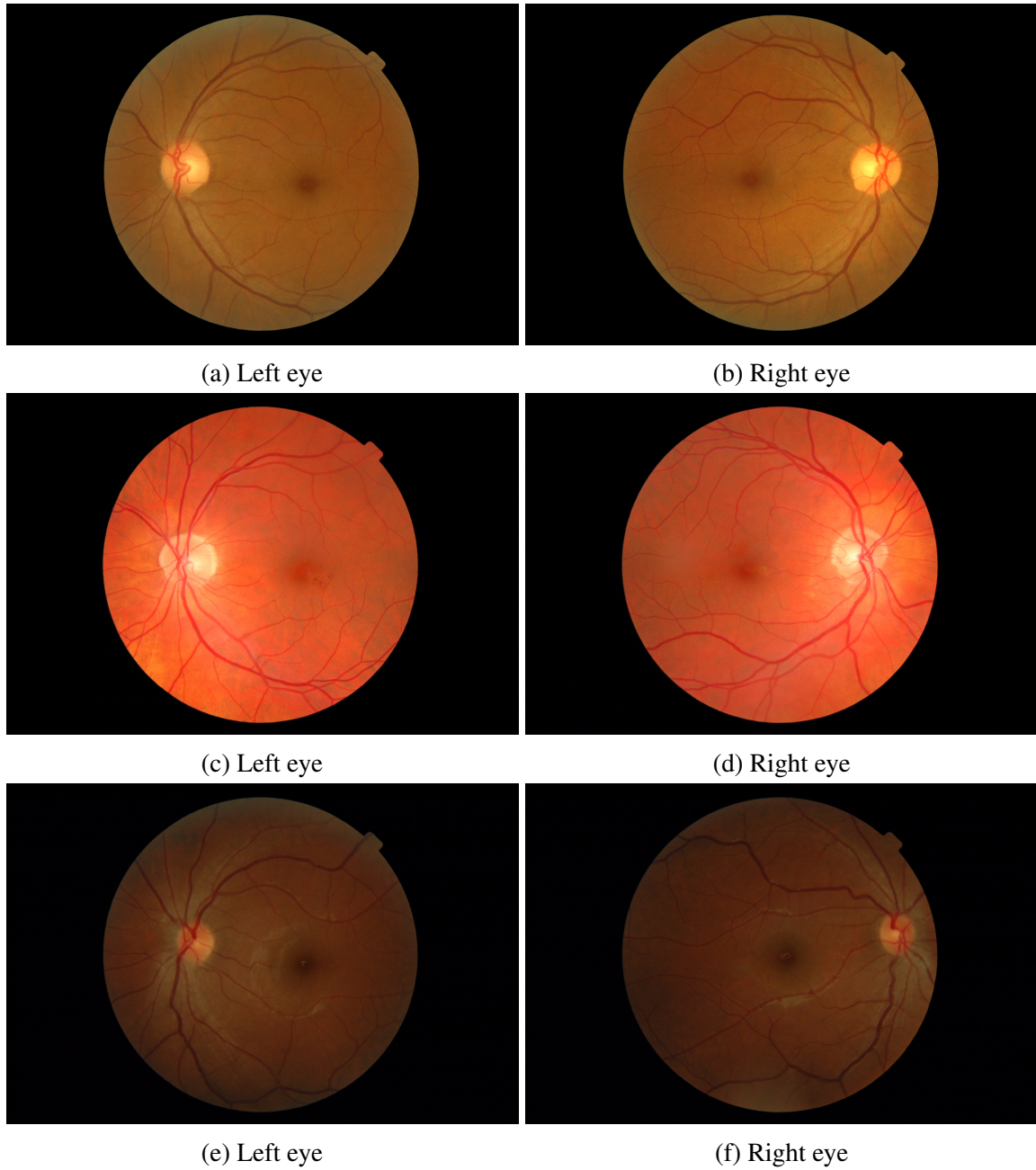


Fig. 3.1 Sample MESSIDOR fundus images [12].

Table 3.1 MESSIDOR criteria for grading for DR and DME [12]; n_MA: number of MA, n_HE: number of HE, NV = 1: NV, NV = 0: no NV.

DR grade	Criteria
0 (Normal)	(n_MA = 0) AND (n_HE = 0)
1	(0 < n_MA <= 5) AND (n_HE = 0)
2	((5 < n_MA < 15) OR (0 < n_HE < 5)) AND (NV = 0)
3	(n_MA >= 15) OR (n_HE >= 5) OR (NV = 1)
DME grade	Criteria
0 (No risk)	No visible hard exudate
1	Shortest distance between macula and EX > one papilla diameter
2	Shortest distance between macula and EX <= one papilla diameter

in Nanded, India, with a Kowa VX-10 α digital fundus camera. The images were captured with a 50° field of view and a resolution of 4288 \times 2848 pixels. The pupils of all subjects were dilated before image acquisition.

Six sample IDRID fundus images representing the overall variation in the image quality are shown in Figure 3.2. The IDRID dataset provided high-resolution retinal fundus images with excellent image quality. In the experiment with IDRID dataset, all retinal images available were utilized without conducting any specific image quality checks. Three types of ground truth are provided by medical experts:

1. *Pixel-level annotation*: provided for 81 fundus images with signs of DR and available in *tif* file format as binary masks for MA, HE, Soft Exudates (SE), EX and Optic Disc (OD).
2. *Image-level grading*: each fundus image is graded on a 5-point scale for DR and on a 3-point scale for DME according to international standards. The scales are reported in Table 3.2.
3. *OD and fovea center co-ordinates*: provided for all 516 images.

Table 3.2 IDRID criteria for grading for DR and DME [13]; NPDR, PDR.

DR grade	Criteria
0: No apparent retinopathy	No visible sign of abnormalities
1: Mild Non-Proliferative Diabetic Retinopathy (NPDR)	Presence of MA only
2: Moderate NPDR	More than just MA but less than severe NPDR
3: Severe NPDR	Any of the following: • >20 intraretinal HEs • Venous beading • Intra-retinal microvascular abnormalities • no signs of Proliferative Diabetic Retinopathy (PDR)
4: PDR	Either or both of the following: NV Vitreous/pre-retinal HE
DME grade	Criteria
0	No Apparent EX
1	Presence of EX outside the radius of one disc diameter from the macula center
2	Presence of EX within the radius of one disc diameter from the macula center

3.3 Clinical dataset: GoDARTS

The clinical dataset, GoDARTS [193], consists of retinal images and linked patient data including demographic information, clinical measurements, drug history, genetic profile, clinical outcomes, and further images like CT and MRI for subsets of patients. The following sections describe the dataset and the pre-processing methods applied to the retinal images.

3.3.1 GoDARTS history and structure

Diabetes Audit and Research in Tayside Scotland (DARTS) was started in 1996 to identify all the patients with diabetes in the Tayside region of Scotland through electronic record linkage and to improve health care. In 1998, consenting patients from DARTS were recruited to GoDARTS and invited to provide a blood sample for DNA extraction to be used for research. Also, phenotypic data (clinical and lifestyle factors) was collected from participants through questionnaires and clinical examinations. The aim of the GoDARTS cohort study is to investigate the contribution of genetic and environmental factors which are associated with

disease onset, progression, and response to treatment in a diabetic population. Full details about the GoDARTS cohort are given in [193].

Individuals were recruited from inception until 2015. After this date, the recruitment continued through several other initiatives like SHARE [194], GoSHARE, and GoDARTS-Scotland [193]. By 2015, participants in GoDARTS totaled 18,306, of which 10,149 were Type 2 Diabetes (T2D) and 8,157 healthy controls at the time of recruitment. GoDARTS medical records for patients with diabetes have been linked to retinal images from the Scottish Diabetic Retinopathy Screening (DRS) program [195].

Access to GoDARTS data for analysis is achieved through the Safe Haven (SH) within the Health Informatics Center (HIC) of the University of Dundee. The data available can be grouped broadly into two categories, described below.

1. Data collected during recruitment: blood pressure/pulse, height/weight (Body Mass Index (BMI)) waist measurement, diabetes history, menopausal history, lifestyle questionnaire, biochemistry - Haemoglobin A1c (HbA1c), Cholesterol (Chol), High-Density Lipoprotein (HDL), low-density lipoprotein (LDL), Creatinine, Tryglicerides (Trig).
2. Data collected through electronic health record linkage: including demography, ethnicity, biochemistry, outpatient appointments, hospital admissions, maternity, psychiatric cases, cancer register, prescribing, National Records of Scotland (NRS) deaths, Community Health Index (CHI) deaths, kidney register.

To protect the confidentiality and identity of individuals, the CHI numerical identifier unique to each patient, issued on first registration with a GP or admission to a hospital in Scotland, was converted to proCHI, a project-specific code [196]. This code was adopted specifically for the GoDARTS project.

Retinal Images

The retinal images for GoDARTS patients were obtained from the Scottish DRS program. Most patients are T2D and the rest Type 1 Diabetes (T1D). Retinal image capture followed the standard Scottish DRS protocol [195]; in brief, a Canon CR6-45NM Non-Mydriatic and

Canon CR-DGi Non-Mydriatic retinal cameras were used with 45° field of view [197]. All images are macula-centered. In total there are 102,455 retinal images from 8,594 individuals, obtained at multiple time points (varying from patient to patient) for about 12 years starting from 2006. Both left and right eye images are available for most but not all individuals, some having only either the left or right one. Multiple images of the same eye are available for many individuals for quality assessment reasons.

As the 102,455 retinal images were acquired in routine clinical practice and not within a study, there are 34 different image resolution levels: 3504×2336 (65,279 images, 63.71%), 3456×2304 (27,139, 26.49%), 2544×1696 (3,171, 3.1%), 3267×2178 (2,328, 2.27%) and 3888×2592 (1,244, 1.21%). These resolutions account for the vast majority of the images, with other resolutions representing only 3.21% of the whole set. Figure 3.3 shows sample left- and right-eye retinal images of different individuals from GoDARTS and the variability in the image quality are clearly visible. During the analysis with GoDARTS fundus images, all available retinal images were utilized without conducting any specific image quality checks.

Patient features used

The following GoDARTS features associated with major risk factors of CVD and microvascular disease outcomes for the analysis were selected.

1. Demographic information: age (at the time of retinal imaging; continuous variable) and sex (binary variable).
2. Clinical measurements (continuous variables): Diastolic Blood Pressure (DBP), Systolic Blood Pressure (SBP), BMI, Glycated Haemoglobin (GH), HDL, Total Cholesterol (TC), Trig, Genome-Wide Polygenic Risk Scores (GWPRS).
3. Disease outcomes (binary variables): Major Adverse Cardiovascular Event (MACE), All Cause Death (ACD), Chronic Kidney Disease (CKD), DR, Diabetic Peripheral Neuropathy (DPN).

Extraction of feature data

The feature data is derived/computed/extracted in different ways depending on the type of the feature, as follows.

1. **Demographic features:** A demography file was used to extract the features from this category.
 - (a) *Age at imaging:* computed as the difference between the date when the retinal image was captured and the date of birth, sourced from the demography file. Age is expressed in years.
 - (b) *Sex:* directly extracted from the column 'sex' in the demography file, renamed this column as 'sex_male' and provided binary labels, replacing the original M and F with 1 for male and 0 for a female to facilitate programming, refer Section [5.2](#).
2. **Clinical measurements:** Computed using the biochemistry file, storing results from blood, urine and other tests. The median values of the various clinical measurements over the 3 years following the date of the first retinal image available for an individual were calculated. These median values were used as labels for training DL models with retinal images (more details in Section [5.4](#)).
3. **Disease outcomes:** Features computed from biochemistry, hospital admissions and NRS deaths files. The criteria for each outcome are:
 - (a) **MACE:** individuals admitted to the hospital due to non-fatal myocardial infarction or acute coronary syndrome, non-fatal stroke or cardiovascular death are marked with a MACE label as 1, otherwise the label is 0. This information is extracted from the hospital admissions file using the International Classification of Diseases (ICD)-10 codes, I21-I23 and I60-I63. Individuals with a history of MACE prior to the first retinal image available were excluded from the MACE analysis.
 - (b) **ACD:** the status was obtained from NRS Death which includes ICD-coded cause of death.

- (c) CKD: individuals who has been diagnosed of having CKD stage 3 or higher if their estimated Glomerular Filtration Rate (eGFR) is less than $60 \text{ ml/min/1.73m}^2$ for ≥ 3 months are labelled as 1 otherwise 0.
- (d) DR: individuals with no visible retinopathy were labeled as 0 and those with DR grade R1 to R4 were labeled as 1. This classification is referred to as any DR vs no DR.
- (e) DPN: individuals are labeled as 1 if 10g monofilament test was reported to be absent, or foot vibration reported to be abnormal ($>25\text{V}$ Vibration Perception Threshold (VPT)) or foot sensation reported to be abnormal at the time of diabetic foot screening otherwise labeled as 0.

The clinical measurements and disease outcomes for the proCHI codes were provided by a team of clinicians and researchers as a part of the collaboration (Section 1.1.1). The descriptive characteristics of the dataset at baseline (first retinal image acquisition) are provided in Table 3.3 and Table 3.4. These characteristics are mean, standard deviation, range (minimum value & maximum value) and Interquartile Range (IQR), the difference between 75^{th} ($Q3$) and 25^{th} ($Q1$) percentiles, $\text{IQR} = Q3 - Q1$. These are provided for the whole data, as well as for males and females to check for possible bias due to sex. The “proportion” column in Table 3.4 shows the percentage of individuals representing the feature in the respective sex category. The column labeled “p-value” in Table 3.3 was calculated using the Welch Two Sample t-test in R to compare each continuous feature against the sex group, whereas in Table 3.4, the “p-value” column was computed using Pearson’s Chi-squared test with Yates continuity correction in R to compare each categorical feature against the sex group. A p-value less than 0.05 indicates a statistically significant difference between the mean values of male and female groups for continuous features or a significant association between two variables for binary features.

Table 3.3 Descriptive characteristics of continuous features in GoDARTS; M = Male; F = Female; n = individuals used for feature

Feature (units)	Sex	n	Mean (std)	Range	IQR	p-value
Age (years)	Overall	8,570	66.11 (11.77)	17.98-96.45	58.8-74.71	0.0009
	Male	4,819	65.73 (11.4)	23.81-96.22	58.53-74.05	
	Female	3,751	66.59 (12.21)	17.98-96.45	59.09-75.38	
DBP (mmHg)	Overall	6,047	76.0 (7.88)	48.0-108.5	70.5-81.0	3.67E-09
	Male	3,386	76.52 (7.99)	48.0-108.5	71.0-82.0	
	Female	2,661	75.33 (7.7)	51.0-106.0	70.0-80.0	
SBP (mmHg)	Overall	6,656	138.9 (11.52)	93.5-208.0	132.5-144.5	0.0045
	Male	3,721	138.55 (11.53)	93.5-208.0	132.0-144.0	
	Female	2,934	139.35 (11.5)	98.0-198.5	133.0-145.0	
BMI (Kg/m ²)	Overall	6,562	31.15 (6.0)	15.2-59.1	27.0-34.3	< 2.2e-16
	Male	3,666	30.52 (5.34)	15.95-59.1	26.85-33.31	
	Female	2,895	31.96 (6.66)	15.2-57.7	27.25-35.7	
GH (%)	Overall	6,214	7.52 (1.15)	6.08-12.8	6.7-8.1	0.1698
	Male	3,473	7.5 (1.12)	6.08-12.7	6.7-8.05	
	Female	2,740	7.54 (1.19)	6.08-12.8	6.69-8.1	
HDL (mmol/L)	Overall	6,656	1.34 (0.35)	0.23-3.94	1.11-1.48	< 2.2e-16
	Male	3,721	1.27 (0.32)	0.23-3.55	1.06-1.38	
	Female	2,934	1.42 (0.37)	0.23-3.94	1.2-1.58	
TC (mmol/L)	Overall	6,656	4.37 (0.85)	1.45-14.28	3.82-4.79	< 2.2e-16
	Male	3,721	4.25 (0.84)	1.45-14.28	3.7-4.68	
	Female	2,934	4.53 (0.84)	2.41-9.59	3.98-4.95	
Trig (mmol/L)	Overall	3,471	2.31 (1.95)	0.31-44.61	1.28-2.71	0.0005
	Male	1,947	2.4 (2.27)	0.31-44.61	1.28-2.79	
	Female	1,524	2.18 (1.43)	0.32-14.51	1.28-2.63	

3.3.2 Pre-processing

GoDARTS retinal images have significant variations in terms of image dimensions (Section 3.3.1), luminosity, color, focus, and in general, quality. To reduce the effect of these variations in the DL training, the images were pre-processed. The block diagram describing the pre-processing steps is shown in Figure 3.4. All pre-processing was done automatically using the cv2 python package [198].

Table 3.4 Descriptive characteristics of binary features in GoDARTS; M = Male; F = Female; n = individuals used for feature.

Feature	Sex	n	Proportion (%)	p-value
sex_male	Overall	8,570	56.23	
	Male	4,819	-	-
	Female	3,751	-	
MACE	Overall	6,656	25.48	
	Male	3,721	26.47	0.04
	Female	2,934	24.23	
ACD	Overall	6,656	35.95	
	Male	3,721	37.87	0.0003
	Female	2,934	33.54	
CKD	Overall	7,562	44.12	
	Male	4,118	60.93	< 2.2e-16
	Female	3,444	24.01	
DR	Overall	7,601	59.57	
	Male	4,271	59.03	0.283
	Female	3,330	60.27	
DPN	Overall	7,688	37.92	
	Male	4,351	37.9	0.991
	Female	3,337	37.94	

A sample retinal image before and after pre-processing is shown in Figure 3.5. The process is as follows. The original image is converted to a gray-level image using the `opencv` [198] library function, `cv2.cvtColor()`, then to a binary image using `cv2.threshold` with a threshold of value 10 (this value was chosen by trials). The bounding square circumscribing the circular retinal region imaged is then computed to remove excess black regions. Next, a circular visibility mask is fitted to the retinal region to eliminate artifacts (small protrusions) appearing in some images. Finally, images are resized to 512×512 pixels, Contrast Limited Adaptive Histogram Equalization (CLAHE) [199] applied on each color channel (R, G, B) and intensities are normalized to $[0,1]$.

3.4 Conclusions

This chapter has presented two public datasets Methods to Evaluate Segmentation and Indexing Techniques in the field of Retinal Ophthalmology (MESSIDOR) and Indian Diabetic Retinopathy Image Dataset (IDRID) and a clinical dataset GoDARTS. Retinal images in MESSIDOR and IDRID were collected from diabetic individuals during routine clinical examinations in France and India respectively. They both provide image-level labels for DR grading. GoDARTS is a cohort study started in 1998 and it includes retinal images obtained from the Scottish DRS program and linkage to Electronic Medical Record (EMR) of recruited participants in Tayside, Scotland.

The main strengths of GoDARTS compared to MESSIDOR are more number of participants, phenotypic and genotypic data, and linkage to EMR. GoDARTS is a rich clinical database with longitudinal phenotypic data (bio-chemistry, prescribing, morbidity and demography). It also includes multiple imaging modalities like retinal images, Computerized Tomography (CT), Magnetic Resonance Imaging (MRI) which can be analyzed for disease outcomes using DL.

For this thesis, a subset of GoDARTS bio-resources namely retinal images, demographic features, clinical measurements related to Cardiovascular (CV) risk factors, and outcomes related to systemic diseases were considered. There are a total of 102,455 retinal color fundus images with 45° field of view from 8,594 individuals, obtained at multiple time points in GoDARTS. The mean age of individuals based on the first retinal image available is 66.11 ± 11.77 years. Image pre-processing was applied to all the retinal images in the GoDARTS to account for significant variation in the images during DL training, described in Chapter 5, Chapter 6 and Chapter 7.

To identify the robust DL architecture that can be trained on retinal images of GoDARTS, the MESSIDOR dataset was used for the generation of the synthetic dataset and fine-tuning the hyper-parameters of DL models and IDRID was used as an independent dataset for validation the DL models, described in Chapter 4.

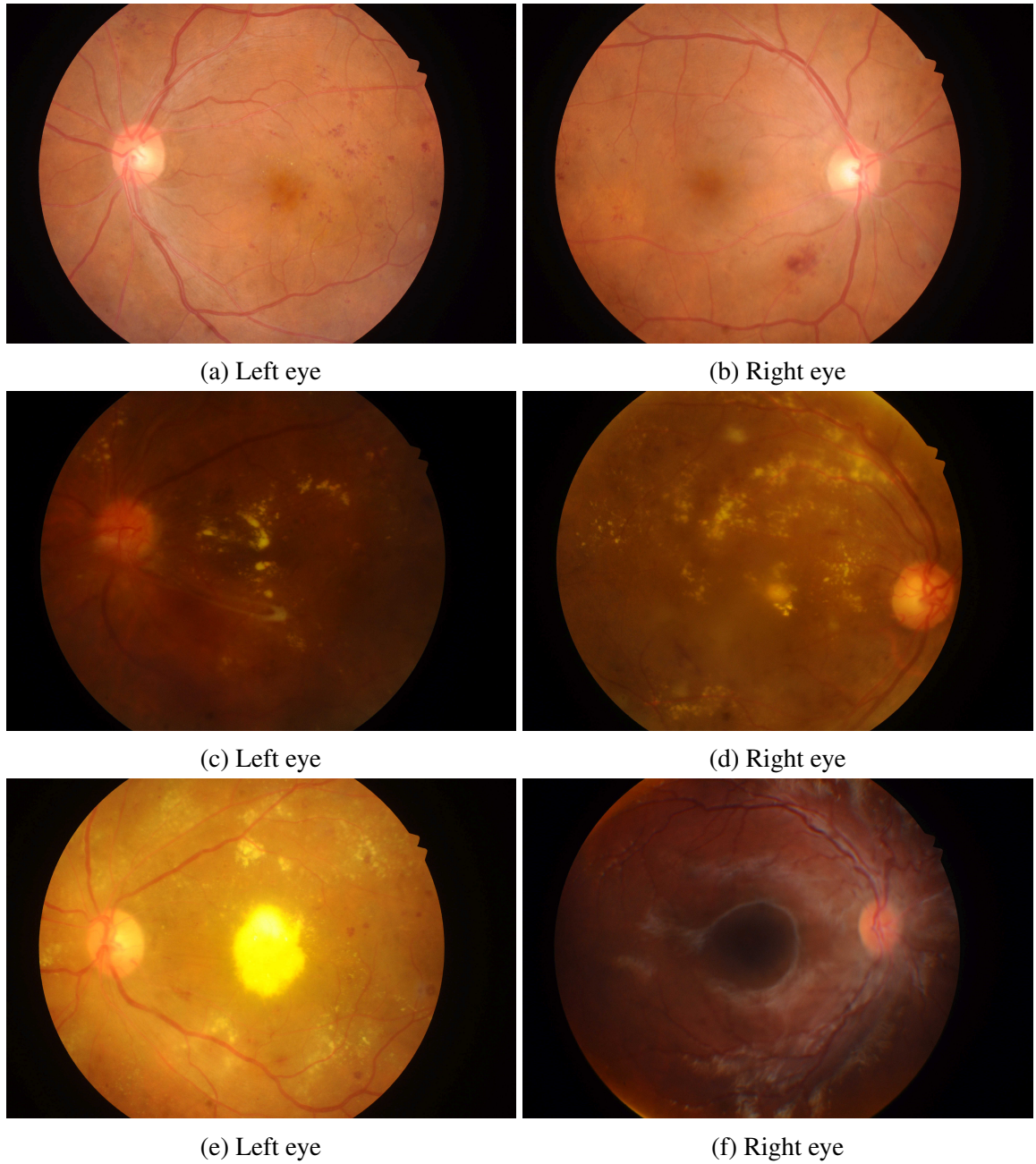


Fig. 3.2 Sample IDRIID fundus images [13].

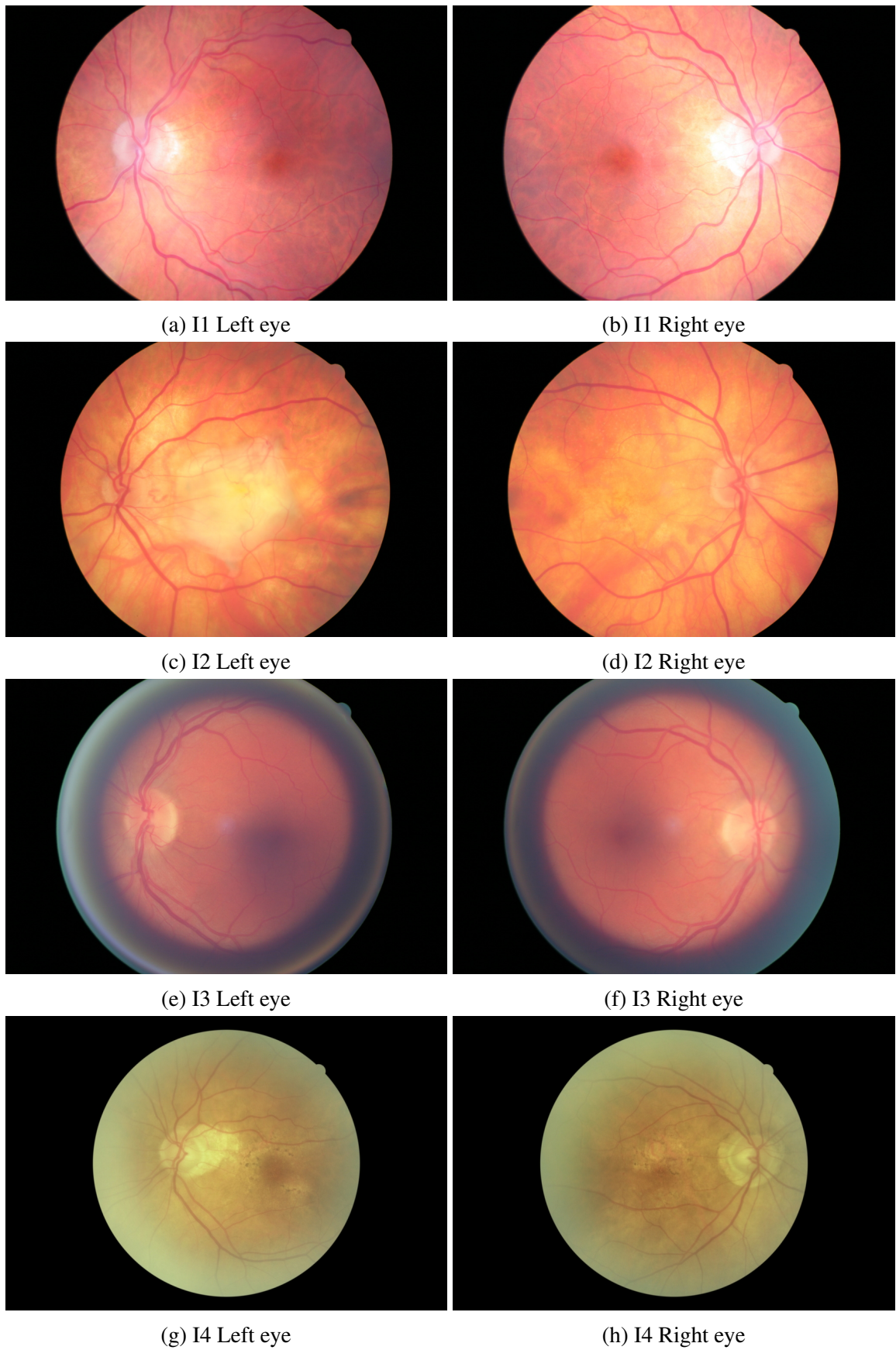


Fig. 3.3 Sample GoDARTS fundus images of left- and right-eyes of different individuals (I).

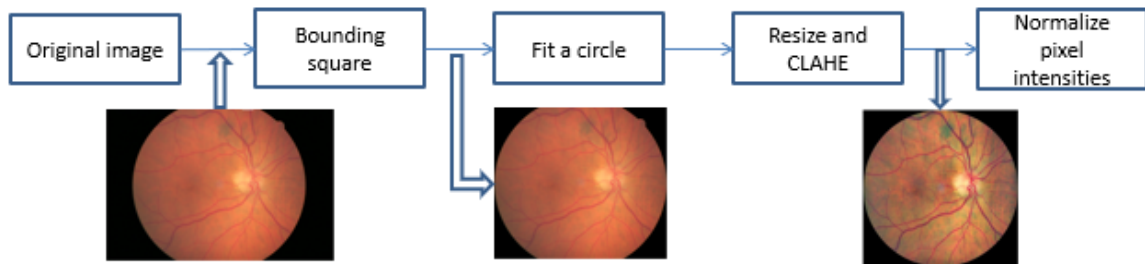
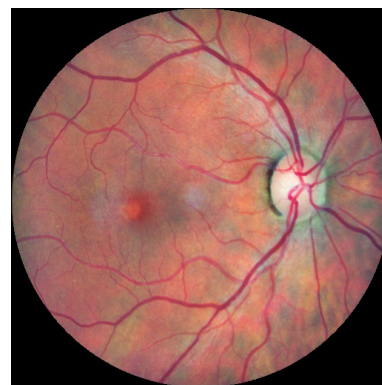


Fig. 3.4 Block diagram for image pre-processing in GoDARTS.



(a) Original image

(b) Pre-processed image

Fig. 3.5 Example of image pre-processing in GoDARTS.

Chapter 4

Identifying robust CNN architecture

4.1 Introduction

In this chapter, a framework is proposed for generating synthetic datasets parameterized by difficulty level to test classifiers ¹. The framework is demonstrated using the MESSIDOR dataset to generate synthetic data, and the image pre-processing applied to MESSIDOR is discussed. Several well-known CNN architectures are employed to identify the best DL architecture using the synthetic dataset from MESSIDOR, and the models are also validated on an independent IDRID dataset to solve a 5-class classification problem.

4.1.1 Context and motivation

The general context is the validation of medical image analysis systems; see ([200], [201], [202]). The motivation for this work is the cost of generating large, annotated datasets of real clinical data to provide ground truth, or gold standard, for testing. The problem is particularly acute for Deep Learning (DL) models containing millions of parameters to be trained, requiring ever-growing training and testing sets.

¹*Syed, M. G., and Trucco, E. (2019). A framework to generate synthetic test image sets parameterized by difficulty level. Medical Image Understanding and Analysis: 23rd Annual Conference, MIUA 2019 (invited talk).* This publication is based on the work presented in this chapter.

Various approaches to this challenge exist. First, recruit and coordinate a cohort of specialist annotators [202]. This is extremely expensive, unfeasible in most cases, and inevitably limited by the annotator's speed [203]. Second, leverage small volumes of annotations to generate much larger ones [204]. Third, resort to transfer learning to deploy models pre-trained in related domains. Second and third, both are time-consuming and also heavily dependent on annotations and computational resources. Fourth, generate synthetic data, addressed here. This chapter does not aim to compare and contrast the various approaches in the literature.

4.1.2 Generating synthetic medical images

Synthetic images provide an interesting alternative to annotations of real data. Provided sufficient realism for the target task can be achieved, one can generate, in principle, arbitrarily large volumes of fully characterized data, as opposed to annotating only specific properties of real images like lesions or image labels. A classic example outside the medical domain is the Kinect module for human pose recognition, which was trained on one million synthetic human poses generated from 8,000+ manually segmented body scans [205]. DL, especially Generative Adversarial Network (GAN) architectures [206], have been used recently to push the realism of synthetic medical images, including prostate lesions in MRI scans [207], CT scans from MR ones [208], colonoscopy frames [209], and fundus retinal images [210, 211], including with lesions [212, 213].

4.1.3 Parameterizing synthetic datasets by difficulty level

All approaches described in the above section have concentrated on achieving maximum realism. However, systematic validation would benefit from structuring a synthetic dataset by difficulty level given a specific task. This requires a quantitative characterization of the difficulty of the target problem. This chapter proposes a protocol identifying a set of parameters characterizing the data, together with their numerical ranges. Sampling the ranges yields a multi-dimensional, discrete parameter space in which each point identifies a dataset

with a well-defined difficulty level. The protocol creates synthetic images by augmenting real images with patches (small regions) sampled from a dictionary of real lesion patches.

4.1.4 Fine-tuning and validation

The synthetic images were created using the retinal images from the MESSIDOR dataset, refer to Section 3.2.1. Different DL architectures were trained on the controlled synthetic images varying the values of model hyper-parameters. These models were validated using 10-fold cross-validation and performance was compared across the models referred to as "cube experiment". As a sanity check, the fine-tuned DL architectures are validated on an independent dataset, IDRID, refer Section 3.2.2.

4.2 Proposed framework for synthetic data generation

This section presents the protocol steps, explaining them with an example leading to a low-dimensional parameter space that can be visualized.

1. Identify the target problem: here, to test a normal-abnormal classifier for fundus camera color retinal images. The definition used was normal = as healthy and abnormal = as containing lesions. In our example, three types of real lesions were considered: Microaneurysms (MA), Hard Exudates (EX), and Hemorrhages (HE). We do not model, at this stage, co-occurring lesions (i.e., their co-occurrence probabilities) nor their position in the retinal images.
2. Identify the set of parameters characterizing difficulty and their ranges. Lesions appear as small image regions (patches) with specific textural properties. In our low-dimensional example, three parameters influencing the difficulty of the classification problem were considered:
 - (a) the *size of the patch* (the smaller the more difficult to classify an image correctly as abnormal), s ;

- (b) the *number of patches* in an image (the smaller the more difficult), n ;
 - (c) the *contrast of the patch (transparency)* with respect to the background (the lower the more difficult; see Section 4.3 for details), α .
3. Define a discrete grid by sampling the parameter space. A number of points are sampled on each axis, yielding a discrete grid in parameter space. The number of grid points must compromise between feasibility, i.e. building and testing classifiers over the whole grid in an acceptable time on the computational platform available, and exhaustiveness, i.e. high sampling frequencies to achieve thick grids. Ranges must be decided considering realistic values. In our example, in the interest of visualization, a $3 \times 3 \times 3$ grid was used.
 4. Generate synthetic images. For each point in parameter space, generate an image set. The set is associated with a precise level of difficulty for the target problem. In our example, a single point generates an image set defined by a specific triple, (s, n, α) . The real image samples were augmented from a dictionary of real lesion patches, adapting them according to (s, n, α) values and blending them in a real image.
 5. Run the classifier on the whole grid. For each point in parameter space, build and test a classifier by cross-validation, dividing the (s, n, α) -th dataset as usual; save the values of the performance criteria for that point. At the end, each point in the grid will be associated with a characterization of performance; in our example, accuracy.
 6. Analyze the results. The model performs well if the most difficult problems in the grid are solved with acceptable values of the performance criteria. The grid allows one to identify regions in parameter space generating hard problems, fully characterized, and which model's hyper-parameters must be improved.

4.3 Materials and methods

The retinal images from MESSIDOR (described in Section 3.2.1) were used for generating synthetic data. 547 healthy images (0 Diabetic Retinopathy (DR) and has no lesions)

were augmented using manually cropped lesion patches from images containing signs of pathologies (DR graded 1,2,3) as described in detail in the following sections.

4.3.1 Pre-processing

The images from MESSIDOR have 3 different sizes in pixels and all images include black corners around the circular retinal region captured. To eliminate the black corners, a circle is fitted on the retinal region, and the largest square is inscribed with the circle computed. Well-known functions (like `measure`, `optimize`) were used from `skimage` [214] and `scipy` [215] libraries in python to fit the circle. The center and radius of the circle were used to obtain the maximum square inscribed in the retina region by applying basic geometry principles as shown in Figure 4.1, where $C(x,y)$ is the center point of the circle and r is the radius. The resulting images are resized to 256×256 pixels as commonly done when using DL algorithms on images. A sample image summarizing the pre-processing steps is shown in Figure 4.2.

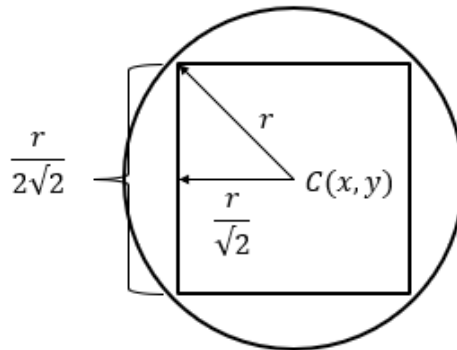


Fig. 4.1 Measurements for maximum square inscribed in a circular retina.

4.3.2 Dictionary of lesion patches

From the 640 pre-processed retinal images graded DR 1 to 3, patches of 3 sizes were cropped manually (9×9 , 17×17 , 25×25 ; target image size 256×256), to account for the different extents of visible lesions. The patch size is designed to capture enough texture to characterize a lesion in most cases. In total, the lesion dictionary contains 120 patches per size and per

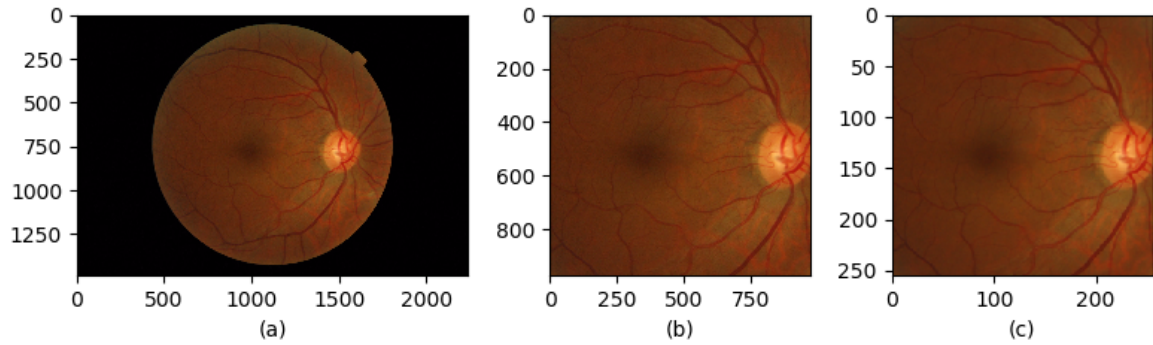


Fig. 4.2 Image pre-processing in MESSIDOR: (a) Original MESSIDOR image; (b) largest inscribed square (retinal region considered); (c) image resized to 256×256 .

lesion. Examples are shown in Figure 4.3. The lesions were cropped manually using GNU Image Manipulation Program (GIMP), a free and open source image editor [216].

4.3.3 Parameter space

This example is restricted to a grid size of 3 values on each one of 3 axes (total 27 grid points), allowing a compact explanation of the key points and feasible processing times: each experiment (one grid point) took 165 minutes to train the *VGG16_transfer_learning* parameters (see Section 4.4 for details) on an NVIDIA GeForce GTX GPU. 1,094 256×256 images for each point (one dataset) in parameter space were generated, of which 547 images are healthy (normal) and 547 augmented with lesion patches (abnormal). The whole parameter space in our low-dimensionality example contains 27 datasets, i.e. $27 \times 1,094 = 29,538$ images. The three dimensions in our parametric space are as follows.

1. *Dimension 1*: size. The size range is sampled at three points: 9×9 , 17×17 , and 25×25 . Briefly, the rationale is that microaneurysms are very small, motivating the 9×9 patches; the largest size (25×25) is approximately 10% of the linear image dimension.
2. *Dimension 2*: number. Experiments indicated that augmenting images with 17, 33, and 50 patches, considering the sampled values of the other parameters, led to a good range of easy to difficult problems.

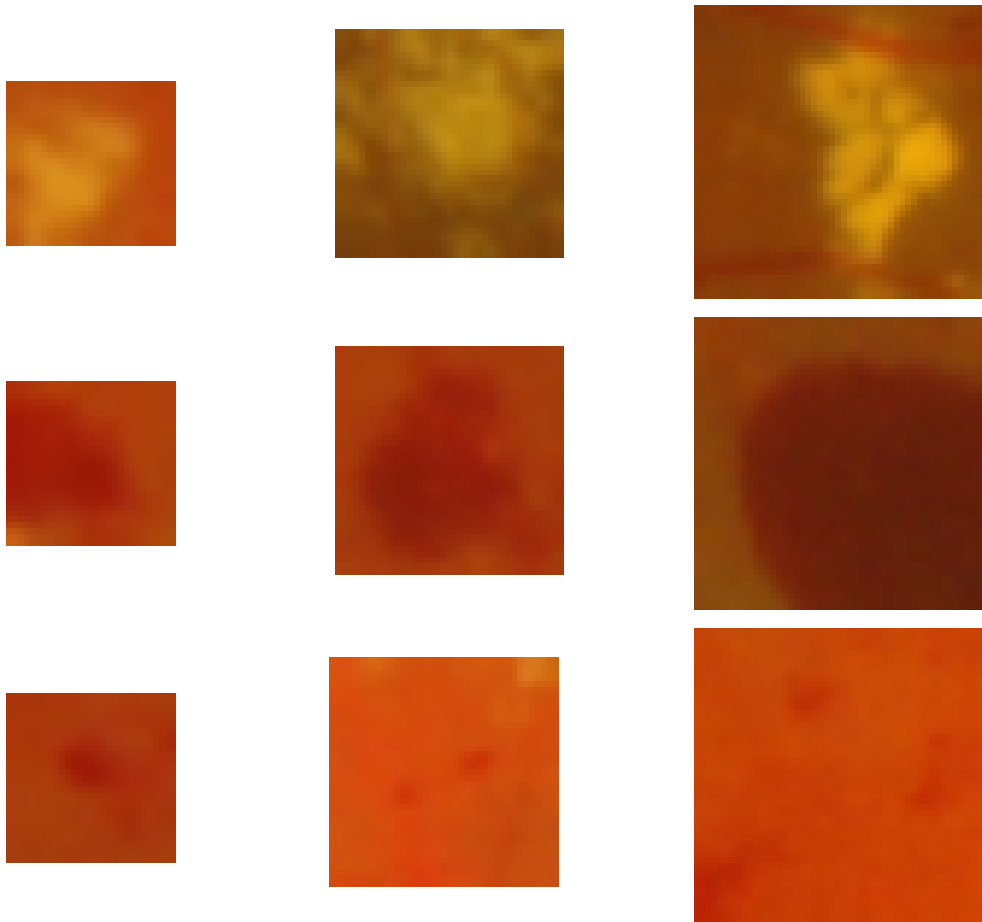


Fig. 4.3 Sample patches for each lesion type, cropped from retinal images graded 1 to 3; Rows (top to bottom): EX, HE, MA; Columns (left to right): patch dimensions - 9×9 , 17×17 , 25×25 .

3. *Dimension 3: transparency.* Transparency models the sharpness of a lesion patch via Equation (4.1), where i, j are pixel coordinates, I_1 is the image to be augmented, I_2 is the image with a lesion patch, I is the augmented image, and α (the transparency level) varies in $[0, 1]$.

$$I(i, j) = \alpha I_1(i, j) + (1 - \alpha) I_2(i, j) \quad (4.1)$$

When $\alpha=0$ (zero transparency), I is the same as I_2 and the lesion patch is blended in with maximum contrast. When α increases, the transparency increases, and the contrast patch-background decreases, making detection more difficult. When $\alpha=1$ (max transparency), I is the same as I_1 , and the lesion is invisible. Useful α values were identified at 0 (to include no visibility), 0.4, and 0.7. The resulting patch I is blended in the original retinal images using Pérez, Gagnat, and Blake's Poisson-guided interpolation [217]. Figure 4.4 shows examples of images from different points (datasets) in the grid.

4.3.4 Classifier

Convolutional Neural Network (CNN) was used as a classifier to demonstrate the framework. The hyper-parameters, such as the number of layers, filters in each layer, weight initialization, fully connected layers/nodes, learning rate, and optimization function, were tuned to improve the classification performance. A modified version of the well-known CNN architectures was used that were proven to perform exceptionally well and achieved state-of-the-art performance when they first participated in the famous computer vision challenge, ImageNet [85]. The CNNs include VGG16 [92], ResNet50 [93], InceptionV3 [218], DenseNet201 [94] and EfficientNet-B2 [98] network.

Inspired by the work of Poplin et al. [23], the CNN model architectures were modified specifically in the fully connected layers by replacing the Fully Connected (FC) layer with a single sigmoid output node to fit our classification task (normal vs abnormal). The top layers such as Convolutional (Conv), pooling, activation, batch normalization, and dropout layers were used as they were proposed by the respective authors. A brief summary, as well as comparisons of these architecture, are provided in Table 4.1. Input images are $256 \times 256 \times 3$ (3 color channels). To expedite the training process and enable the CNN to learn features such as edge enhancement in its initial layers, The standard practice in the deep learning community was followed to utilize pre-trained weights from Imagenet, which were trained on images of size $224 \times 224 \times 3$ (similar to our images). These pre-trained weights were applied to the convolution layers and trained the fully connected layer, which contained a single node

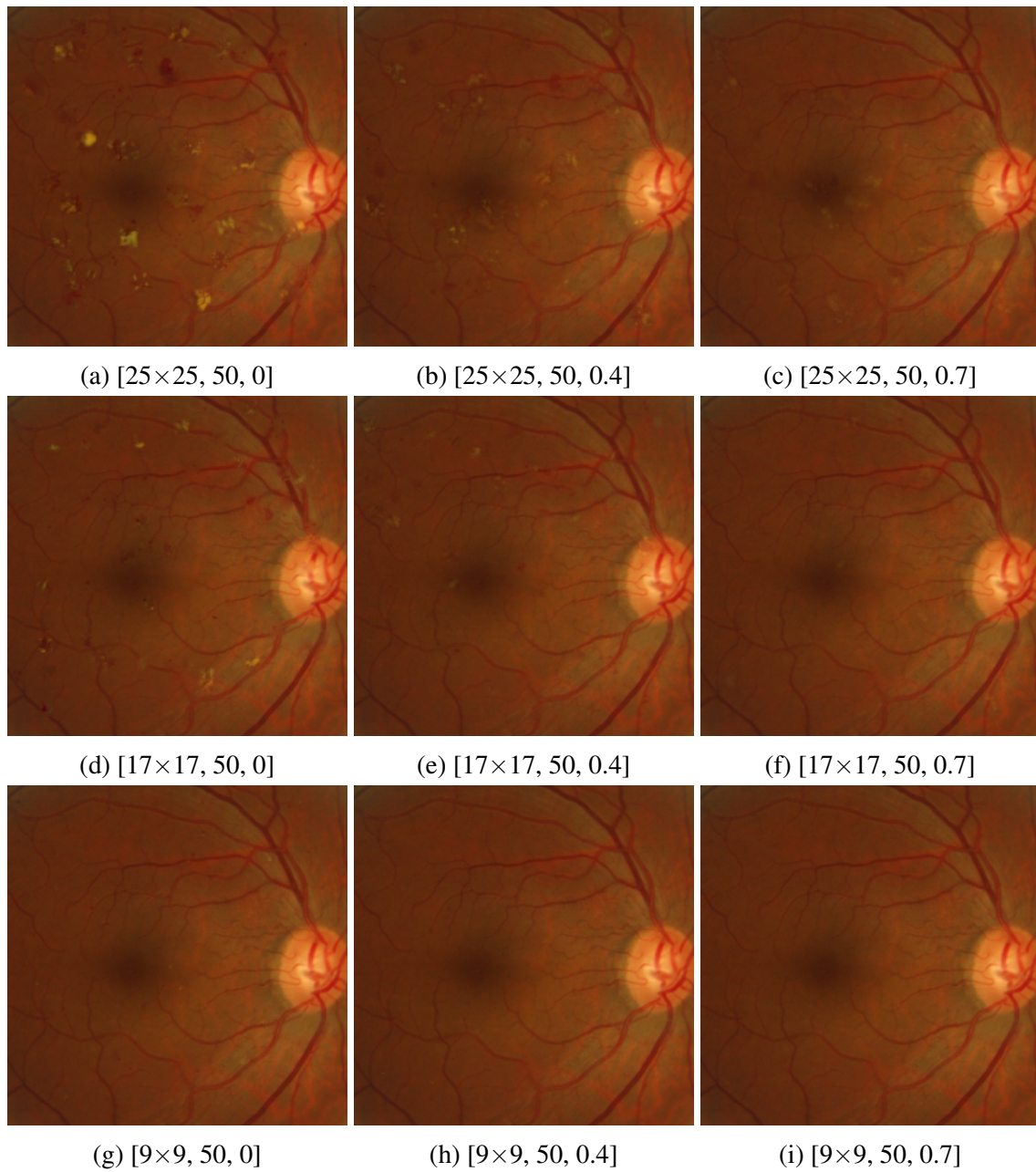


Fig. 4.4 Examples of augmented images from datasets at various points in parameter space, with coordinates [patch size, number of patches, transparency].

Table 4.1 Summary of modified CNN models.

Model	Year	Base unit	Trainable parameters
VGG16	2014	Convolution	14,715,201
ResNet50	2015	Residual block	23,536,641
InceptionV3	2015	Inception module	21,770,401
DenseNet201	2017	Dense block	18,094,849
EfficientNet-B2	2019	Inverted residual block	7,702,403

[92]. A sigmoid layer is followed by the last single node in the fully connected layer. The total number of trainable parameters in these modified CNNs are provided in Table 4.1.

The first experiment used the modified VGG16. The actual VGG16 has 13 Conv layers and 3 fully connected layers; it has been shown to be a simple but powerful model for image classification and detection tasks [92]. Subsequent experiments involved the other CNN models mentioned above.

4.3.5 CNN Training

The Conv layers of all the models were initialized with pre-trained weights from ImageNet. Following [86] and [92], a random uniform initialization was used for the non-pre-trained weights. Each point in the parameter-space grid generates an independent dataset (and a problem of well-characterized difficulty) of 1,094 images (half normal, half abnormal). A 10-fold cross-validation was applied on each point and this provides 984 random images (90% of total images) for model development referred to as *development set* and 110 random images (10% of total images) for testing, referred as *test set* in each fold and this is repeated for all the 10-folds. The *development set* was further randomly partitioned to 90% as *train set* and 10% as *validation set*.

The hyper-parameter tuning process resulted in the selection of the following parameters for further experimentation: a batch size of 16 with retinal images as input and corresponding normal/abnormal status as output label, binary cross-entropy as the loss function, Adam optimization and Nesterov Accelerated Gradient momentum (Nadam) with an initial learning rate of 0.001, reduced by a factor of 0.1 if the validation loss did not improve within 10

consecutive epochs (minimum learning rate 10^{-5}). All models were trained for 50 epochs as the learning curves indicated that the model training quickly saturated due to the initialization of the model parameters with pre-trained weights from ImageNet. Additionally, to avoid overfitting, training was stopped if there was no improvement in the validation loss for 20 epochs, and the weights with the best validation performance were saved.

To illustrate the framework, only accuracy is used here, defined as the percentage of correctly classified images to the total number of images. For each data point (dataset), the mean accuracy of *test set* is recorded from the 10-fold cross-validation schedule. The training specifications are summarized in Table 4.2.

Table 4.2 Summary of training specifications on MESSIDOR data.

Category	Specification
Input and Output	
Input	Color fundus image
Input dimensions	256×256
Output	Binary classification (Sigmoid activation)
Performance metric	Mean test accuracy (10-fold CV)
Training	
Weight Initialization	ImageNet and random uniform
Epochs	50
Batch size	16
Loss function	Binary cross-entropy
Optimizer	Nadam
Learning rate	0.001 reduced by a factor of 0.1
Avoid overfitting	
Early stopping	on validation loss
Weights	Best validation loss

All training was done on a single NVIDIA GeForce GTX GPU, using Python 3.6 for code development and Keras 2.2.2 [219] with TensorFlow 1.9.0 as the back-end for training and validating DL models.

4.4 Results

The modified VGG16 version was trained in two steps. First, freezing the ImageNet pre-trained parameters of the Conv modules and training only the parameters of the fully-connected layers, referred to as *VGG16_partially_trained*. Next, training all network parameters by initializing the Conv module parameters with ImageNet pre-trained parameters, referred to as *VGG16_transfer_learning*. Figures 4.5 and 4.6 visualize the parameter space defining all datasets and color-code model accuracy at each point (*VGG16_partially_trained* and *VGG16_transfer_learning* respectively). Difficulty levels are clearly organized in the space; the most difficult problem (the fewest and smallest patches with high transparency) is situated, as expected, opposite the easiest one (the most numerous and largest patches with zero transparency).

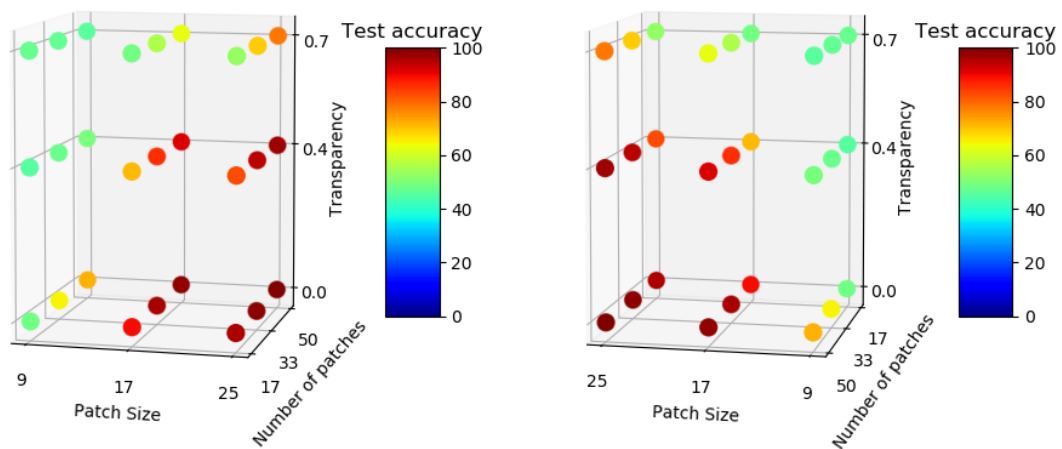


Fig. 4.5 Visualization of the performance of *VGG16_partially_trained* in 3-D parameter space, using color coding and two viewpoints for clarity (180° rotation around a vertical axis through the center of the ground plane). The region of the parameter space challenging the classifier is clearly visible.

The improvement in the performance is clearly visible in Figure 4.6 when compared with Figure 4.5. hence a similar training strategy was followed with other CNN models, i.e., initializing the Conv layers parameters with ImageNet pre-trained parameters.

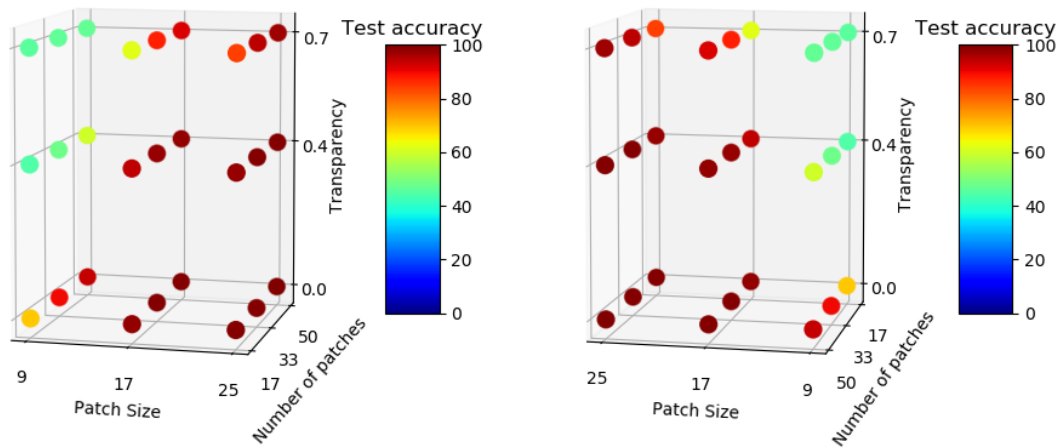


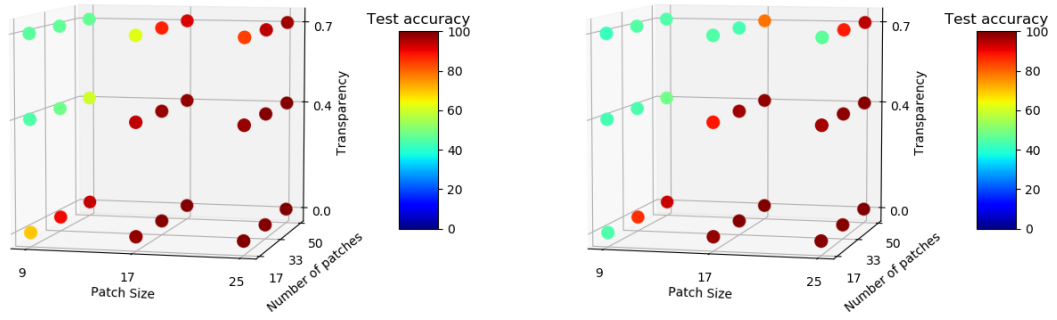
Fig. 4.6 Visualization of the performance of *VGG16_transfer_learning* in 3-D parameter space. Color coding and viewpoints as per the previous figure. The improvement with respect to Figure 4.5 is immediately visible.

Figure 4.7 shows the performance visualization of all modified CNN models namely VGG16, ResNet50, InceptionV3, DenseNet201 and EfficientNet-B2 in the 3D parameter space. From Figure 4.7, it can be clearly seen that all the models perform very well in the easiest classification problem (bottom right side of the cube) and struggle to learn the features as the difficulty level gradually increases (top left side of the cube). In the overall comparison among all the models, EfficientNet-B2 outperforms the other models.

Assuming the values sampled for each parameter are representative of real data, the grid suggests which parameters most influence the difficulty of the task, for a given classifier. This allows one to intervene specifically, e.g., change specific hyper-parameters, or design annotations for real-data experiments.

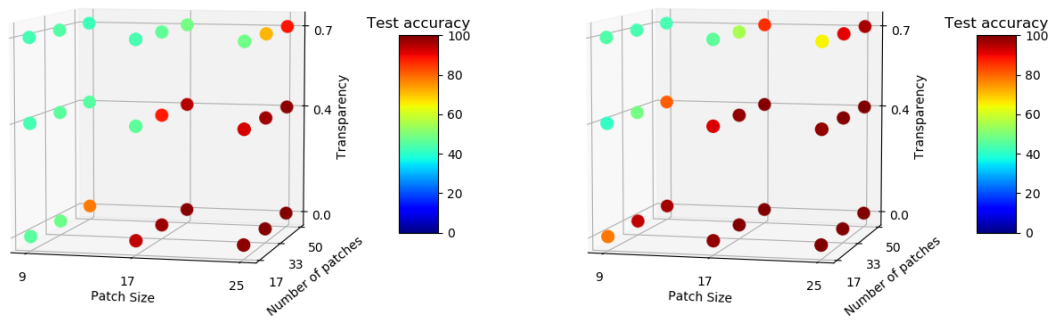
4.5 Validating CNN models on IDRID

As a further check, all CNN models were trained as described in Section 4.3.4, on the real images from the Indian Diabetic Retinopathy Image Dataset (IDRID) dataset. Here, the task is a 5-class classification problem, where each class represents grades of DR, described



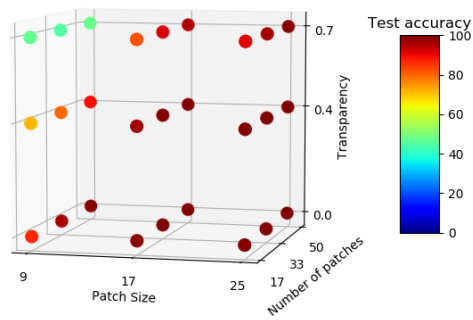
(a) VGG16

(b) ResNet50



(c) InceptionV3

(d) DenseNet201



(e) EfficientNet-B2

Fig. 4.7 Comparative performance visualization of all modified CNNs in 3-D parameter space.

in Section 3.2.2. This required some changes to the CNN architectures, e.g., changing the number of nodes from 1 to 5 in the output layer and the whole retinal image was considered (no black corners eliminated) for training, validation, and testing. The pre-processing on IDRID is described in the next section.

4.5.1 Pre-processing

The IDRID images have significant variations in terms of luminosity, color, focus, and general quality. To reduce the effect of these variations on the DL training, the images were pre-processed automatically using the cv2 [198] package of python.

The extraction of the circular retina is similar to what has been applied to Genetics of Diabetes Audit and Research in Tayside Scotland (GoDARTS) images (Section 3.3.2). In brief, a bounding rectangle was inscribed on the retina to crop the circular retina. Following [220], the color image was converted into YCrCb color space and Contrast Limited Adaptive Histogram Equalization (CLAHE) was applied in Y channel and the resultant image is converted back to RGB color space. Finally, the images were resized to 256×256 and the pixel intensities were normalized to $[0,1]$. Figure 4.8 shows a sample image with the pre-processing.

4.5.2 CNN Training

IDRID consists of 413 retinal images as a training set and 103 retinal images as the testing set, split from the total 516 images. DR grades were provided on a 5-point scale: 0-No DR, 1-4 different stages of DR (Table 3.2), leading to a 5-class classification problem. The training set was further randomly split to 80% as the train set (329 images) and 20% as the validation set (84 images). Table 4.3 shows the details of the splits.

A similar protocol as used in the previous experiment (Section 4.3.5) was followed with a few modifications, e.g., adapting the algorithm on a 5-class (not binary) classification problem: hence the last single node output layer was replaced with 5 nodes. The sigmoid and binary cross entropy were also replaced with softmax as activation and categorical cross

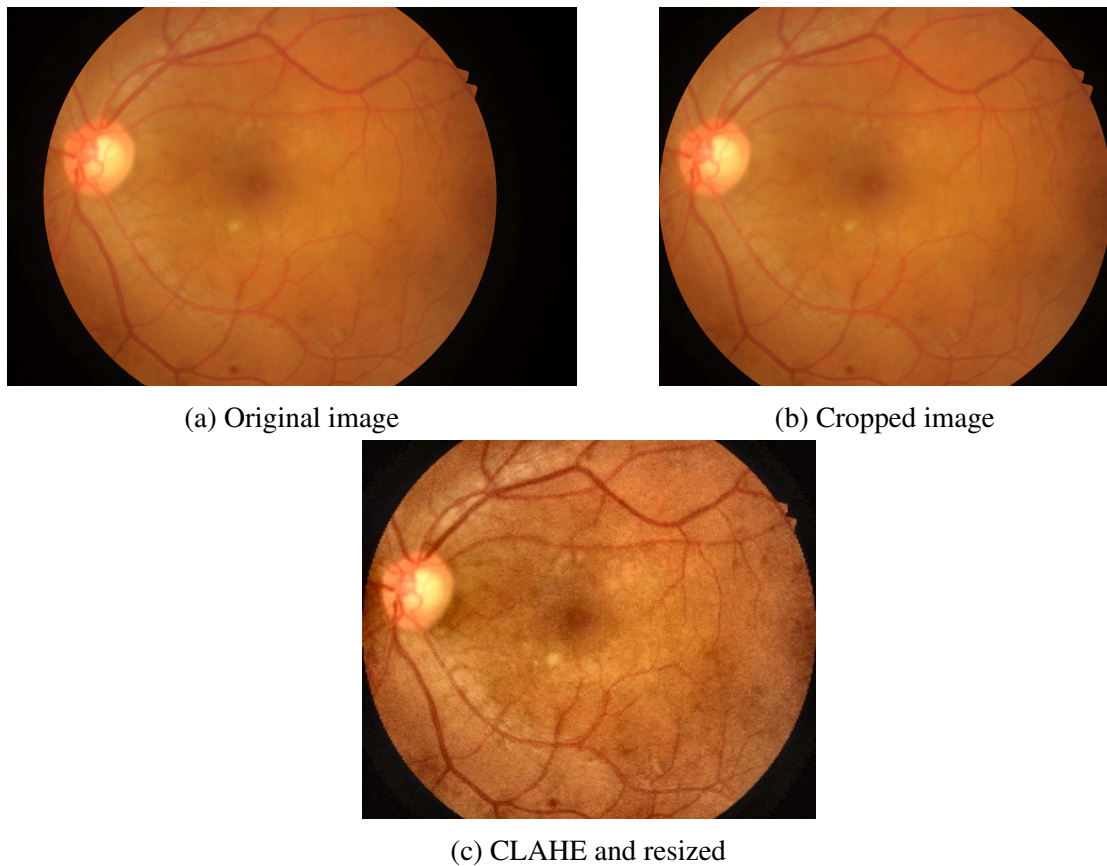


Fig. 4.8 Example of image pre-processing in IDRID.

entropy as loss function. The model was trained for 200 epochs with a batch size of 16. Accuracy was considered as the performance metric to validate model performance on test data. The rest of the protocol was the same as for the cube experiment. Key values are summarized in Table 4.4.

4.5.3 Results

Five modified CNN models were trained to classify a 5-class DR grade from retinal images. The accuracy performance of these models on test data is shown in Table 4.5. The EfficientNet-B2 model gave comparative best accuracy of 65.05% on the test data followed by VGG16 with an accuracy of 62.14% and the remaining three CNN models gave an accuracy less than or equal to 60%.

Table 4.3 DR class distribution in the IDRID data splits.

DR grade	Count (proportion, %)		
	Train	Validation	Test
0	107 (32.52)	27 (32.14)	34 (33.01)
1	16 (4.86)	4 (4.76)	5 (4.85)
2	108 (32.83)	28 (33.33)	32 (31.07)
3	59 (17.93)	15 (17.86)	19 (18.45)
4	39 (11.85)	10 (11.9)	13 (12.62)
Total	329	84	103

Table 4.4 Summary of training specifications on IDRID data.

Category	Specification
Input and Output	
Input	Color fundus image
Input dimensions	256×256
Output	5-class classification (Softmax activation)
Performance metric	Test accuracy
Training	
Weight Initialization	ImageNet and random uniform
Epochs	200
Batch size	16
Loss function	Categorical cross-entropy
Optimizer	Nadam
Learning rate	0.001 reduced by a factor of 0.1
Avoid overfitting	
Early stopping	on validation loss
Weights	Best validation loss

For comparison, the three top performers on DR classification in the IDRID leaderboard [221, 13] are, at the time of experiments, programs by teams called LzyUNCC, VRT, and Mammoth. They achieved accuracies of 74.76%, 59.22%, and 54.37%, respectively, in the onsite challenge [13]. These teams used external retinal image datasets from the Kaggle DR challenge [191] for pre-training their respective models. LzyUNCC and VRT pooled the results from 5 and 10 different models respectively to obtain the final prediction. The input

image sizes used were 896×896 , 640×640 , 512×512 for LzyUNCC, VRT and Mammoth respectively. Therefore it was concluded that the solution provided here is competitive given the state-of-the-art on images from the IDRID challenge.

Model	Test Accuracy (%)
VGG16	62.14
ResNet50	46.6
InceptionV3	60.19
DenseNet201	56.31
EfficientNet-B2	65.05

Table 4.5 Performance of modified CNNs on IDRID test data.

Recent literature [222, 164, 94, 93, 223, 90] and EfficientNet authors from their experiments show that the DL network accuracy improves with scaling up any of the network dimensions (width, depth, or resolution) but the accuracy gain reduces for bigger DL models. In brief, depth is a network parameter for scaling the network depth to go deeper; width is again a network parameter for scaling the network width to capture more fine-grained features; the resolution is the input image parameter for image width and height to capture fine-grained patterns. The superior performance of EfficientNet-B2 [98] compared to other DL architectures might be due to their new compound model scaling approach in the dimensions of depth, width, and resolution, where all the dimensions are uniformly scaled using compound coefficients as described in [98].

4.6 Discussions

This chapter has presented a framework for generating synthetic datasets organized systematically in multi-dimensional parameter space by levels of difficulty for a target problem. The main aim was to support the process of tuning parameters and hyper-parameters of a medical image analysis classifier to achieve a robust model.

The focus of this work was not the generation of realistic retinal images containing lesions, as e.g. is in [212, 213]. Instead, the focus was on defining a parameter space providing a

quantitative characterization of difficulty. Synthetic images were built by augmenting images of real, healthy retinas with patches containing lesions, sampled from a dictionary of patches cropped from images of real, diseased retinas. More realistic models of difficulty should arguably consider the position and co-occurrence of lesions in populations of real patients.

A simple 3-D (27 grid points in all) was chosen to enable visualization of the parameter space. More complete models of difficulty (more parameters) will be higher-dimensional preventing visualization, but the systematic organization of the datasets based on a quantitative definition of difficulty will not be affected. While the dimension of the parameter space depends on the representation of difficulty for a target problem and image domain, the size of the grid (number of points, related to the sampling frequencies along each axis) depends ultimately on the computational resources available vis-à-vis the number of parameters to train.

To use multiple performance criteria, one would replace the information at each point in parameter space, now a single number (accuracy), with a vector of numbers (values of multiple criteria). The spatial ordering of points by difficulty would then require an ordering over a discrete vector field.

The initial experiments of solving the binary classification task of normal vs abnormal retinal fundus images in the proposed 3-D cube started with a 2-layered CNN architecture, which was extensively tuned for its hyper-parameters. However, this CNN showed excellent performance only for the easy problem (the bottom layer of the cube) and failed to perform at the medium-difficulty problem. As a result, in the rest of our work, we used existing state-of-the-art CNN architectures such as VGG16, ResNet50, InceptionV3, DenseNet201, and EfficientNet-B2, which were known to perform well in the ImageNet challenge.

4.7 Conclusions

This experiment was done to identify an optimal DL algorithm under certain conditions, not because the real data was unaccessable. There was some delay in obtaining access to real retinal images from GoDARTS as linked patient data needed moving to a secured

Safe Haven (SH) environment of the Health Informatics Center (HIC). This secured SH environment required a brand new setup of GPU resources to carry out computationally heavy DL algorithms. Meanwhile, open-source retinal image data was used from Methods to Evaluate Segmentation and Indexing Techniques in the field of Retinal Ophthalmology (MESSIDOR) for our initial experiments.

A low-dimensionality example (the "cube experiment") was provided to illustrate the framework, within the evaluation of a few CNN models to classify normal vs. abnormal retinal fundus camera images. The proposed framework is general and has applications beyond retinal image analysis.

The comparative performance evaluation of multiple CNN models tried in the cube experiment indicates that EfficientNet-B2 [98] provides better performance than other models. These results were validated with independent real images in a 5-class problem of the IDRID challenge dataset.

The EfficientNet-B2 architecture has been therefore adopted in the subsequent work reported in this thesis, including classification and regression tasks with real retinal image data from GoDARTS. In Chapter 5, The retinal images were investigated for predicting demographic and clinical features using EfficientNet-B2 and Chapter 6 presents the results of predicting Cardiovascular (CV) risk scores and its associations with Cardiovascular Disease (CVD) from retinal images using EfficientNet-B2.

Chapter 5

Predicting demographic and clinical features

5.1 About this chapter

The Genetics of Diabetes Audit and Research in Tayside Scotland (GoDARTS) bio-resources, described in Section 3.3, were used to train the Deep Learning (DL) models for predicting demographic features, clinical measurements, and disease outcomes from the retinal images ^{1 2 3}. EfficientNet-B2, the best-performing model from the cube experiment, refer to Chapter 4, was used for training and testing. All the available retinal images in GoDARTS were used for training and testing the model for predicting demographic features. For clinical measurements and disease outcomes, the retinal images at baseline (first retina available

¹Syed, M. G., Wang, H., Trucco, E., Huang, Y., Mordi, I., and Doney, A. *Biological Vascular Age from Retinal Photographs Predicts All-Cause Death and Cardiovascular Events: a GoDARTS study*. This manuscript is currently being prepared for submission to a journal and is based on the research described in Section 5.2 and 5.3.

²Soca, A., Syed, M. G., Trucco, E., Harvey, J., and Doney, A. *Prediction of dementia outcome from retinal and genomic data in the GoDARTS cohort*. This manuscript was accepted at AD/PD 2023, International Conference on Alzheimer's and Parkinson's Diseases and related neurological disorders and is currently being prepared for submission to the conference and is based on the research described in Section 5.2 and 5.3.

³Huang, Y., Syed, M. G., et al., *Genomic Determinants of Biological Age Estimated By Deep Learning Applied to Retinal Images*. This manuscript is currently being prepared for submission to a journal and is based on the research described in Section 5.2 and 5.3.

date) were used. This chapter describes the materials, methods, and results of the DL model for the prediction of demographic and clinical features.

5.2 Demographic features

This section describes the prediction of demographic features (age and sex) from retinal images in GoDARTS using the Convolutional Neural Network (CNN). All the 102,455 retinal images available in GoDARTS were used for this experiment.

5.2.1 Materials

Dataset

As mentioned in Section 3.3.1, there are 102,455 retinal images in total from 8,594 individuals obtained at multiple time points. The date of birth and sex information for 24 individuals were not available, so these individuals were excluded, resulting in 102,082 images from 8,570 individuals. The descriptive characteristics for age and sex in the whole GoDARTS at baseline (at first retinal image) are provided in Table 3.3 and Table 3.4 of Section 3.3.1. The descriptive characteristics for the data split (more details in Section 5.2.2) are shown in Table 5.1 and Table 5.2 shows the data distribution in different age sub-groups.

Outcome variables: age and sex

The sex information is directly extracted from the demography database in GoDARTS bio-resources. The age at retinal imaging is computed using the date of birth and date of retinal imaging information. Refer to Section 3.3.1 for more details on feature extraction.

5.2.2 Methods

Image pre-processing

The image pre-processing steps applied on the GoDARTS retinal images are described in Section 3.3.2.

Table 5.1 Dataset characteristics of whole cohort and data splits used for predicting demographic features.

	Overall	Train	Validation	Test
Participants	8,570	5,999	849	1,722
Images	102,082	71,434	9,954	20,694
Right eye Images (%)	50,719 (49.68)	35,476 (49.66)	4,954 (49.77)	10,289 (49.72)
Sex - Count (proportion in %)				
Male	4,819 (56.23)	3,398 (56.64)	471 (55.48)	950 (55.17)
Female	3,751 (43.77)	2,601 (43.36)	378 (44.52)	772 (44.83)
Age at imaging (in years) – Mean (std, IQR)				
All	66.11 (11.77, 58.8-74.71)	66.09 (11.71, 58.76-74.68)	66.67 (11.65, 59.58-75.04)	65.89 (12.0, 58.51-74.71)
Male	65.73 (11.4, 58.53-74.05)	65.62 (11.37, 58.33-74.04)	67.14 (11.0, 60.61-75.08)	65.42 (11.65, 58.14-73.72)
Female	66.59 (12.21, 59.09-75.38)	66.7 (12.13, 59.42-75.44)	66.07 (12.39, 57.73-75.02)	66.48 (12.4, 58.95-75.4)

Deep learning architecture and training

The EfficientNet-B2 DL model, which was the top-performing model in the cube experiment (as described in Chapter 4) was utilized. To achieve optimal performance based on the compound scaling mechanism, an input image size of 260x260 were used, as recommended by the authors [98]. Two separate models were trained, one for predicting age at retinal imaging and the other for predicting sex. The fully connected layer of EfficientNet-B2 was replaced with a global average pooling layer, followed by a single output node with linear activation for age prediction and sigmoid activation for sex prediction. The two models had a total of approximately 7.7 million trainable parameters each.

The training procedures for both models were similar in terms of the number of epochs, batch size, weight initialization, image augmentation, and measures to prevent overfitting. The pre-trained weights from ImageNet were used to initialize the model, and the dataset was split randomly into 70% for training, 10% for validation, and 20% for testing. The characteristics of the dataset at baseline for all splits are provided in Table 5.1. During the

Table 5.2 Data distribution for age subgroups in the whole cohort and data splits. In parenthesis, the value includes the proportion of data in the respective column.

Age group	Overall		Train		Validation		Test	
	# of individuals	# of Im-ages	# of individuals	# of Im-ages	# of individuals	# of Im-ages	# of individuals	# of Im-ages
0-10	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)
10-20	1 (0.01%)	6 (0.01%)	1 (0.02%)	6 (0.01%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)
20-30	27 (0.32%)	180 (0.18%)	14 (0.23%)	95 (0.13%)	6 (0.71%)	31 (0.31%)	7 (0.41%)	54 (0.26%)
30-40	180 (2.1%)	1,411 (1.38%)	120 (2.0%)	938 (1.31%)	13 (1.53%)	111 (1.12%)	47 (2.73%)	362 (1.75%)
40-50	641 (7.48%)	5,892 (5.77%)	454 (7.57%)	4,134 (5.79%)	52 (6.12%)	433 (4.35%)	135 (7.84%)	1,325 (6.4%)
50-60	1,562 (18.23%)	16,312 (15.98%)	1,108 (18.47%)	11,579 (16.21%)	148 (17.43%)	1,598 (16.05%)	306 (17.77%)	3,135 (15.15%)
60-70	2,623 (30.61%)	31,932 (31.28%)	1,848 (30.81%)	22,724 (31.81%)	259 (30.51%)	2,911 (29.24%)	516 (29.97%)	6,297 (30.43%)
70-80	2,630 (30.69%)	33,251 (32.57%)	1,808 (30.14%)	22,755 (31.85%)	278 (32.74%)	3,573 (35.9%)	544 (31.59%)	6,923 (33.45%)
80-90	858 (10.01%)	12,313 (12.06%)	613 (10.22%)	8,704 (12.18%)	86 (10.13%)	1,199 (12.05%)	159 (9.23%)	2,410 (11.65%)
90-100	48 (0.56%)	785 (0.77%)	33 (0.55%)	499 (0.7%)	7 (0.82%)	98 (0.98%)	8 (0.46%)	188 (0.91%)

data split, care was taken to ensure that retinal images of the same individual were not present in different splits to avoid information leaks during training. The model was fine-tuned by training all the parameters for a total of 50 epochs with a batch size of 32, using retinal images as input and corresponding age at imaging and sex as output labels. Simple image augmentation methods, such as horizontal flip and rotation, were randomly applied to each image during training.

The DL models for age and sex prediction used mean squared error loss and binary cross-entropy loss, respectively. Adam optimization and Nesterov Accelerated Gradient momentum were applied with an initial learning rate of 0.001, which was reduced by a factor of 0.1 if the validation loss did not improve within 5 consecutive epochs, with a minimum learning rate of 10^{-5} . To avoid overfitting, training was stopped if there was no improvement in the validation loss for 20 epochs. The weights leading to the best validation performance were saved. A summary of the training specifications can be found in Table 5.3, and the performance evaluation metrics are described in Section 5.2.2, Evaluation metrics.

Table 5.3 Summary of training specifications for predicting age and sex using all retinal images in GoDARTS.

Category	Specification
Input and Output	
DL architecture	EfficientNet-B2
Input	Color fundus image
Input dimensions	260 x 260
Output	
<i>for age</i>	Linear activation
<i>for sex</i>	Sigmoid activation
Performance metric	
<i>for age</i>	Mean Absolute Error (MAE), R^2
<i>for sex</i>	Area Under Receiver Operating Characteristic (ROC) Curve (AUC), accuracy, sensitivity and specificity
Training	
Weight Initialization	ImageNet and random uniform
Epochs	50
Batch size	32
Loss function	
<i>for age</i>	Mean squared error
<i>for sex</i>	Binary cross-entropy
Optimizer	Nadam
Learning rate	0.001 reduced by a factor of 0.1
Avoid overfitting	
Early stopping	on validation loss
Weights	Best validation loss

The experiments were conducted on the NVIDIA TITAN Xp GPU provided by the Safe Haven (SH) environment offered by Health Informatics Center (HIC) services, adhering to the guidelines of the University of Dundee, UK [25]. The DL model was developed using Python 3.6 and Keras 2.2.2 [219], with TensorFlow 1.9.0 as the backend for training and testing.

Evaluation metrics

In light of recent publications on deep learning research related to retinal biomarkers [23, 158, 155, 162], the MAE (Equation (5.1)), and the coefficient of determination (R^2 , Equation (5.2)), were used as the evaluation metric for age prediction as it is a real number. The best possible metric values are 0 for MAE and 1 for R^2 . Ideally, R^2 value ranges from 0 to 1 but it can sometimes give negative values if the model's performance is worse than the mean prediction. Below, y_i is the true value, and \hat{y}_i is the predicted value of the i th sample; \bar{y} is the mean of true values.

$$MAE = \left(\frac{1}{n}\right) \sum_{i=1}^n |y_i - \hat{y}_i| \quad (5.1)$$

$$R^2 = 1 - \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2} \quad (5.2)$$

For sex prediction accuracy, sensitivity, specificity and AUC were computed as evaluation metrics. These metrics are computed using the predicted probability scores from the model and actual labels. The definitions for accuracy, sensitivity, and specificity are given in Equation 5.3, 5.4, and 5.5 respectively. Their components are defined in Table 5.4.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (5.3)$$

Table 5.4 Confusion matrix terminology.

		Actual	
		Positive	Negative
Predicted	Positive	True Positive (TP)	False Positive (FP)
	Negative	False Negative (FN)	True Negative (TN)

$$\text{Sensitivity}(or)TPR = \frac{TP}{TP + FN} \quad (5.4)$$

$$\text{Specificity}(or)TNR = \frac{TN}{TN + FP} \quad (5.5)$$

The ROC is a plot of the True Positive Rate (TPR) against the False Positive Rate (FPR) (or 1-True Negative Rate (TNR)), generally, obtained by varying the threshold on the predicted probability scores. The AUC is the measure of the classifier's ability to distinguish between the classes. Its value ranges from 0 to 1, with 1 corresponding to a classifier able to perfectly distinguish between positive and negative classes correctly, and 0 to a classifier predicting all positives as negatives and vice versa. The AUC value of 0.5 indicates that the classifier predicts at random.

Statistical significance

Deep learning algorithms do not generate statistical significance directly. Therefore, non-parametric bootstrap sampling was employed to estimate the statistical significance of the model's performance on the test data. 2,000 randomly sampled with replacements were used from the test data, with each sample size being the same as that of the test data. The performance metrics were then calculated for each bootstrap sample. The 95% confidence interval was determined from the distributions of the performance metrics as the range between the 2.5 and 97.5 percentile points, following the approach used in [23].

Activation visualization

The technique of Gradient-based Class Activation Mapping (Grad-CAM) [99] was employed to visually identify the regions in input images that contain important information for the classifier. Grad-CAM achieves this by using the gradient of the loss function with respect to the feature maps in intermediate layers (such as convolutional layers) as weights. These weighted feature maps in the layers of interest are then averaged and upsampled to the original input size to visualize the critical regions.

The last convolutional layer of EfficientNet-B2 was selected for applying Grad-CAM. The feature maps in this layer have spatial dimensions of $9 \times 9 \times 1048$, where 1048 represents the number of channels. The feature maps were weighted using the gradient of the loss function and rescaled to the input image dimensions (260×260) to generate heatmaps, as shown in Figure 5.2. However, the substantial upscaling may introduce errors such as masking out important regions or creating false positives.

5.2.3 Results

A total of 71,434 retinal images were used for training, 9,954 images for validation, and 20,694 for testing. Two separate models were trained, validated, and tested for predicting age and sex.

Age prediction using deep learning

The performance results on complete test data for age predictions are shown in Table 5.5. The model achieved MAE of 3.951 (95% CI 3.908, 3.995) years and R^2 of 0.809 (0.804, 0.814) on the complete test data. MAE for all the left eye retinal image predictions in the test data is 3.91 (3.849, 3.969) years and R^2 is 0.813 (0.806, 0.82). For all the right eye predictions in the test data, MAE is 3.944 (3.933, 4.054) years and R^2 is 0.804 (0.797, 0.812). From Table 5.5, the model performance appears to be consistent while predicting age using the left retina and right retina. Figure 5.1 shows a scatter plot for predicted age and actual age in the whole test data, only left eye images and only right eye images in the test data.

Table 5.5 Model's performance for predicting age on complete test data. 95% CI values are computed using bootstrap samples.

Category	# of Images	MAE (95% CI) years	R^2 (95% CI)
Complete test data	20,694	3.951 (3.908, 3.995)	0.809 (0.804, 0.814)
Only left eye in test data	10,405	3.91 (3.849, 3.969)	0.813 (0.806, 0.82)
Only right eye in test data	10,289	3.944 (3.933, 4.054)	0.804 (0.797, 0.812)

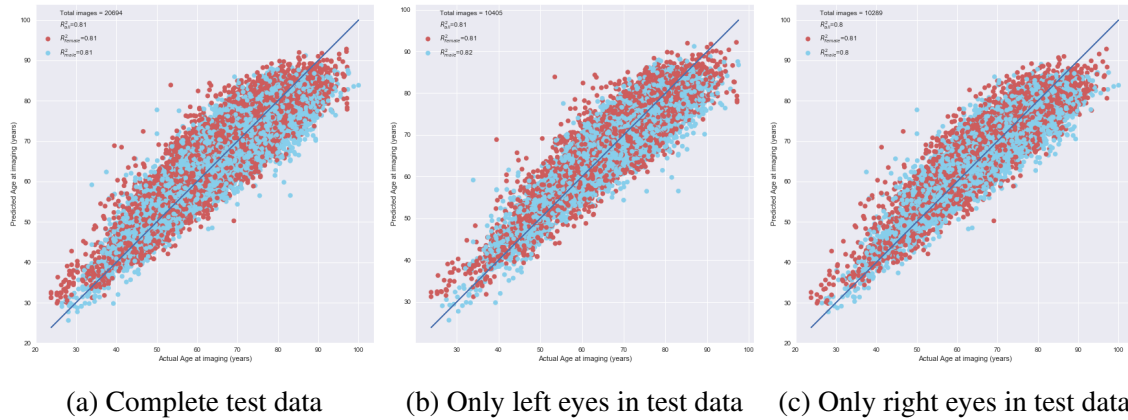


Fig. 5.1 Scatter plot for actual age at imaging and predicted age at imaging in the complete test data and its subset.

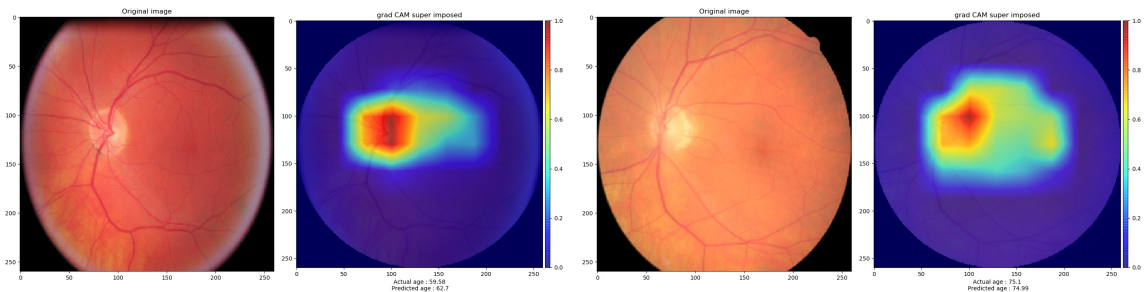
The MAE for age sub-groups with an interval of 10 years in the test data is shown in Table 5.6. The MAE is 3.5 years in the age group of 60-80 years (13,220 images) and the MAE is slightly more (~ 4 years) in the age group from 30-60 years (4,822 images). The MAE increases in the older age group, >90 years, arguably because of the decreasing number of people in the groups and similar to the youngest groups (20-30). The age sub-group wise performance considering only the left eye retina or only the right eye retina in the test data is provided in the Appendix A, Table A.1 and A.2.

Heatmaps were generated using Grad-CAM for all retinal images in the test data from the age prediction model. The resulting Grad-CAM heatmaps, which indicate the important features in the input image for making the age prediction, were analyzed. Figure 5.2 shows two sample retinal images with their corresponding Grad-CAM heatmaps. Each sample heatmap shows the original image to the left and Grad-CAM heatmap superimposed on the original image on the right. The visual inspection across several Grad-CAM heatmaps

Table 5.6 Age sub-group-wise model performance with mean actual age, mean predicted age, MAE, and its 95% CI for the age sub-groups with an interval of 10 years on the complete test data.

Age group	# of Images	Mean actual age	Mean predicted age	MAE (95% CI)
0-10	0	-	-	-
10-20	0	-	-	-
20-30	54	27.457	32.899	5.567 (4.741, 6.372)
30-40	362	36.638	40.372	4.455 (4.03, 4.883)
40-50	1,325	45.883	48.984	4.367 (4.168, 4.572)
50-60	3,135	55.679	58.567	4.152 (4.037, 4.268)
60-70	6,297	65.158	66.224	3.566 (3.494, 3.639)
70-80	6,923	74.801	73.652	3.576 (3.511, 3.638)
80-90	2,410	83.317	78.897	5.048 (4.903, 5.194)
90-100	188	92.432	83.575	8.867 (8.364, 9.375)

indicates that Optic Disc (OD) and macula are the most important features for predicting age at imaging in all age groups; additionally, the retinal vasculature is also important in younger and middle-age groups.



(a) Male individual, actual age 59.58, predicted age 62.7

(b) Female individual, actual age 75.1, predicted age 74.99

Fig. 5.2 Sample grad-CAM heatmaps from model trained for age prediction.

Sex prediction using deep learning

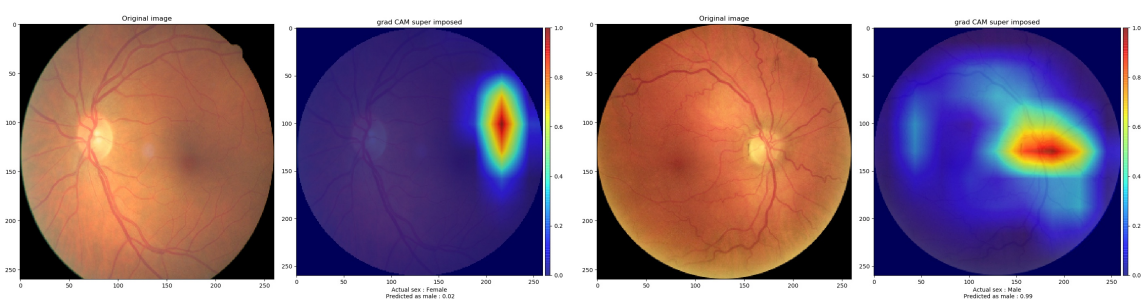
The performance results on complete test data for sex predictions are shown in Table 5.7. The model achieved AUC of 0.899 (0.895, 0.903), the accuracy of 0.811 (0.806, 0.817), the sensitivity of 0.886 (0.88, 0.891), and specificity of 0.717 (0.708, 0.727) on the complete

test data. For all the left eye retinal image predictions in the test data AUC of 0.897 (0.891, 0.903), the accuracy of 0.811 (0.803, 0.819), the sensitivity of 0.886 (0.878, 0.894) and specificity of 0.716 (0.703, 0.73) were achieved. For all the right eye predictions in the test data, AUC of 0.902 (0.896, 0.908), the accuracy of 0.811 (0.804, 0.819), the sensitivity of 0.885 (0.877, 0.894) and specificity of 0.717 (0.704, 0.731) was achieved. The model performance appears to be consistent for sex prediction using images from only the left-eye retina and only the right-eye retina.

Table 5.7 Model’s performance for predicting sex using all test data images. 95% CI values are computed using bootstrap samples.

Category	# of Images	AUC (95% CI)	Accuracy (95% CI)	Sensitivity (95% CI)	Specificity (95% CI)
Complete test data	20,694	0.899 (0.895, 0.903)	0.811 (0.806, 0.817)	0.886 (0.88, 0.891)	0.717 (0.708, 0.727)
Only left eyes in the test data	10,405	0.897 (0.891, 0.903)	0.811 (0.803, 0.819)	0.886 (0.878, 0.894)	0.716 (0.703, 0.73)
Only right eyes in the test data	10,289	0.902 (0.896, 0.908)	0.811 (0.804, 0.819)	0.885 (0.877, 0.894)	0.717 (0.704, 0.731)

The Grad-CAM method was used to generate activation heatmaps for all the input retinal images in the test dataset for predicting sex. Figure 5.3 displays two sample retinal images accompanied by their corresponding Grad-CAM heatmaps, highlighting the important regions in the input image for determining sex. From the visual inspection across several Grad-CAM heatmaps, it appears that OD and the region around the macula are the important features for the sex prediction.



(a) Actual sex female, predicted sex female

(b) Actual sex male, predicted sex male

Fig. 5.3 Sample grad-CAM heatmaps from model trained for predicting sex (correct predictions).

Grad-CAM heatmap consistency

Interestingly, it was noticed from the Grad-CAM heatmaps that in the majority of male predictions by the DL model, the OD region is activated; whereas for the majority of female predictions the temporal vascular arcade region is activated. A systematic analysis was performed to check the consistency of the heatmap activations in both the male and female predictions made by the model.

In this analysis, the non-normalized heatmaps were considered for all the correct predictions (actual label and predicted label both are the same) in the test data. Non-normalized heatmaps provide a comparison between raw values, hence a direct one.

In fundus images approximately centered on the macula, or halfway between the macula and the OD, as in GoDARTS, the OD region appears on the left side for left-eye images and on the right side for right-eye ones. The positional difference of OD in both eyes can be clearly seen in Figure 3.1. In the interest of a consistent comparison, all the Grad-CAM heatmaps are grouped based on right/left eye and male/female resulting in four categories. Table 5.8 summarizes the number of images that fall under these four categories.

Pixel-wise means and standard deviations are computed for all the heatmaps in the respective categories and shown in Figure 5.4 and 5.5. The difference is that Figure 5.4 shows the heatmaps using a common color range and Figure 5.5 shows the heatmaps using individual color ranges for the categories, to see activation levels which are not clearly visible using the global range.

Table 5.8 Number of images in sub-groups.

Category	Images count
Male right eye	5,083
Male left eye	5,147
Female right eye	3,263
Female left eye	3,291

Observations from heatmap analysis

From 5.4 with a common color bar, the mean heatmap shows that OD is highly activated for correct predictions with male images, but not for the correct predictions with female images. From 5.5 with individual color bars, the mean heatmap shows that the temporal vascular arcade region is more activated than the OD region for correct predictions with female images.

5.3 Retinal predicted age and chronological age

5.3.1 Introduction

The progression of biological aging can vary from person to person. Biological age (also referred to as physiological age) may be a better indicator than chronological age for susceptibility to long-term degenerative diseases, particularly vascular ones like coronary artery disease, stroke, and vascular death [224]. To estimate biological age, the retina is a candidate source of biomarkers, as is the brain [225, 226]. Recent DL investigations [23, 158, 155, 162] have reported successful chronological age predictions using retinal fundus camera images.

This section presents our investigation on how *the difference between chronological age and retinal vascular age predicted by DL* associates with Major Adverse Cardiovascular Event (MACE) and All Cause Death (ACD) of Type 2 Diabetes (T2D) population in GoDARTS. Note that the age prediction analysis in Section 5.2 was using *all* the available retinal images in GoDARTS, whereas this section is focused on T2D population in GoDARTS.

5.3.2 Materials

Individuals considered are with type 2 diabetes above 30 years of age at the time of their first available retinal photograph from GoDARTS and excluded those who had a previous hospitalization history for Myocardial Infraction (MI) or stroke, identified using International Classification of Diseases (ICD)-10 codes I21-I23 and I60-I63. This resulted in a dataset

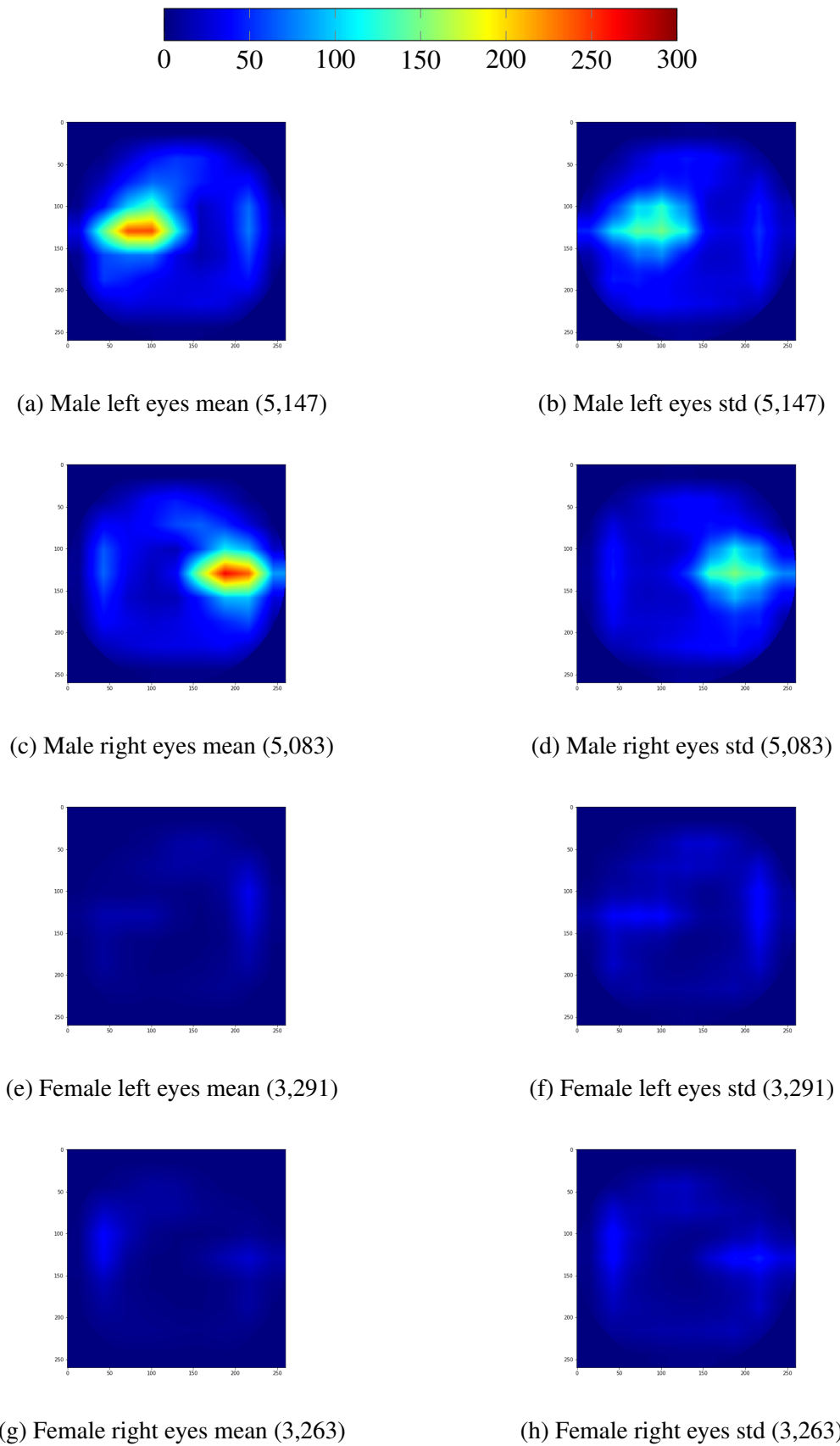
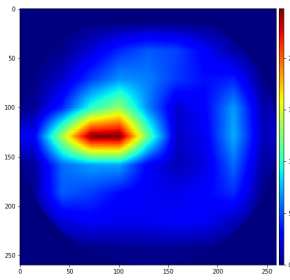
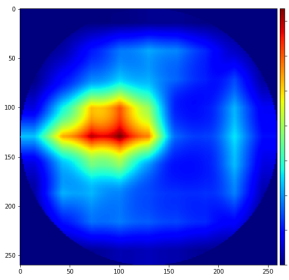


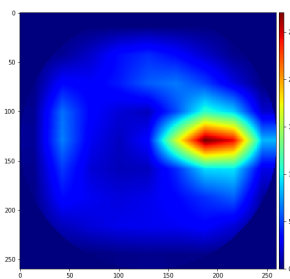
Fig. 5.4 Pixel wise mean and standard deviation heatmaps with a common color bar.



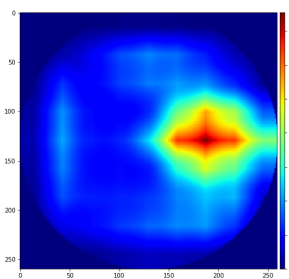
(a) Male left eyes mean (5,147)



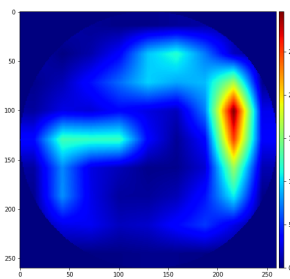
(b) Male left eyes std (5,147)



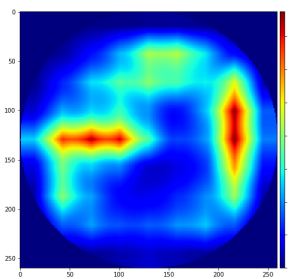
(c) Male right eyes mean (5,083)



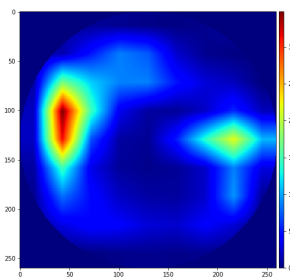
(d) Male right eyes std (5,083)



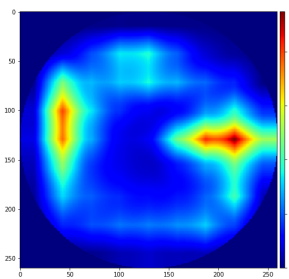
(e) Female left eyes mean (3,291)



(f) Female left eyes std (3,291)



(g) Female right eyes mean (3,263)



(h) Female right eyes std (3,263)

Fig. 5.5 Pixel wise mean and standard deviation heatmaps with individual color bars.

of 81,260 retinal images obtained at multiple time points over a span of 12 years, from 6,646 individuals (males = 3,717, females = 2,929). The descriptive characteristics for the data splits are shown in Table 5.9. Table 5.10 shows the data distribution in different age sub-groups.

Table 5.9 Dataset characteristics of whole cohort and data split in T2D participants.

	Overall	Train	Validation	Test
Participants	6,646	4,652	664	1,330
Images	81,260	56,944	8,191	16,125
Right eye Images (%)	40,390 (49.7)	28,298 (49.69)	4,077 (49.77)	8,015 (49.71)
Sex - Count (proportion in %)				
Male	3,717 (55.93)	2,610 (56.1)	361 (54.37)	746 (56.09)
Female	2,929 (44.07)	2,042 (43.9)	303 (45.63)	584 (43.91)
Age at imaging (in years) – Mean (std, IQR)				
All	66.92 (11.06, 59.82-75.06)	66.81 (11.02, 59.76-74.98)	67.62 (11.17, 60.65-76.01)	66.93 (11.15, 59.58-75.06)
Male	66.62 (10.68, 59.67-74.54)	66.52 (10.57, 59.71-74.26)	67.72 (10.82, 60.35-76.51)	66.43 (10.97, 59.19-74.55)
Female	67.29 (11.53, 60.07-75.56)	67.18 (11.57, 59.91-75.51)	67.51 (11.6, 61.38-75.66)	67.56 (11.35, 60.3-75.96)

Outcome variables: age at imaging, MACE and ACD

The procedures for obtaining chronological age (also referred to as actual age or age at imaging), outcomes for MACE and ACD of the individuals are described in Section 3.3.1.

5.3.3 Methods (survival analysis)

Image pre-processing and data split

The retinal images used have 30 different sizes in pixels, four of which form the vast majority (96%) of the data: 2336×3504 pixels (53,266 images, 65%), 2304×3456 (21,099 images, 26%), 1696×2544 (2,245 images, 2.7%) and 2178×3267 (1,893 images, 2.3%). Smaller

Table 5.10 Data distribution for age subgroups in the whole cohort and data splits of T2D individuals. In parenthesis, the value includes the proportion of data in the respective column.

Age group	Overall		Train		Validation		Test	
	# of individuals	# of Im-ages	# of individuals	# of Im-ages	# of individuals	# of Im-ages	# of individuals	# of Im-ages
0-10	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)
10-20	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)
20-30	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)
30-40	91 (1.37%)	521 (0.64%)	63 (1.35%)	344 (0.6%)	7 (1.05%)	48 (0.59%)	21 (1.58%)	129 (0.8%)
40-50	431 (6.49%)	3,970 (4.89%)	305 (6.56%)	2,792 (4.9%)	44 (6.63%)	364 (4.44%)	82 (6.17%)	814 (5.05%)
50-60	1,171 (17.62%)	12,011 (14.78%)	828 (17.8%)	8,467 (14.87%)	101 (15.21%)	1,105 (13.49%)	242 (18.2%)	2,439 (15.13%)
60-70	2,087 (31.4%)	25,730 (31.66%)	1,473 (31.66%)	18,173 (31.91%)	207 (31.17%)	2,565 (31.31%)	407 (30.6%)	4,992 (30.96%)
70-80	2,140 (32.2%)	27,678 (34.06%)	1,497 (32.18%)	19,545 (34.32%)	214 (32.23%)	2,750 (33.57%)	429 (32.26%)	5,383 (33.38%)
80-90	685 (10.31%)	10,665 (13.12%)	456 (9.8%)	7,108 (12.48%)	87 (13.1%)	1,305 (15.93%)	142 (10.68%)	2,252 (13.97%)
90-100	41 (0.62%)	685 (0.84%)	30 (0.64%)	515 (0.9%)	4 (0.6%)	54 (0.66%)	7 (0.53%)	116 (0.72%)

sizes account for only 4% of the images. The image pre-processing steps applied on these retinal images are described in Section 3.3.2.

Deep learning architecture and training

The deep learning architecture and training strategy followed is described in Section 5.2.2. The dataset was randomly partitioned into three subsets: 70% for training, 10% for validation, and 20% for testing. The dataset characteristics at baseline in all the splits were provided in Table 5.9. The training specifications are summarized in Table 5.11.

Table 5.11 Summary of training specifications for predicting age using all retinal images in T2D individuals of GoDARTS.

Category	Specification
Input and Output	
DL architecture	EfficientNet-B2
Input	Color fundus image
Input dimensions	260×260
Output	Linear activation
Performance metric	MAE, R^2
Training	
Weight Initialization	ImageNet and random uniform
Epochs	50
Batch size	32
Loss function	Mean squared error
Optimizer	Nadam
Learning rate	0.001 reduced by a factor of 0.1
Avoid overfitting	
Early stopping	on validation loss
Weights	Best validation loss

Activation visualization

The Grad-CAM [99] method is applied to the last convolutional layer of the trained DL model for visualizing the heatmaps. More details are available in Section 5.2.2.

Evaluation metrics

The evaluation metrics, MAE and R^2 are defined in Section 5.2.2, Evaluation metrics.

Predicted age difference

The Predicted Age Difference (PAD) is defined as the difference between the age predicted by the DL model from the retinal image and the chronological age of an individual on the date of the image captured. Following [227], we refer to a positive retinal PAD as ‘older’ than

chronological age and to a negative retinal PAD as ‘younger’ than chronological age retina. The estimated PAD was further used for survival analysis with ACD event and MACE.

Rate of change of PAD

The follow-up time considered was 5 years and the individuals whose retinal image is not available within the follow-up period were excluded from the analysis. Predicted Age Last First Difference (PALFD) (τ) is defined as the difference between the predicted age at the last available retinal image (in the follow-up period) and the predicted age from the first available retinal image (baseline). The PALFD rate (τ_{rate}) (Equation (5.6)) is defined as the change in predicted age from the last and first available retinal image during the follow-up window divided by the duration (in years) between the last and first date of retinal image acquisition.

$$\tau_{rate} = \frac{\text{Predicted age from last available image} - \text{Predicted age from first available image}}{\text{Duration between last and first image capture dates}} \quad (5.6)$$

The interpretation of the τ_{rate} (Equation (5.6)) is that if the value is >1 then the age is progressing faster than Chronological Rate (CR), $=1$ then the age is progressing in the CR, <1 then the age is progressing slower than CR. The τ_{rate} was used in the survival analysis to find the association with mortality event.

$$\tau_{rate} = \begin{cases} > 1, & \text{the age progression faster than CR} \\ = 1, & \text{the age progression as per CR} \\ < 1, & \text{the age progression slower than CR} \end{cases}$$

Survival analysis

As the mortality (ACD) event and MACE are associated to each individual rather than a left or right eye retinal image, All the individuals with both left and right eye retina available

at baseline in the test data were considered. To find the association between the PAD score and right censored ACD event and MACE, the Kaplan-Meier (KM) estimator was used on the upper and lower tertiles of the computed PAD score from the test data. To quantify the association further, Cox proportional hazard (Coxph) regression analysis with adjustment for age at retinal imaging and sex, and Pooled Cohort Equations (PCE) Atherosclerotic Cardiovascular Disease (ASCVD) risk score [30] was performed.

Recently, Raghu et al. [228] have shown that a 5-year increase in biological age predicted from chest radiograph images using DL indicates a higher risk of all-cause mortality than a 5-year increase in chronological age. Therefore, following [228], a similar procedure to find the association of retinal predicted age with mortality event and MACE were investigated in GoDARTS. KM curves were computed in retinal predicted age groups (<60, 60-69, 70-79, 80+) and Coxph regression were used.

To find the association between the retinal τ_{rate} and right censored mortality, the KM estimator was used on the upper and lower tertiles of the computed τ_{rate} from the test data. Coxph regression was used to quantify the association with mortality events by adjusting for predicted age at the first retinal image and sex.

Survival analysis was carried out using R (3.2.5) with the ‘survival’ package and ‘survminer’ package (for the plots) in the SH environment provided by HIC services [25].

5.3.4 Results

The mean age of retinal imaging at baseline (earliest image available) was 66.92 ± 11.06 years and the individuals were aged between 30.5 and 100 years. For this experiment, a total of 56,944 retinal images were used for training, 8,191 images for validation to avoid overfitting and 16,125 images for testing.

Age prediction using deep learning

For training, validation, and testing an EfficientNet-B2 model, GoDARTS retinal images of the left and right eyes captured at multiple time points were used. Table 5.12 shows the model’s performance on the age predictions from all the retinal images in the test data as well

as in its data subsets, which include left and right eye retinal images at baseline (first image available), only left eye images at baseline, only right eye images at baseline and average of predictions from left and right eye images at baseline. Gerrits et al. [162] was followed to compute the average of predictions from left and right eye images at baseline to represent individual-level prediction and these were used for survival analysis (Section 5.3.4, Early Mortality and MACE in older retina group).

Table 5.12 Model's performance for predicting age in T2D individuals. 95% CI values are computed using bootstrap samples.

Category	# Images	MAE (95% CI) years	R^2 (95% CI)
Complete test data	16,125	4.088 (4.037, 4.135)	0.77 (0.764, 0.776)
Baseline images - test data (left and right eye)	2,632	3.973 (3.862, 4.09)	0.795 (0.78, 0.808)
Only left eye images in test at baseline	1,316	4.094 (3.921, 4.267)	0.783 (0.76, 0.802)
Only right eye images in test at baseline	1,316	3.855 (3.698, 4.012)	0.805 (0.785, 0.823)
Average of left and right eye predictions from base- line images - test data	1,316 in- dividuals	3.692 (3.543, 3.844)	0.823 (0.806, 0.839)

The model achieved MAE of 4.088 (95% CI 4.037, 4.135) years and R^2 of 0.77 (0.764, 0.776) on the complete test data. MAE for the baseline images in test data by considering one left eye and one right image per individual is 3.973 (3.862, 4.09) years and R^2 is 0.795 (0.78, 0.808). Only left eye images in test data at baseline give MAE of 4.094 (3.921, 4.267) years and R^2 of 0.783 (0.76, 0.802). Only right eye images in test data at baseline give MAE of 3.855 (3.698, 4.012) years and R^2 of 0.805 (0.785, 0.823). The performance with the average of the individual's left and right eye age prediction at baseline in test data is MAE of 3.692 (3.543, 3.844) years and R^2 of 0.823 (0.806, 0.839). Figure 5.6 shows a scatter plot for predicted age and actual age in the whole test data and its subsets. The R^2 for male and female individuals in all these categories of the test dataset shows almost consistent performance.

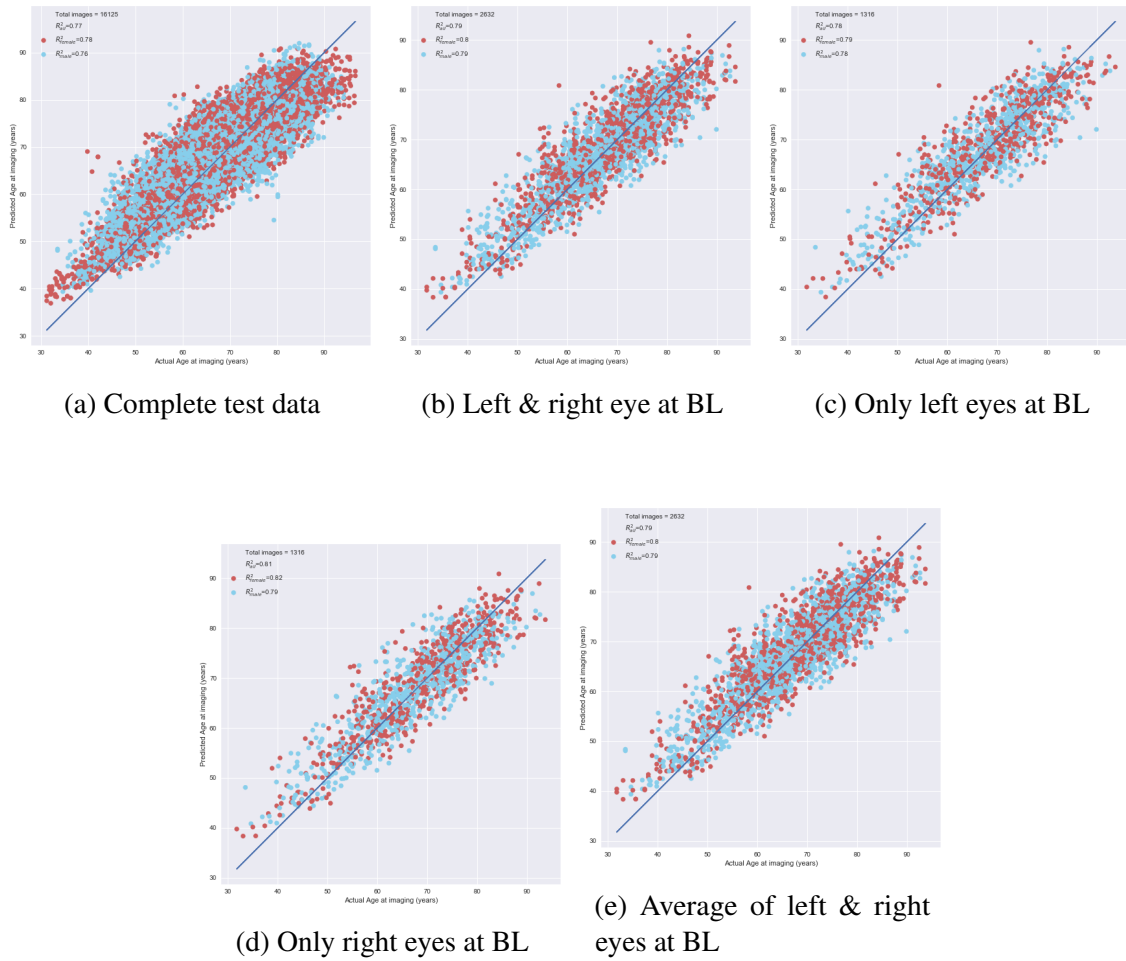


Fig. 5.6 Scatter plots for actual age at imaging and predicted age at imaging in the whole test data and its subset in T2D individuals. BL = baseline.

The MAE for age sub-groups with an interval of 10 years in the whole test data is shown in Table 5.13. The MAE is 3.5 years in the age group of 70-80 years (5,383 images), 3.8 years in the age group of 60-70 (4,992 images). The MAE is a little higher in the younger age groups between 30-50 years and older age groups >90 years and this might be because of the lesser number of individuals in these age groups. The age sub-group wise performance considering for other subset categories of test data as shown in Table 5.12 are provided in the Appendix A, Table A.3, A.4, A.5 and A.6.

The age sub-group wise performance trend i.e. MAE is lower in the groups with more images (60-80 years) and higher in the groups with fewer images (<50 years and >90 years),

in T2D individuals from Table 5.13 appear to be similar to that of the test performance using all the retinal images in GoDARTS which is shown in Table 5.6.

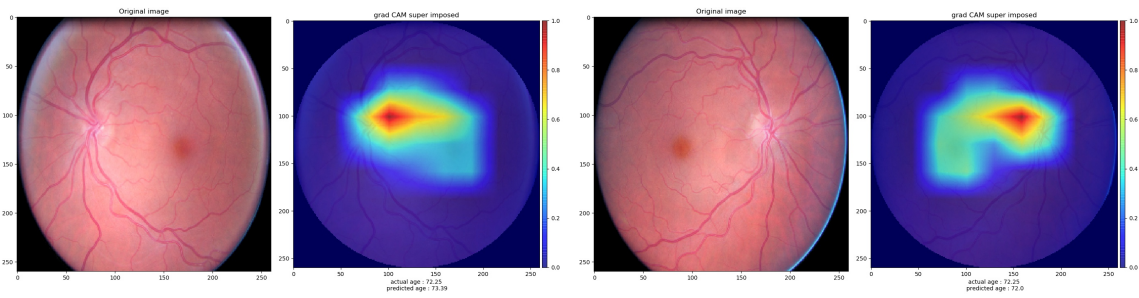
Table 5.13 Age sub-group-wise model performance with mean actual age, mean predicted age, MAE, and its 95% CI for the age sub-groups with an interval of 10 years on the complete test data of T2D individuals.

Age group	# Images	Mean actual age	Mean predicted age	MAE (95% CI)
0-10	0	-	-	-
10-20	0	-	-	-
20-30	0	-	-	-
30-40	129	36.904	43.270	6.374 (5.781, 7.04)
40-50	814	46.177	51.380	5.515 (5.24, 5.811)
50-60	2,439	55.835	59.577	4.707 (4.563, 4.858)
60-70	4,992	65.186	66.838	3.81 (3.729, 3.894)
70-80	5,383	74.842	74.131	3.468 (3.396, 3.537)
80-90	2,252	83.372	79.541	4.62 (4.487, 4.748)
90-100	116	92.379	83.377	9.0 (8.376, 9.661)

Activation heatmaps were generated for all the input retinal images in the test data for predicting age in individuals with type 2 diabetes (T2D) using Grad-CAM. Figure 5.7 shows four sample retinal images for two individuals (left eye and right eye). The visual inspection across several Grad-CAM heatmaps indicates that OD and macula are the most important features for predicting age at imaging in all age groups; additionally, the retinal vasculature is also important in younger and middle-age groups. These observations results are similar to the model performance using all the retinal images in GoDARTS, refer Section 5.2.3. More Grad-CAM heatmaps including both left and right eye images of three individuals are provided in Appendix A, Figure A.1.

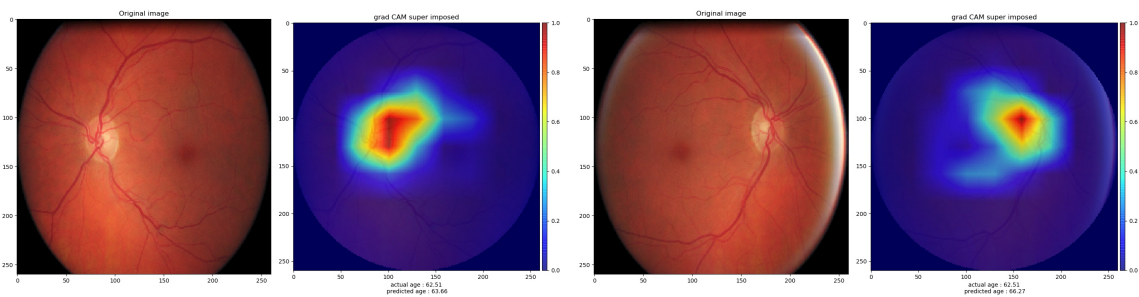
Early Mortality and MACE in older retina group

There are 466 mortality events and 305 MACE in the test population of 1,316 individuals used for the survival analysis. KM curves for mortality and MACE are shown in Figure 5.8, computed for the upper and lower tertiles based on the retinal PAD score computed from individual-level prediction at baseline. The time in years from the date of imaging (DoI) to the event is considered as the time variable for both mortality event and MACE. For



(a) I1 left eye, actual age 72.25, predicted age 73.39

(b) I1 right eye, actual age 72.25, predicted age 72.0

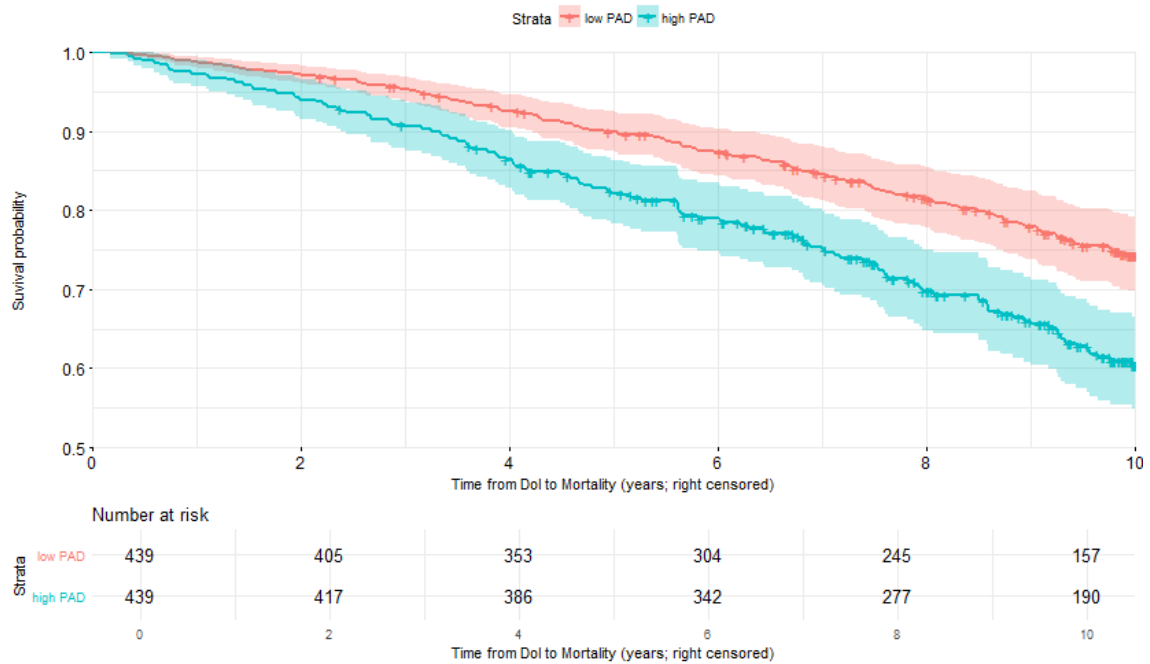


(c) I2 left eye, actual age 62.51, predicted age 63.66

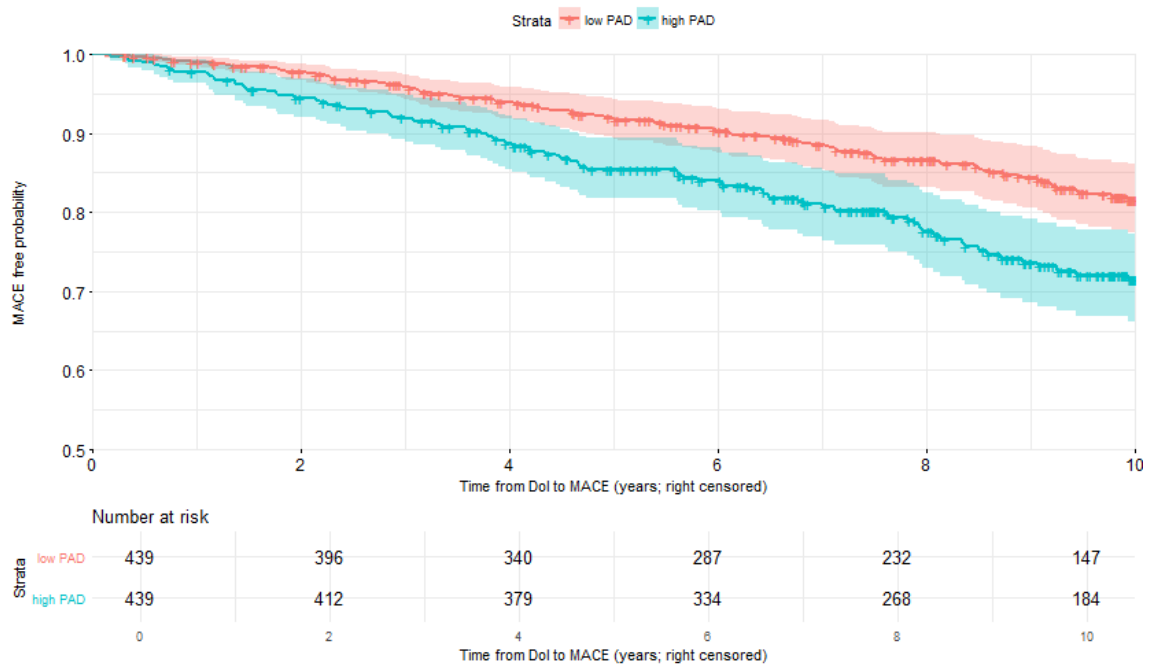
(d) I2 right eye, actual age 62.51, predicted age 66.27

Fig. 5.7 Sample grad-CAM heatmaps of two individuals from a model trained for age prediction in T2D individuals. I1 = Individual 1, I2 = Individual 2.

KM curves the time considered was until 10 years from the DoI for the event as shown in the x-axis. The KM curves are adjusted for actual age at retinal imaging to account for the PAD score as the individuals were included in the study at different ages. The KM curves show that the individuals in the upper tertile PAD (high PAD) group are significantly more associated with early mortality than those in the lower tertile PAD (low PAD) group. The range of PAD in the upper and lower tertile group are (2.41 to 16.12 years) and (-13.78 to -1.49 years) respectively. Similarly for MACE, the individuals in the high PAD group are significantly more associated to develop MACE than those in the low PAD group. KM curves plotted using only left-eye predictions and only right-eye predictions also show significant survival rates in low and high PAD groups similar to that of individual-level predictions. These plots are provided in the Appendix A, Table A.2 and A.3.



(a) KM curves for mortality



(b) KM curves for MACE

Fig. 5.8 KM curves of right-censored survival data for upper (high PAD) and lower (low PAD) tertiles of retinal PAD groups computed from the average of the left eye and right eye image predictions for age.

The relationship between mortality event/MACE and retinal PAD was tested using Coxph regression analysis (coxph R function) adjusted for age at imaging, and sex. A 1-year increase in retinal PAD score increases the risk of mortality by 5.9% (hazard ratio (HR) = 1.0597, 95% CI = 1.034 – 1.085, P = 1.62e-06). Individuals in the high PAD group have a higher risk of mortality than those in the low PAD by 79% (HR = 1.798, 95% CI = 1.385 – 2.335, P = 1.04e-05). The risk of developing MACE increases by 5.8% (HR = 1.0587, 95% CI = 1.028 - 1.089, P = 1.06e-4) with a 1-year increase in retinal PAD score and the individuals in the high PAD group are in high risk of developing MACE than the low PAD group by 73% (HR = 1.732, 95% CI = 1.258 – 2.386, P = 7.59e-4).

Coxph regression model was also adjusted for ASCVD risk score [30] alone as it considers age, sex, systolic and diastolic blood pressure, total and high-density lipoprotein cholesterol, diabetes history, and smoking status in calculating the risk score. The associations for mortality event and MACE with PAD score still remain significant. Coxph results adjusted for ASCVD risk score alone are, a 1-year increase in retinal PAD score increases the risk of mortality by 3.3% (hazard ratio (HR) = 1.033, 95% CI = 1.01 – 1.058, P = 0.0043). Individuals in the high PAD group have a higher risk of mortality than those in the low PAD by 46% (HR = 1.46, 95% CI = 1.126 – 1.894, P = 0.0043). The risk of developing MACE increases by 4% (HR = 1.04, 95% CI = 1.011 - 1.069, P = 0.006) with a 1-year increase in retinal PAD score and the individuals in the high PAD group are in high risk of developing MACE than the low PAD group by 50% (HR = 1.50, 95% CI = 1.086 – 2.059, P = 0.013).

The Coxph regression results with only left-eye predictions and only right-eye predictions are provided in the Appendix A, Table A.7 and A.8 and they show significant results similar to individual-level predictions.

Predicted age for MACE and mortality

KM curves for mortality and MACE are shown in Figure 5.9, computed, for four age groups in years (<60, 60-69, 70-79, 80+) based on the predicted age computed from individual-level prediction at baseline. Time from the date for imaging (DoI) to the event in years is considered as the time variable for both mortality events and MACE. For KM curves

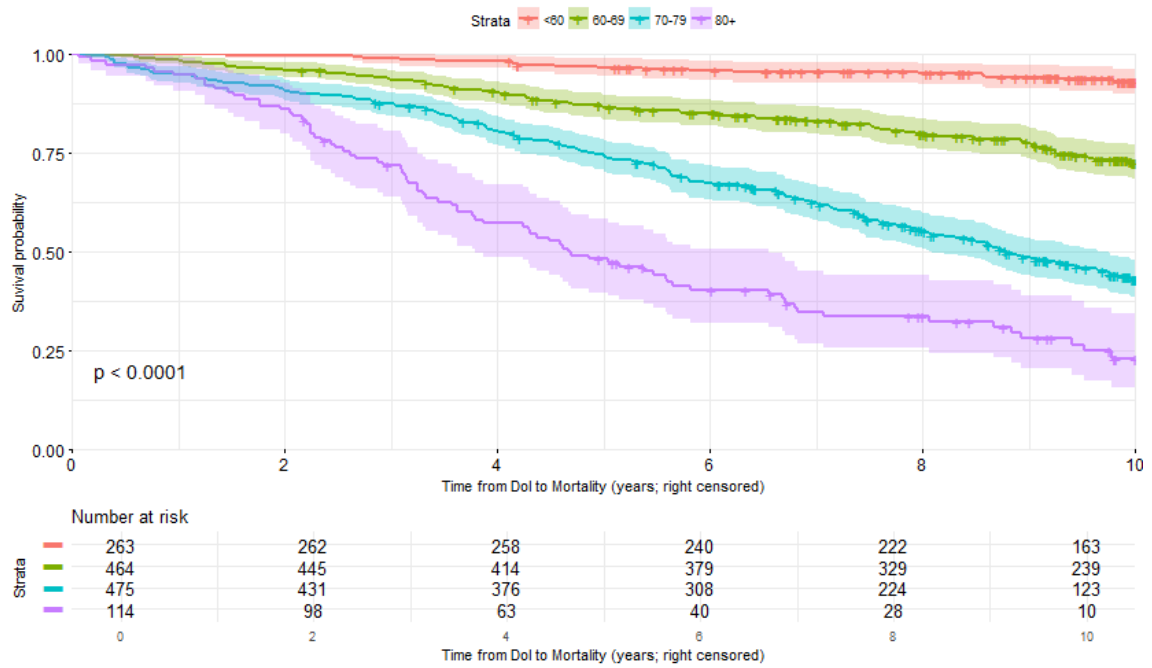
the time considered was until 10 years from the date of imaging for the event as shown in the x-axis. The KM curves show that the individuals in higher age groups are significantly more associated with early mortality events or MACE than those in the lower age groups. KM curves plotted using only left-eye predictions and only right-eye predictions also show significant associations in all four age groups similar to that of individual-level predictions. These plots are provided in the Appendix A, Figure A.4 and A.5.

From coxph regression, a 1-year increase in retinal predicted age is associated with higher mortality than the 1-year increase in the chronological age (retinal predicted age (HR = 1.112, 95% CI = 1.098 – 1.126, $p < 2e-16$) and chronological age (HR = 1.095, 95% CI = 1.083 – 1.106, $p < 2e-16$)). For MACE, a 1-year increase in retinal predicted age has a higher risk of developing MACE than 1-year increase in the chronological age (retinal predicted age (HR = 1.102, 95% CI = 1.085 – 1.118, $p < 2e-16$) and chronological age (HR = 1.086, 95% CI = 1.072 – 1.099, $p < 2e-16$)). The results are robust after adjusting for sex and are provided in the Appendix A, Table A.9 along with only left eye predictions and only right eye predictions as they show significant results like individual-level predictions.

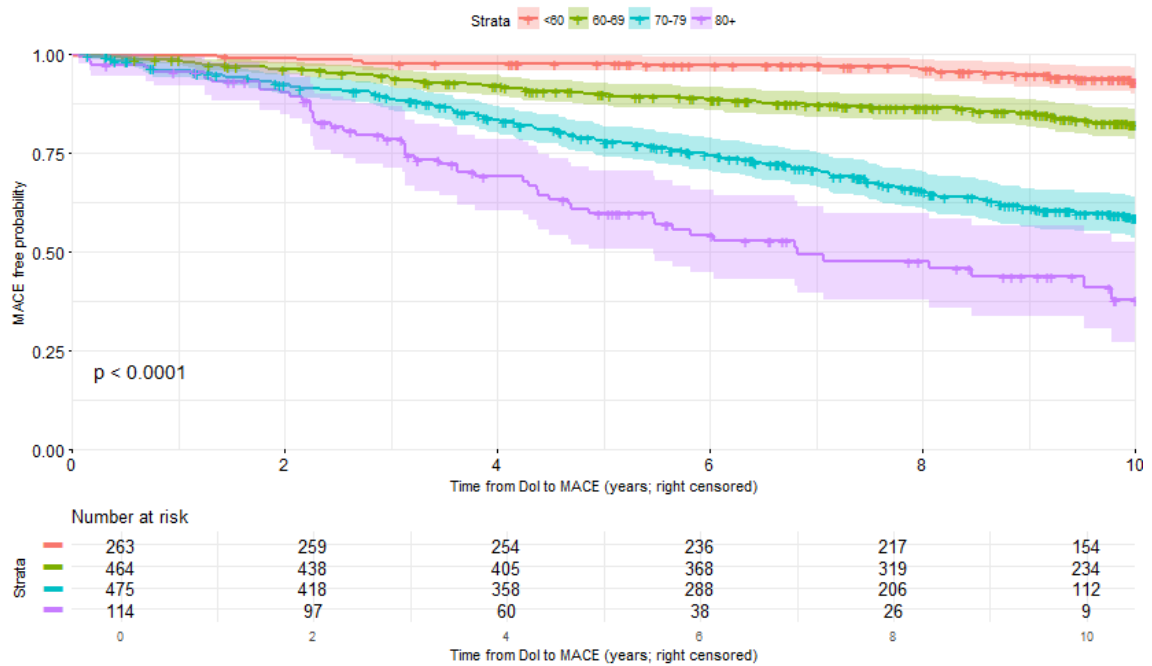
Rate of change of PAD

A follow-up period of 5 years was considered to compute τ_{rate} . The individuals who do not have at least one follow-up visit after the baseline or if they died before the last available retinal image were excluded. After applying the exclusion criteria, 326 mortality events from 1,172 individuals were obtained in the test dataset. The predicted age for the last available retinal image in the follow-up period was computed using the DL model.

KM curves for mortality event are shown in Figure 5.10, computed, for the upper and lower tertiles groups based on the retinal τ_{rate} computed from the individual-level prediction in the 5-year follow-up period. The time in years from the last available retinal date of imaging (DoI) in the follow-up period to the mortality event is considered as the time variable. For KM curves the time considered was until 5 years from the last date of imaging for the event(x-axis). KM estimates were adjusted for the predicted age at the first available image. The KM curves show that the individuals in the upper tertile τ_{rate} group (high τ_{rate})



(a) KM curves for mortality



(b) KM curves for MACE

Fig. 5.9 KM curves of right-censored survival data for groups of predicted age computed from the average of the left eye and right eye image predictions for age.

are significantly more associated with early mortality than those in the lower tertile τ_{rate} group (low τ_{rate}). The median value of τ_{rate} in the upper and lower tertile group are 1.61 and 0.25 respectively. KM curves plotted using only left-eye predictions show significant survival rates in the high and low tertile τ_{rate} groups similar to that of individual-level predictions. KM curves using only right-eye predictions do not show significant survival rates. These plots are provided in Appendix A, Figure A.6 and A.7.

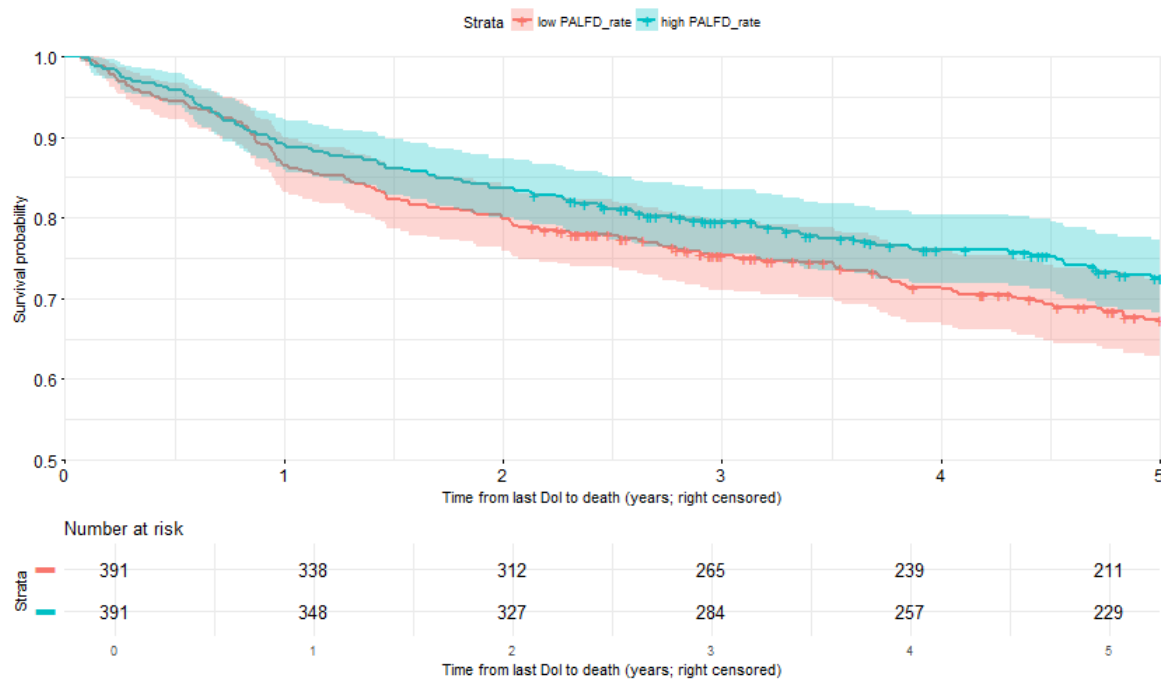


Fig. 5.10 KM curves for mortality from right censored survival data for upper (high τ_{rate}) and lower (low τ_{rate}) tertiles of τ_{rate} groups computed from the average of the left eye and right eye image predictions for age at last and first available retinal images. τ_{rate} = PALFD rate.

The Coxph regression analysis adjusted for predicted age at first available retinal image and sex indicates that individuals in the high τ_{rate} group have a higher risk of mortality than those in the low τ_{rate} group by 57% (HR = 1.575, 95% CI = 1.196 – 2.074, P = 0.0011). Details of coxph results using only left eye and only right eye predictions are provided in the Appendix A, Table A.10. Only left-eye predictions show significant results similar to individual-level predictions but no significant results were observed using only right-eye predictions.

5.4 Clinical measurements

This section describes materials and results for predicting continuous clinical measurements that act as risk factors for Cardiovascular (CV) related diseases. We investigated whether the retina can predict various known CV risk factors, namely Systolic Blood Pressure (SBP), Diastolic Blood Pressure (DBP), High-Density Lipoprotein (HDL), Total Cholesterol (TC), Glycated Haemoglobin (GH), Body Mass Index (BMI) and Tryglicerides (Trig) using deep learning.

5.4.1 Materials

A subset of baseline (earliest available) retinal images were selected from individuals in GoDARTS who had no prior history of hospital admissions due to MI or stroke, using ICD-10 codes I21-I23 and I60-I63. The dataset consisted of a total of 13,964 retinal images from 6,656 individuals, including both left- and right-eye images. In some cases, multiple images were available for the same individual from the same imaging session for quality assessment purposes. The distribution of images in the whole cohort and data splits is presented in Table 5.14.

Table 5.14 Retinal Image distribution of whole cohort and data splits for predicting continuous clinical measurements.

	Overall	Train	Validation	Test
Participants	6,656	4,659	665	1,332
Images	13,964	9,786	1,392	2,786
Right eye Images (%)	6,928 (49.61)	4,852 (49.58)	690 (49.57)	1,386 (49.75)

Outcome variables

The outcome variables considered in this experiment are continuous measurements of quantities regarded as the risk factors for CV diseases [23, 158, 162]. These features are SBP, DBP, HDL, TC, GH, BMI and Trig. Their values are computed as the median values over the 3 years of available measurements from the baseline imaging date as described in Section

3.3.1. The dataset characteristics for the whole data as well as data splits are shown in Table 5.15.

Table 5.15 Baseline characteristics of clinical measurements in GoDARTS; n = individuals used for the feature.

Feature (units)	sex	Overall						Train						Validation						Test					
		n	Mean (std)	Range	IQR	n	Mean (std)	Range	IQR	n	Mean (std)	Range	IQR	n	Mean (std)	Range	IQR	n	Mean (std)	Range	IQR				
Age (years)	All	6,655	67.21 (11.18)	17.98-96.45	60.08-75.42	4,659	67.0 (11.1)	17.98-96.45	59.9-75.12	665	67.32 (11.32)	25.31-93.48	59.84-75.97	1,331	67.88 (11.4)	26.95-94.57	60.8-76.1								
	Male	3,721	66.92 (10.78)	27.4-96.22	59.99-75.04	2,610	66.72 (10.65)	27.4-96.22	59.82-74.52	349	67.19 (11.06)	31.12-91.74	60.04-75.94	762	67.47 (11.1)	29.61-93.59	60.46-75.77								
	Female	2,934	67.57 (11.67)	17.98-96.45	60.25-75.96	2,049	67.35 (11.64)	17.98-96.45	60.03-75.74	316	67.46 (11.61)	25.31-93.48	59.28-76.06	569	68.44 (11.79)	26.95-94.57	61.39-76.48								
DBP (mmHg)	All	6,047	76.0 (7.88)	48.0-108.5	70.5-81.0	4,224	76.06 (7.83)	51.0-108.5	70.5-81.0	603	75.59 (7.97)	52.0-102.0	70.0-80.0	1,220	75.99 (8.03)	48.0-108.0	70.0-81.0								
	Male	3,386	76.52 (7.99)	48.0-108.5	71.0-82.0	2,368	76.62 (7.95)	54.0-108.5	71.0-82.0	320	76.29 (8.2)	52.0-102.0	70.0-81.0	698	76.31 (8.02)	48.0-108.0	70.0-81.0								
	Female	2,661	75.33 (7.7)	51.0-106.0	70.0-80.0	1,856	75.34 (7.61)	51.0-104.0	70.0-80.0	283	74.8 (7.64)	53.5-91.0	69.5-80.0	522	75.56 (8.02)	53.0-106.0	70.0-81.0								
SBP (mmHg)	All	6,656	138.9 (11.52)	93.5-208.0	132.5-144.5	4,659	138.8 (11.33)	98.0-208.0	132.5-144.0	665	138.17 (11.53)	100.0-205.0	132.0-144.0	1,332	139.63 (12.14)	93.5-198.0	132.5-146.0								
	Male	3,721	138.55 (11.53)	93.5-208.0	132.0-144.0	2,610	138.67 (11.33)	100.0-208.0	132.0-144.0	349	137.27 (11.89)	100.0-205.0	131.0-143.0	762	138.7 (12.01)	93.5-198.0	132.0-145.0								
	Female	2,934	139.35 (11.5)	98.0-198.5	133.0-145.0	2,049	138.96 (11.33)	98.0-198.5	132.5-144.5	316	139.16 (11.04)	111.0-184.0	134.0-145.0	569	140.86 (12.23)	101.0-189.5	134.0-147.0								
BMI (Kg/m ²)	All	6,562	31.15 (6.0)	15.2-59.1	27.0-34.3	4,596	31.18 (6.05)	16.0-59.1	27.06-34.31	650	31.03 (5.9)	15.2-52.7	26.92-34.39	1,316	31.13 (5.91)	17.1-56.81	26.8-34.22								
	Male	3,666	30.52 (5.34)	15.95-59.1	26.85-33.31	2,570	30.56 (5.36)	16.0-59.1	27.0-33.44	340	30.28 (5.38)	15.95-51.5	26.5-32.94	756	30.47 (5.26)	18.45-55.15	26.6-33.26								
	Female	2,895	31.96 (6.66)	15.2-57.7	27.25-35.7	2,026	31.96 (6.73)	16.9-57.7	27.21-35.7	310	31.85 (6.32)	15.2-52.7	27.56-35.59	559	32.03 (6.59)	17.1-56.81	27.29-35.74								
GH (%)	All	6,214	7.52 (1.15)	6.08-12.8	6.7-8.1	4,357	7.51 (1.14)	6.08-12.8	6.65-8.05	621	7.53 (1.18)	6.08-12.7	6.65-8.15	1,236	7.56 (1.15)	6.08-12.7	6.7-8.15								
	Male	3,473	7.5 (1.12)	6.08-12.7	6.7-8.05	2,444	7.5 (1.12)	6.08-12.6	6.7-8.02	325	7.58 (1.19)	6.1-12.5	6.7-8.2	704	7.49 (1.11)	6.08-12.7	6.7-8.0								
	Female	2,740	7.54 (1.19)	6.08-12.8	6.69-8.1	1,913	7.53 (1.18)	6.08-12.8	6.65-8.05	296	7.48 (1.18)	6.1-12.7	6.6-8.1	531	7.64 (1.21)	6.1-12.3	6.7-8.3								
HDL (mmol/L)	All	6,656	1.34 (0.35)	0.23-3.94	1.11-1.48	4,659	1.34 (0.35)	0.23-3.94	1.12-1.48	665	1.32 (0.37)	0.23-3.02	1.09-1.45	1,332	1.32 (0.34)	0.53-2.99	1.11-1.47								
	Male	3,721	1.27 (0.32)	0.23-3.55	1.06-1.38	2,610	1.28 (0.32)	0.23-3.55	1.07-1.4	349	1.23 (0.3)	0.52-2.54	1.02-1.34	762	1.26 (0.32)	0.53-2.99	1.06-1.36								
	Female	2,934	1.42 (0.37)	0.23-3.94	1.2-1.58	2,049	1.43 (0.37)	0.47-3.94	1.2-1.59	316	1.42 (0.41)	0.23-3.02	1.18-1.57	569	1.41 (0.34)	0.59-2.84	1.21-1.56								
TC (mmol/L)	All	6,656	4.37 (0.85)	1.45-14.28	3.82-4.79	4,659	4.38 (0.86)	1.45-14.28	3.82-4.79	665	4.35 (0.84)	2.11-10.89	3.84-4.78	1,332	4.35 (0.82)	1.89-9.24	3.81-4.8								
	Male	3,721	4.25 (0.84)	1.45-14.28	3.7-4.68	2,610	4.26 (0.85)	1.45-14.28	3.71-4.68	349	4.23 (0.89)	2.11-10.89	3.66-4.66	762	4.21 (0.81)	1.89-9.24	3.68-4.68								
	Female	2,934	4.53 (0.84)	2.41-9.59	3.98-4.95	2,049	4.53 (0.86)	2.41-9.59	3.97-4.95	316	4.48 (0.77)	2.51-7.54	4.03-4.9	569	4.55 (0.81)	2.73-8.21	3.99-5.01								
Trig (mmol/L)	All	3,471	2.31 (1.95)	0.31-44.61	1.28-2.71	2,424	2.34 (2.07)	0.32-44.61	1.27-2.72	353	2.2 (1.52)	0.45-14.39	1.29-2.73	694	2.25 (1.68)	0.31-21.12	1.32-2.61								
	Male	1,947	2.4 (2.27)	0.31-44.61	1.28-2.79	1,362	2.43 (2.42)	0.33-44.61	1.27-2.8	198	2.33 (1.69)	0.5-14.39	1.33-2.87	387	2.33 (1.97)	0.31-21.12	1.32-2.6								
	Female	1,524	2.18 (1.43)	0.32-14.51	1.28-2.63	1,062	2.22 (1.5)	0.32-14.51	1.28-2.64	155	2.05 (1.26)	0.45-9.7	1.26-2.54	307	2.14 (1.22)	0.55-9.45	1.3-2.61								

5.4.2 Methods

Image pre-processing and data split

The 13,964 retinal images used have 14 different sizes in pixel, three of which form the vast majority (98.8%) of the data: 2336×3504 pixels (12,936 images, 92.6%), 2304×3456 (560 images, 4%), 1696×2544 (310 images, 2.2%). Smaller sizes account for only 1.2% of the images. The image pre-processing steps applied on these retinal images are described in Section 3.3.2.

Deep learning architecture and training

The deep learning architecture and training strategy followed is described in Section 5.2.2. The dataset was randomly partitioned into three subsets: 70% for training, 10% for validation, and 20% for testing. The dataset characteristics at baseline in all the splits were provided in Table 5.15. The training specifications are summarized in Table 5.16.

Evaluation metrics

As the output labels are continuous features MAE and R^2 were used as the evaluation metrics. These are defined in Section 5.2.2.

5.4.3 Results

A total of 9,786 retinal images were utilized for training, 1,392 images for validation, and 2,786 images for testing. All the models were trained independently for the prediction of continuous clinical features.

Model performance for predicting continuous clinical features

The performance results on complete test data for predicting continuous clinical features are shown in Table 5.17. The model achieved MAE of 5.88 (95% CI 5.71, 6.07) and R^2 of 0.14 (0.1, 0.17) for the prediction of DBP on the complete test dataset. For the other

Table 5.16 Summary of training specifications for predicting continuous clinical features using retinal images.

Category	Specification
Input and Output	
DL architecture	EfficientNet-B2
Input	Color fundus image
Input dimensions	260×260
Output	Linear activation
Performance metric	MAE, R^2
Training	
Weight Initialization	ImageNet and random uniform
Epochs	50
Batch size	32
Loss function	Mean Squared Error (MSE)
Optimizer	Nadam
Learning rate	0.001 reduced by a factor of 0.1
Avoid overfitting	
Early stopping	on validation loss
Weights	Best validation loss

clinical feature, the R^2 values are approximately equal to zero specifying that there is no information available in the retina for the corresponding feature predictions using this dataset. Figure 5.11 displays scatter plots illustrating the predicted and actual labels for all continuous features in the test data. The green line represents the diagonal line between the x- and y-axis. Figure 5.11a shows a positive correlation between the predicted and actual values of DBP, indicating that the DL model could extract some information or signal from the fundus image to predict DBP. However, for the remaining clinical features, there is no observed correlation between the predicted and actual values, suggesting that the DL model was unable to extract information for predicting these features. Moreover, the scatter plots highlight that the DL model merely learns the mean value of the clinical feature it is trained on, as the scatter plots mainly concentrate on the mean value of the feature provided in Table 5.15.

Table 5.17 Model's performance for predicting continuous clinical features. Bootstrap samples are utilized to estimate the 95% confidence interval (CI) values.

Feature	No. of Images	MAE (95% CI)	R^2 (95% CI)
DBP	2,555	5.88(5.71,6.07)	0.14(0.1,0.17)
SBP	2,786	9.28(8.97,9.59)	-0.07(-0.1,-0.05)
BMI	2,754	4.39(4.24,4.53)	0.04(0.01,0.07)
GH	2,590	0.87(0.84,0.9)	0.05(0.02,0.09)
HDL	2,786	0.25(0.25,0.26)	-0.01(-0.04,0.01)
TC	2,786	0.62(0.6,0.64)	0.02(0.01,0.04)
Trig	1,444	1.02(0.96,1.09)	-0.0(-0.02,0.02)

5.5 Disease outcomes

Here the retinas were investigated for predicting the binary disease outcomes within 12 years from the date of retinal imaging using deep learning in the GoDARTS diabetic cohort. The disease events considered are macrovascular complications (MACE), microvascular complications (Chronic Kidney Disease (CKD), Diabetic Peripheral Neuropathy (DPN), Diabetic Retinopathy (DR)) and ACD. This section describes the materials and results for predicting disease outcomes, represented by binary labels (disease 1, no disease 0), from retinal images.

5.5.1 Materials

For the classification of MACE and ACD events, the retinal images described in 5.4.1 were considered. For the classification of microvascular complications (CKD, DPN, DR), a total of 17,139 retinal images were used which were obtained from 8,222 individuals and it includes both left- and right-eye retinal images. The image distribution of the whole data and the data splits used for microvascular complications is shown in Table 5.18.

Outcome variables

The outcome variables in this experiment are categorical and binary. The extraction of these features (MACE, ACD, CKD, DPN and DR) from the GoDARTS bio-resources is described

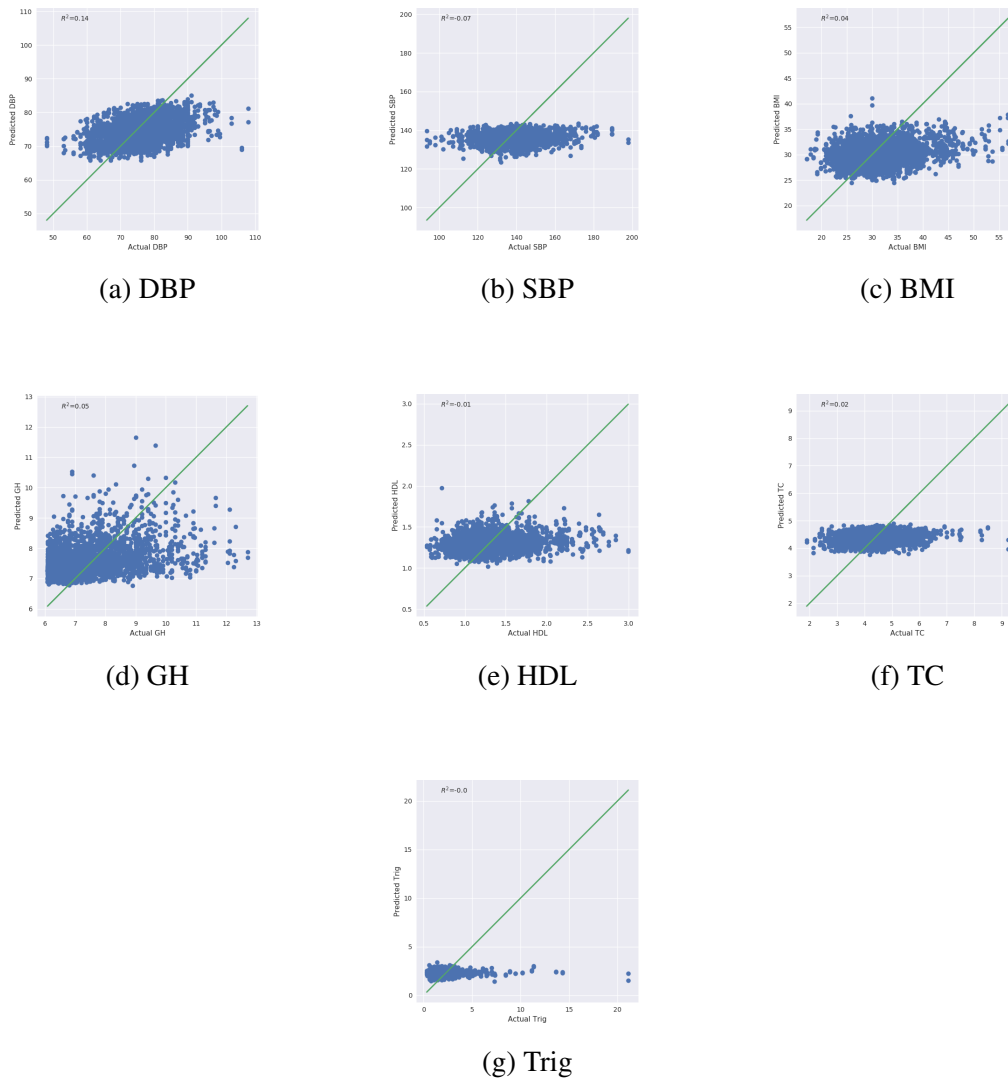


Fig. 5.11 Scatter plot for actual and predicted labels of clinical measurements in the whole test data. The green line is a diagonal line.

Table 5.18 Retinal Image distribution of whole cohort and data splits for predicting microvascular complications.

	Overall	Train	Validation	Test
Participants	8,222	5,755	822	1,645
Images	17,139	11,988	1,718	3,433
Right eye Images (%)	8,499 (49.6%)	5,944 (49.6%)	850 (49.5%)	1,705 (49.67%)

in Section 3.3.1. The dataset characteristics for the whole data as well as data splits are shown in Table 5.19.

Table 5.19 Baseline characteristics of Disease outcomes in GoDARTS; n = individuals used for the feature. prop. = proportion.

Feature	sex	Overall		Train		Validation		Test	
		n	events count (% prop.)	n	events count (% prop.)	n	events count (% prop.)	n	events count (% prop.)
MACE	All	6,656	1,696 (25.48)	4,659	1,186 (24.46)	665	166 (24.96)	1,332	344 (25.83)
	Male	3,721	985 (26.47)	2,610	677 (25.94)	349	93 (26.65)	762	215 (28.22)
	Female	2,934	711 (24.23)	2,049	509 (24.84)	316	73 (24.84)	569	129 (22.67)
ACD	All	6,656	2,393 (35.95)	4,659	1,653 (35.48)	665	239 (35.94)	1,332	501 (37.61)
	Male	3,721	1,409 (37.87)	2,610	976 (37.39)	349	127 (36.39)	762	306 (40.16)
	Female	2,934	984 (33.54)	2,049	677 (33.04)	316	112 (35.44)	569	195 (34.27)
DR	All	7,601	4,528 (40.43)	5,323	3,141 (40.99)	768	463 (39.71)	1,510	924 (38.81)
	Male	4,271	2,521 (40.97)	3,007	1,753 (41.7)	434	261 (39.86)	830	507 (38.92)
	Female	3,330	2,007 (39.73)	2,316	1,388 (40.07)	334	202 (39.52)	680	417 (38.68)
CKD	All	7,562	3,336 (44.12)	5,273	2,328 (44.15)	765	348 (45.49)	1,524	660 (43.31)
	Male	4,118	2,509 (39.07)	2,883	1,754 (39.16)	418	265 (36.6)	817	490 (40.02)
	Female	3,444	827 (24.01)	2,390	574 (24.02)	347	83 (23.92)	707	170 (24.05)
DPN	All	7,688	2,915 (37.92)	5,389	2,035 (37.76)	770	270 (35.06)	1,529	610 (39.9)
	Male	4,351	1,649 (37.9)	3,066	1,162 (37.9)	434	140 (32.26)	851	347 (40.78)
	Female	3,337	1,266 (37.94)	2,323	873 (37.58)	336	130 (38.69)	678	263 (38.79)

5.5.2 Methods

Image pre-processing and data split

The image pre-processing steps applied on these retinal images are described in Section 3.3.2.

Deep learning architecture and training

The deep learning architecture and training strategy followed is described in Section 5.2.2. The dataset was randomly partitioned into three subsets: 70% for training, 10% for validation, and 20% for testing. The dataset characteristics at baseline in all the splits were provided in Table 5.19. The training specifications are summarized in Table 5.20.

Evaluation metrics

As the output labels are binary features AUC, accuracy, sensitivity and specificity were used as the evaluation metrics which are defined in Section 5.2.2, Evaluation metrics.

5.5.3 Results

Five models were trained, validated and tested individually for predicting the binary disease outcomes, MACE, ACD, DR, CKD and DPN.

Table 5.20 Summary of training specifications for predicting disease outcomes using retinal images.

Category	Specification
Input and Output	
DL architecture	EfficientNet-B2
Input	Color fundus image
Input dimensions	260×260
Output	Sigmoid activation
Performance metric	AUC, accuracy, sensitivity and specificity
Training	
Weight Initialization	ImageNet and random uniform
Epochs	50
Batch size	32
Loss function	binary cross-entropy
Optimizer	Nadam
Learning rate	0.001 reduced by a factor of 0.1
Avoid overfitting	
Early stopping	on validation loss
Weights	Best validation loss

Prediction of MACE

The results for predicting MACE are shown in Table 5.21. The model achieved AUC of 0.642 (95% CI 0.619, 0.665), accuracy of 0.738 (0.722, 0.754), sensitivity of 0.065 (0.048, 0.084) and specificity of 0.973 (0.965, 0.98) on the complete test data.

Prediction of ACD

The results for predicting ACD are shown in Table 5.21. The model achieved AUC of 0.741 (0.722, 0.759), accuracy of 0.7 (0.682, 0.718), sensitivity of 0.543 (0.513, 0.573) and specificity of 0.794 (0.775, 0.813) on the complete test data.

Table 5.21 Model’s performance for predicting disease outcomes. 95% CI values are computed using bootstrap samples.

Feature	# images	AUC (95% CI)	Accuracy (95% CI)	Sensitivity (95% CI)	Specificity (95% CI)
MACE	2,786	0.642 (0.619, 0.665)	0.738 (0.722, 0.754)	0.065 (0.048, 0.084)	0.973 (0.965, 0.98)
ACD	2,786	0.741 (0.722, 0.759)	0.7 (0.682, 0.718)	0.543 (0.513, 0.573)	0.794 (0.775, 0.813)
DR	3,143	0.633 (0.614, 0.653)	0.624 (0.607, 0.641)	0.885 (0.87, 0.899)	0.211 (0.187, 0.233)
CKD	3,177	0.711 (0.693, 0.728)	0.657 (0.641, 0.674)	0.578 (0.551, 0.605)	0.718 (0.698, 0.738)
DPN	3,195	0.57 (0.55, 0.591)	0.601 (0.584, 0.618)	0.009 (0.005, 0.015)	0.997 (0.994, 0.999)

Prediction of microvascular disease outcomes

The results for predicting DR, CKD and DPN are shown in Table 5.21. For CKD prediction, the model achieved AUC of 0.711 (0.693, 0.728), accuracy of 0.657 (0.641, 0.674), sensitivity of 0.578 (0.551, 0.605) and specificity of 0.718 (0.698, 0.738) on the complete test data. For DR and DPN, the AUC is 0.633 and 0.57 respectively. The detailed performance results are provided in Table 5.21

5.6 Discussions

Recent research indicates that retinal images, which are located at the back of the eye, are crucial for vision, highly vascularized, and can offer valuable insights into overall vascular health and biological age. This is particularly relevant for diabetic individuals, who are at greater risk of age-related complications due to metabolic dysfunction associated with diabetes. Retinal images provide a window for detecting early signs of aging and predicting age-related illnesses, allowing for early intervention and prevention. The use of deep learning to predict age and sex from fundus images has significant clinical implications in ophthalmology, enabling ophthalmologists to identify patients who are at a higher risk of developing certain eye diseases and facilitating early detection and diagnosis. Consequently, retinal images are an increasingly important tool for evaluating overall health and age-related changes in the body.

As outlined in Section 3.3.2, the retinal images in the GoDARTS dataset underwent preprocessing by fitting a circle to the retinal region, which removed surface artifacts.

Although this resulted in the clipping of a small portion of the outer surface of the retina and potential loss of retinal information, the central area of the fundus image contains the most critical information, rendering this loss negligible. The clipping process was essential to eliminate artifacts that could lead to bias during deep learning (DL) training.

The deep learning model was trained on retinal images of 8,570 individuals from the GoDARTS dataset, using 70% of the images for training, 10% for validation, and 20% for testing. The model achieved a mean absolute error (MAE) of 3.951 years and an R^2 of 0.81 for age prediction, and an area under the curve (AUC) of 0.9 for sex prediction on the test dataset. These results are comparable to other studies in the literature using larger retinal image datasets. Poplin et al. [23] achieved a MAE of 3.26 years and R^2 of 0.81 for age prediction, and an AUC of 0.97 for sex prediction using UK Biobank's retinal data from 48,101 individuals for training and 12,026 individuals for testing. Kim et al. [155] achieved a MAE of 3.55 years and R^2 of 0.75 for age prediction, and an AUC of 0.97 for sex prediction in individuals with diabetes using fundus images from 155,449 individuals for training and testing a DL model. Rim et al. [158] reported a MAE of 4.5 years and R^2 of 0.51 for age prediction, and an AUC of 0.8 for sex prediction in the UK Biobank used as an external test dataset, using retinal images from 72,890 individuals for training and testing DL models. It should be noted that the mean age of individuals in the GoDARTS dataset (66.11 ± 11.77 years) is higher than the relevant studies in the literature, such as Poplin et al. [23] (56.9 ± 8.2 years), Kim et al. [155] (46.64 ± 15.83), and Rim et al. [158] (53.0 ± 7.67). In another experiment reported in Section 5.3.4, it was observed that the performance improved for age prediction at the individual level (average of left- and right-eye predictions) compared to single-eye predictions. The reason for this improvement is unknown and further investigation is required.

Based on the findings presented in Table 5.6 and Table 5.13, it appears that the DL model tends to overestimate the age of younger individuals (below 60 years) and underestimate the age of older individuals (above 80 years). This may be due to the saturation of aging changes in the retinal fundus image in older individuals, while aging changes may occur more rapidly in younger individuals. Furthermore, the imbalanced distribution of data in different age

groups may have impacted the training of the DL model in extreme age groups, resulting in this pattern as a potential artifact of the DL model itself, which has been reported by other studies with similar cohorts [155, 229]. Additional research and consultation with domain experts are required to explore this further.

Zhu et al. [170] demonstrated a significant correlation between PAD and mortality from non-cardiovascular and non-cancer causes using retinal images from UK Biobank. Similarly, Section 5.3.4 of this thesis found stronger associations between MACE and ACD. A major advantage of the current study is that the retinal images were acquired in actual clinical settings as part of a diabetic screening program, rather than being derived from study-level data like UK Biobank.

The results presented in Section 5.4.3 indicate a moderate R^2 value for predicting DBP using retinal images, but not for SBP and other cardiovascular (CV) risk factors. Comparable studies in the literature [23, 158, 155, 162] have also reported the potential of retinal data in predicting SBP, DBP, BMI, and lipid profiles to some extent. However, it should be noted that most of these studies have primarily focused on relatively younger age groups with a mean age of around 45-55 years, while our study includes individuals with a mean age of 66 years. Therefore, further investigation is required to better understand the potential of retinal data in predicting CV risk factors in our age group. Future analyses could benefit from information on medications used by patients for treating hypertension or lipid profiles.

The DL model using Grad-CAM heatmap algorithm identified the optic disc (OD) and macula as the most significant features for predicting age from fundus images. With age, the size of the optic disc can increase, and the shape of its border can become irregular [230]. Similarly, the macula is a small, highly sensitive area in the center of the retina that is responsible for sharp, detailed vision. The macula can undergo changes that result in decreased visual acuity and color perception as a result of aging [230]. Therefore, it makes sense that the deep learning model identified the OD and macula as the most significant features for predicting age from fundus images. Expert opinions are further needed to assess the reasoning.

This study has yielded interesting results, but it is not without limitations. One limitation is that all retinal images from the GoDARTS cohort have a Field of View (FoV) of 45 degrees, which may affect the generalizability of the results to retinal images with different FoV. Second, the volume of retinal image data used in this study (8,570 individuals) is relatively small compared to the UK Biobank and EyePACS datasets used in other studies (297,360 individuals) [23]. Additionally, there were many missing values for Trig compared to other CV risk factors available for this study. Lastly, these experiments were only conducted on an elderly diabetic cohort in Scotland, and further validation of the experiments is necessary in other health conditions, including existing disease conditions, medications, various age groups, different geographical backgrounds, and ethnicities.

5.7 Conclusions

In this chapter, we presented our investigation results on predicting demographic and clinical features from only fundus images using the EfficientNet-B2 DL model in GoDARTS. For the prediction of age and sex a total of 102,082 images from 8,570 individuals for training, validating, and testing EfficientNet-B2 were used. The model achieved MAE of 3.951 (95% CI 3.908, 3.995) and R^2 of 0.809 (0.804, 0.814) for age prediction and an AUC of 0.899 (0.895, 0.903) for sex prediction on the complete test data. These results are comparable and support the recent literature on age and sex prediction from retinal images using DL. We observed a little higher error for age predictions and a little lower AUC for sex prediction.

The Grad-CAM heatmap generated for the retinal images in the test data shows that OD and macula are the critical regions for the age prediction in all age groups and the retinal vasculature is additionally important in younger and middle age groups. For sex prediction, from Grad-CAM heatmaps it appears that OD and the region around the macula are the important features. Further, we noticed from the Grad-CAM heatmaps that for the majority of male predictions and female predictions, the OD region and the temporal vascular arcade region are activated respectively. To assess this observation we performed a systematic

analysis to check the consistency of the heatmap activations in all the male and female predictions separately.

The PAD computed for all the T2D individuals with age over 30 years in the test data shows statistically significant associations with the risk of developing 10 years MACE and risk of 10 years mortality. Coxph regression shows that associations remain significant even after adjusting the regression model with ASCVD score which considers age, sex, systolic and diastolic blood pressure, total and high-density lipoprotein cholesterol, diabetes history and smoking status in calculating the risk score. The longitudinal analysis by computing τ_{rate} (see Section 5.3.3, Rate of change of PAD) with a follow-up period of 5 years show a statistically significant association with ACD but not with MACE. One possible reason might be because of the less number of MACE in the dataset, the predictive power of the model might be less.

We trained DL model to predict CV risk factors namely SBP, DBP, HDL, TC, GH, BMI and Trig independently from retinal fundus images in GoDARTS, refer Section 5.4. The model achieved MAE of 5.88 (95% CI 5.71, 6.07) and R^2 of 0.14 (0.1, 0.17) for the prediction of DBP on the complete test dataset with 2,786 images. R^2 values for the rest of the CV risk factors are approximately equal to zero specifying that there is no information available in the retina for the corresponding feature predictions using this dataset, refer Section 5.4.3.

The model's performance in terms of AUC for the risk stratification of systemic disease outcomes within 12 years from the date of retinal imaging using EfficientNet-B2 is 0.74, 0.71, 0.642, 0.633, and 0.57 for CKD, ACD, MACE, DR and DPN respectively. The results show that the retina image has some signal for stratifying CKD, ACD, MACE, DR but no signal for DPN stratification using the cross-sectional image analysis in GoDARTS, refer Section 5.5.3.

In this chapter, we observed that retinal images can be very well used to predict age and sex. The Grad-CAM heatmap also highlights the OD, macula, and vasculature regions for the predictions but the exact visualizations of the important regions are not very clear. Further studies are required in the direction of visualizing the CNN predictions. For the prediction of CV risk factors, we do not find any significant results except for DBP. For systemic conditions

the results show that retina images can be used for predictions of CKD, ACD, MACE, DR but not for DPN. In Chapter 6 we investigate predicting Genome-Wide Polygenic Risk Scores (GWPRS) and clinical risk scores for Cardiovascular Disease (CVD); the association of predicted risk score with 10 years CVD from retinal images using EfficientNet-B2.

Chapter 6

Predicting cardiovascular risk scores

6.1 About this chapter

This chapter presents the introduction, materials, methods, and performance results of the model for the prediction of clinical Cardiovascular (CV) risk score and Genome-Wide Polygenic Risk Scores (GWPRS) ¹. It also provides the investigation results on how the predicted clinical CV risk score stratifies individuals with Major Adverse Cardiovascular Event (MACE) and Cardiovascular death (CV death)².

6.2 Introduction

Due to the increasing global burden of Cardiovascular Disease (CVD), there is an urgent need to identify individuals at risk in a rapid and cost-effective manner, to enable effective prevention at both individual and population levels. Currently, clinical risk assessments, such as the Pooled Cohort Equations (PCE) Atherosclerotic Cardiovascular Disease (ASCVD)

¹**Syed, M. G., Doney, A., George, G., Mordi, I., and Trucco, E. (2021).** *Are cardiovascular risk scores from genome and retinal image complementary? A deep learning investigation in a diabetic cohort.* In *International Workshop on Ophthalmic Medical Image Analysis - OMIA (MICCAI workshops)*, pages 109–118. Springer. This publication is based on the work presented in Section 6.5.

²**Syed, M. G., Trucco, E., Doney, A., and Mordi, I.** *Integrating a Deep-Learning Cardiovascular Risk Score Derived from Retinal Images and a Coronary Heart Disease Polygenic Risk Score to Predict Clinical Outcomes.* This manuscript is currently being prepared for submission to a journal and is based on the research described in Section 6.6.

risk score, have limited performance in practice and do not incorporate many well-established markers of CV risk, such as Body Mass Index (BMI), leading to overestimation of risk in some populations and underestimation in others. This highlights the need for improved refinements in CV risk prediction tools in clinical practice. Frequent assessments, such as blood pressure and cholesterol checks, required for these risk scores, also place an additional burden on healthcare services and patients, further emphasizing the need for more efficient and effective risk assessment strategies.

CVD can largely be prevented through lifestyle modifications and medical management. Accurately predicting CVD risk at an early stage, conveniently and simply, can enable timely intervention and have important clinical benefits. CVD risk is determined by a combination of genetic and environmental/lifestyle factors. Recently, GWPRS have been shown to predict CVD risk with similar accuracy to conventional clinical risk scoring methods, such as the PCE ASCVD risk score. Combining a clinical score with a GWPRS may further increase prediction accuracy. GWPRS can be easily and inexpensively determined through chip-based assays, whereas determining a clinical risk score is comparatively more complex and resource-intensive, requiring a clinic visit to obtain a range of clinical measures to be combined with other patient information.

Recent Deep Learning (DL) approaches have shown that the retina may provide information indicative of cardiovascular disease (CVD) risk [169, 162, 23, 158]. This has led to increasing interest in the retina as a potential source of biomarkers for CVD risk, as discussed in detail in Section 2.4. Notably, retinal images can be easily and efficiently captured, including with portable devices that utilize mobile-phone technology. However, it remains unclear to what extent retinal information complements clinical and genomic risk factors in predicting CVD risk. Therefore, this study aimed to investigate:

1. the potential of a DL approach applied to retinal images in predicting clinical risk score and a GWPRS for CVD.
2. how the clinical risk score predicted from retinal images using DL stratifies individuals with MACE and CV death in 10 years.

6.3 Materials

The subset of baseline (earliest available) images used for this experiment is the same as the ones used for the prediction of clinical measurements experiments, described in Section 5.4.1. The image distribution is shown in Table 5.14 for the whole cohort and data splits.

6.3.1 Outcome variables

clinical risk

The ASCVD clinical risk was determined by computing the PCE risk score [30]. This score considers variables such as age, sex, systolic and diastolic blood pressure, total and High-Density Lipoprotein (HDL) cholesterol, diabetes history, and smoking status to estimate the percentage risk of ASCVD over 10 years. The necessary variables were obtained from the electronic health record data available at the time of the baseline retinal image for each participant in Genetics of Diabetes Audit and Research in Tayside Scotland (GoDARTS), and the PCE risk score was calculated accordingly.

GWPRS

The GWPRS was derived from the genotyping data available in the GoDARTS bio-resource, utilizing previously published methods [231]. The resulting score was standardized using z-scores. Descriptive characteristics of the dataset, including those of the entire dataset and its subsets, are presented in Table 6.1.

Table 6.1 Baseline characteristics of CV risk scores in GoDARTS; n = individuals used for the feature.

Feature	sex	Overall			Train			Validation			Test					
		n	Mean (std)	Range	n	Mean (std)	Range	n	Mean (std)	Range	n	Mean (std)	Range	IQR		
Age	All	6,655	67.21(11.18)	17.98-96.45	60.08-75.42	4,659	67.0(11.1)	17.98-96.45	59.9-75.12	665	67.32(11.32)	25.31-93.48	59.84-75.97	67.88(11.4)	26.95-94.57	60.8-76.1
	Male	3,721	66.92(10.78)	27.4-96.22	59.99-75.04	2,610	66.72(10.65)	27.4-96.22	59.82-74.52	349	67.19(11.06)	31.12-91.74	60.04-75.94	67.47(11.1)	29.61-93.59	60.46-75.77
	Female	2,934	67.57(11.67)	17.98-96.45	60.25-75.96	2,049	67.35(11.64)	17.98-96.45	60.03-75.74	316	67.46(11.61)	25.31-93.48	59.28-76.06	68.44(11.79)	26.95-94.57	61.39-76.48
Clinical risk	All	6,638	0.34(0.2)	0.0-1.0	0.18-0.48	4,647	0.34(0.2)	0.0-1.0	0.18-0.47	663	0.34(0.21)	0.01-0.99	0.16-0.48	0.35(0.2)	0.01-0.97	0.19-0.5
	Male	3,711	0.36(0.18)	0.0-0.92	0.23-0.49	2,602	0.36(0.17)	0.0-0.9	0.23-0.48	348	0.36(0.18)	0.01-0.88	0.21-0.5	0.37(0.18)	0.01-0.92	0.23-0.51
	Female	2,927	0.32(0.22)	0.01-1.0	0.13-0.46	2,045	0.31(0.22)	0.01-1.0	0.13-0.45	315	0.32(0.23)	0.01-0.99	0.13-0.46	0.33(0.23)	0.01-0.97	0.14-0.48
GWPRS	All	6,441	6.95(0.61)	4.4-9.26	6.63-7.35	4,508	6.95(0.61)	4.4-8.74	6.64-7.36	645	6.94(0.62)	4.65-8.63	6.63-7.34	6.93(0.63)	4.6-9.26	6.62-7.34
	Male	3,607	6.93(0.63)	4.4-9.26	6.6-7.35	2,534	6.95(0.62)	4.4-8.74	6.61-7.36	334	6.92(0.62)	5.14-8.63	6.6-7.33	6.9(0.64)	4.6-9.26	6.59-7.31
	Female	2,833	6.96(0.6)	4.65-8.63	6.67-7.35	1,974	6.96(0.59)	4.75-8.63	6.67-7.35	311	6.96(0.63)	4.65-8.51	6.67-7.37	6.98(0.62)	4.65-8.57	6.68-7.37
													Count (%)			
sex_male(%)	Male		3,721 (55.91)				2,610 (56.02)				349 (52.48)				762 (57.25)	
sex_male(%)	Female		2,934 (44.09)				2,049 (43.98)				316 (47.52)				569 (42.75)	

6.4 Methods

6.4.1 Pre-processing

Different sizes of retinal images used in this experiment and pre-processing applied is described in Section 5.4.2.

6.4.2 Deep learning architecture and training

The deep learning architecture and training strategy followed are described in Section 5.4.2. For the prediction of PCE ASCVD risk score the linear activation function in the output node was replaced with a sigmoid activation as the risk scores lie between 0 and 1. The training specifications are summarized in Table 6.2.

Table 6.2 Summary of training specifications for predicting CV risk scores using retinal images.

Category	Specification
Input and Output	
DL architecture	EfficientNet-B2
Input	Color fundus image
Input dimensions	260×260
Output	
<i>for PCE ASCVD</i>	Sigmoid activation
<i>for GWPRS</i>	Linear activation
Performance metric	Mean Absolute Error (MAE), R^2
Training	
Weight Initialization	ImageNet and random uniform
Epochs	50
Batch size	32
Loss function	Mean Squared Error (MSE)
Optimizer	Nadam
Learning rate	0.001 reduced by a factor of 0.1
Avoid overfitting	
Early stopping	on validation loss
Weights	Best validation loss

Evaluation metrics

The evaluation metrics, MAE and R^2 are defined in Section 5.2.2.

6.5 Prediction of CV risk scores

6.5.1 Results

9,786 retinal images were utilized for training, 1,392 for validation, and 2,786 for testing to prevent overfitting. Two separate models were developed for estimating the PCE ASCVD risk score and the genetic risk score, which were trained, validated, and tested individually. The bootstrap results for the test data are presented in Table 6.3, showing MAE and R^2 scores. The model achieved an R^2 of 0.554 (95% CI 0.528, 0.579) and MAE of 0.107 (0.104, 0.11) for predicting the PCE ASCVD risk score. For GWPRS, the R^2 was -0.005 (-0.019 , 0.009) with an MAE of 0.484 (0.467, 0.5). Figure 6.1 displays the scatter plots for the actual and predicted labels in the test data.

Table 6.3 Model performance on estimating PCE ASCVD and genetic risk scores in the test data. 95% CI values are computed using 2,000 bootstrap samples.

Feature	No. of Images	R^2 (95% CI)	MAE (95% CI)
PCE ASCVD risk score	2,778	0.554 (0.528, 0.579)	0.107 (0.104, 0.11)
Genetic risk score	2,688	-0.005 (-0.019 , 0.009)	0.484 (0.467, 0.5)

Figure 6.1a depicts a positive correlation between the actual and predicted labels for PCE ASCVD risk score estimation, indicating that the model successfully learned to estimate this score from retinal images. However, the model did not learn any significant associations between retinal images and genetic risk score beyond the average of the genetic risk score. This lack of correlation is evident in Figure 6.1b, where the scatter plot primarily centers around the mean value of GWPRS (value: 7) on the y-axis.

Gradient-based Class Activation Mapping (Grad-CAM) heatmaps were generated from the model trained to estimate the PCE ASCVD risk score using test images. These heatmaps were generated at the last convolutional layers. Figure 6.2 displays four sample retinal images

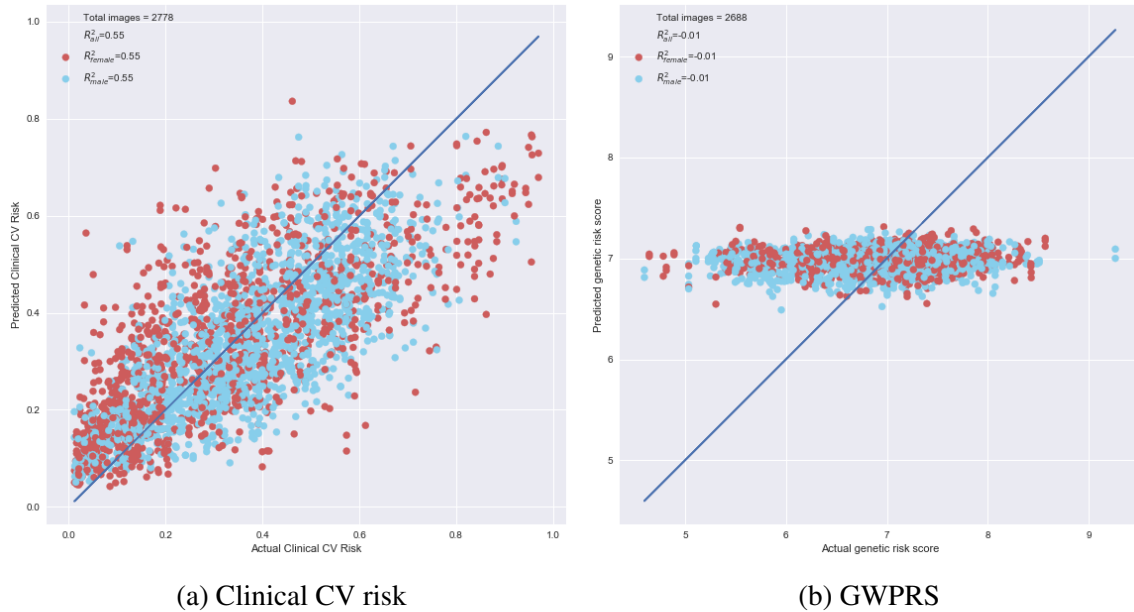


Fig. 6.1 Scatter plot depicting the actual and predicted labels in the test data for CV risk score and GWPRS. The blue line is a diagonal line.

and their respective heatmaps generated for the prediction. Appendix A, Figure A.8 presents more example heatmaps. The most important features for classifying the PCE ASCVD risk score from retinal images were the optic disc, macula, and vasculature. Additionally, the major branching point in the bottom half of the retina vasculature seemed to be highlighted as shown in Figure 6.2. However, the heatmap visualizations were not very clear, and further investigation is required in the future.

6.6 Prediction of MACE and CV death

Further to the model performance results for predicting the clinical CV risk scores, the predicted risk scores from the test data using the trained DL model were used to analyze the MACE and CV death events within 10 years from the date of retinal imaging.

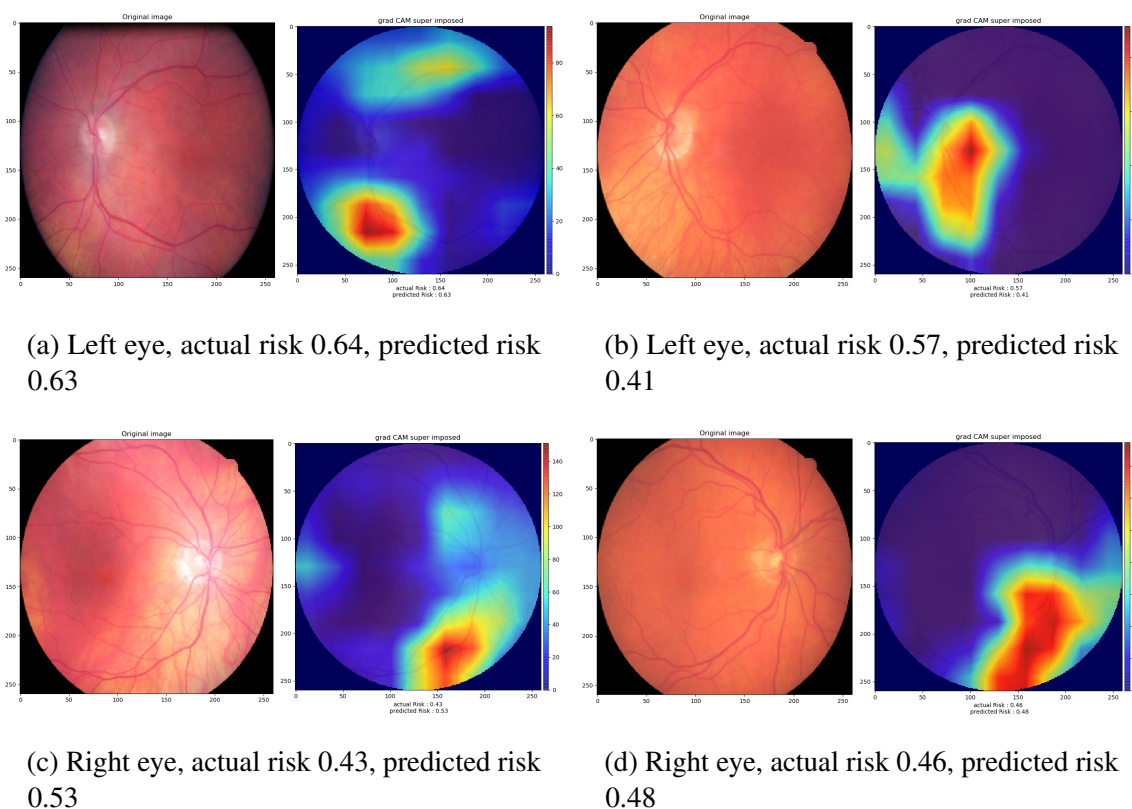


Fig. 6.2 Sample grad-CAM heatmaps for PCE risk prediction. The top row is right-eye images. The bottom row is left-eye images.

6.6.1 Materials

For this analysis individuals at baseline who had both left and right eye retinas available in the test dataset were considered. By applying the inclusion criteria, 1,317 individuals with 2,634 retinal images were obtained (one left and one right eye image per individual).

Outcome variables

MACE: Details about MACE definition and feature extraction are described in Section 3.3.1.

CV death: Causes of death were obtained from the Scottish Mortality records, General Register of Scotland, with any ICD-10 code from I00-I99 within the first two causes of death being classified as a CV death.

6.6.2 Methods

The methods used in this section for survival analysis are similar to the ones used for Predicted Age Difference (PAD) analysis, described in Section 5.3.3. This section presents the methods which are specific to the analysis of predicted CV risk scores.

Predicted risk score

The Predicted Risk Score (PRS) is defined as the prediction of the PCE ASCVD risk score from a retinal image, using the trained deep learning model. The PRS was further used in the survival analysis for MACE and CV death events.

Rate of change of PRS

The follow-up time considered was 3 years and the individuals whose retinal image was not available within the follow-up period were excluded from the analysis. Predicted Risk Last First Difference (PRLFD) (Ω) was defined as the difference between the predicted risk score at the last available retinal image (in the follow-up period) and the predicted risk score from the first available retinal image (baseline). The PRLFD rate (Ω_{rate}) (Equation 6.1) is defined as the change in predicted risk score from the last and first available retinal image during the follow-up window divided by the duration (in years) between the last and first date of retinal image acquisition.

$$\Omega_{rate} = \frac{\text{Predicted risk score from last available image} - \text{Predicted risk score from first available image}}{\text{Duration between last and first image capture dates}} \quad (6.1)$$

The interpretation of the Ω_{rate} (Equation 6.1) is as follows: if the value is positive the risk is progressing faster; if zero, there is no change in the risk progression rate; if negative, the risk is progressing slower. The Ω_{rate} was used in the survival analysis to find the association with MACE and CV death events.

$$\Omega_{rate} = \begin{cases} > 0, & \text{the risk progression is faster} \\ = 0, & \text{no change in risk} \\ < 0, & \text{the risk progression is slower} \end{cases}$$

Survival Analysis

To find the association between the retinal PRS and right censored MACE and CV death events, the Kaplan-Meier (KM) estimator was used on the upper and lower tertiles of the computed PRS from the test data. To quantify the association further, we performed Cox proportional hazard (Coxph) regression with adjustment for age at retinal imaging, sex, and GWPRS.

To find the association between the retinal Ω_{rate} and right censored MACE and CV death events, the KM estimator was used. Considering the ranges of Ω_{rate} values in the test data, the data was split as the top 20% and the bottom 80%. Coxph regression was used to quantify the association with MACE and CV death event.

6.6.3 Results on PCE Risk prediction at baseline

The mean age at baseline was 67.21 ± 11.19 years and the interquartile range is 60.19-75.42 years. Male and female distribution in the data was 56% and 44% respectively. Table 6.1 provides a comprehensive statistical summary of the dataset for the entire cohort as well as for the data splits at baseline.

Table 6.4 shows the model's performance on the risk predictions in the test dataset with left and right eye images per individual at baseline and individual-level predictions at baseline. MAE for the baseline images (including both left and right eyes) is 0.106 (0.103, 0.109) and R^2 is 0.561 (0.534, 0.588). The performance from individual-level predictions at baseline is MAE 0.101 (0.096, 0.105) and R^2 is 0.609 (0.577, 0.639). The R^2 for male and female individuals in all these categories of the test dataset is consistent. Scatter plots for actual PCE risk score and predicted PCE risk score for all these categories are shown in Figure 6.3.

PRS generated from the individual-level predictions in the test dataset was used for survival analysis, described in the next Section.

Table 6.4 The performance of the models in estimating ASCVD risk scores using the test data, along with 95% CI computed using 2,000 bootstrap samples.

Feature	No. of Images	R^2 (95% CI)	MAE (95% CI)
Left and right eye images per individual at baseline	2,634	0.561 (0.534, 0.588)	0.106 (0.103, 0.109)
individual-level predictions at baseline	1,317 individuals	0.609 (0.577, 0.639)	0.101 (0.096, 0.105)

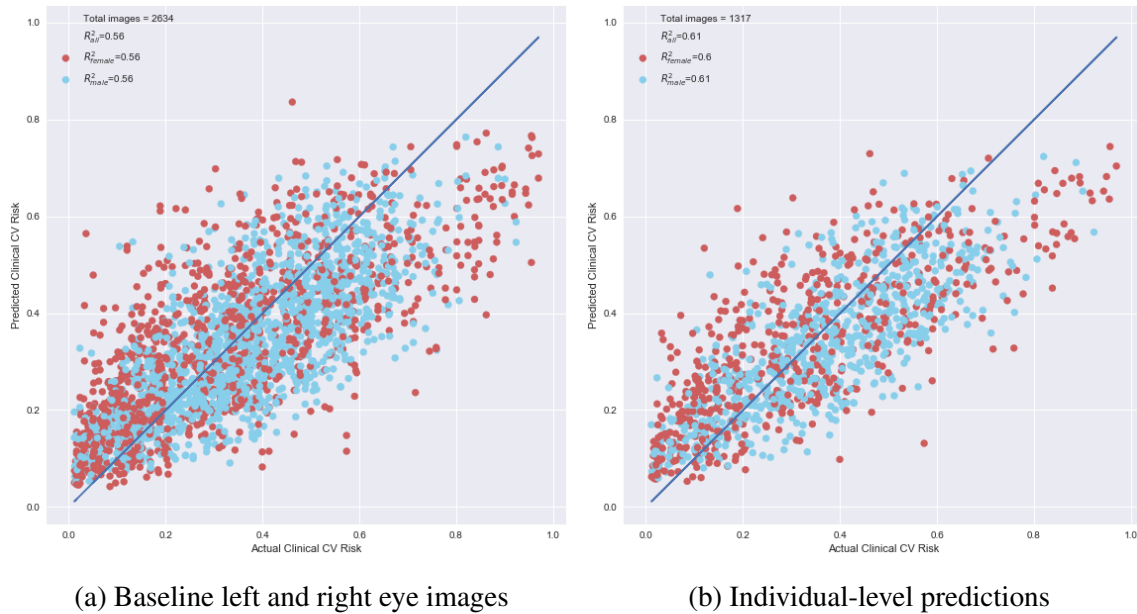


Fig. 6.3 Scatter plot depicting the actual and predicted labels in the test data for clinical CV risk score. The blue line is a diagonal line.

6.6.4 Results on MACE and CV death in high PRS group

MACE (329 events) and CV deaths (267 events) within 10 years from the date of retinal imaging were considered for the analysis. KM curves for 10-year MACE and CV deaths are shown in Figure 6.4, computed for the tertile split based on the retinal PRS derived from individual-level prediction at baseline in the test data. The range of PRS in the upper and

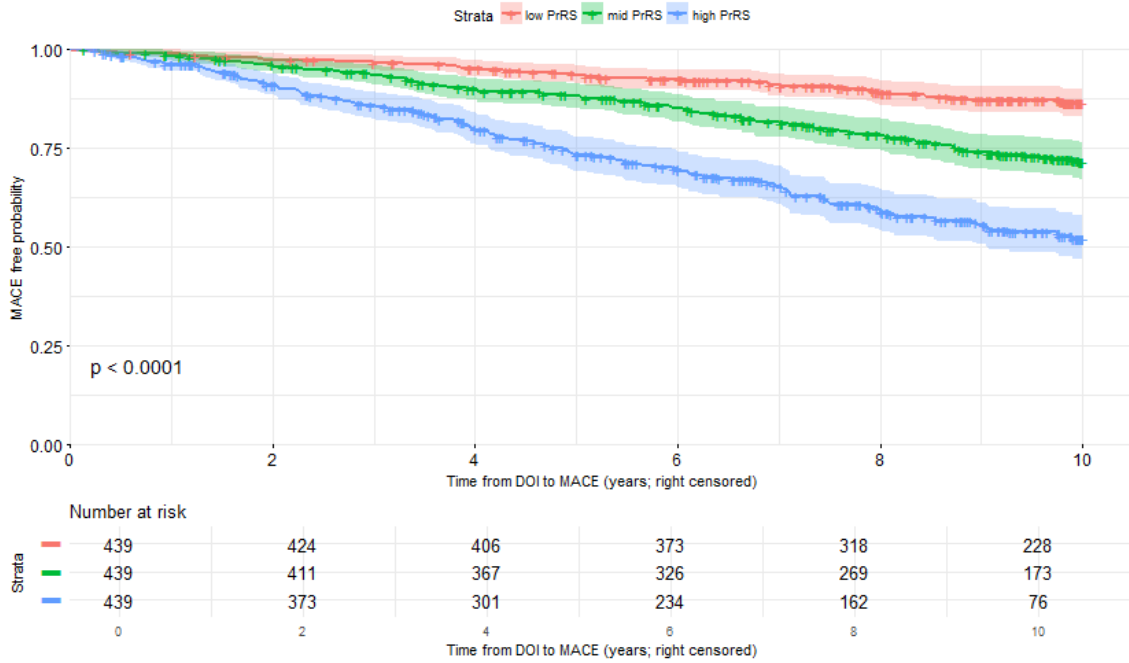
lower tertile groups are (0.4 to 0.74) and (0.05 to 0.25) respectively. Time from the date of imaging (DoI) to the event in years is considered as the time variable for both events. The KM curves show that the individuals in the upper PRS tertile (high PRS) group are significantly more associated with developing MACE than those in the lower PRS tertile (low PRS) group. Similarly, for CV death, the individuals in the high PRS group are significantly more associated with early CV death than those in the low PRS group. KM curves plotted using only left-eye and only right-eye predictions for CV risk scores also show significant survival rates in the tertile groups similar to that of individual-level predictions. These plots are provided in the Appendix A, Figure A.9 and A.10.

The relationship between the events and retinal PRS was tested using Coxph regression analysis (coxph R function) by adjusting for age at imaging, sex, and GWPRS. The risk of developing MACE increases by 2.9% (HR = 1.029, 95% CI = 1.015 - 1.042, P = 3.4e-5, n=1,273, events=317) with 1% increase in retinal PRS and the individuals in the high PRS group are in high risk of developing MACE than the low PRS group by 80.1% (HR = 1.805, 95% CI = 1.156 – 2.817, P = 0.009, n=1,273, events=317). One percent increase in the retinal PRS increases the risk of CV death by 1.9% (hazard ratio (HR) = 1.019, 95% CI = 1.004 – 1.034, P = 0.009, n=1,273, events=255). Individuals in the high PRS group have a higher risk of CV death than those in the low PRS by 62.5% (HR = 1.625, 95% CI = 0.965 – 2.736, P = 0.06, n=1,273, events=255).

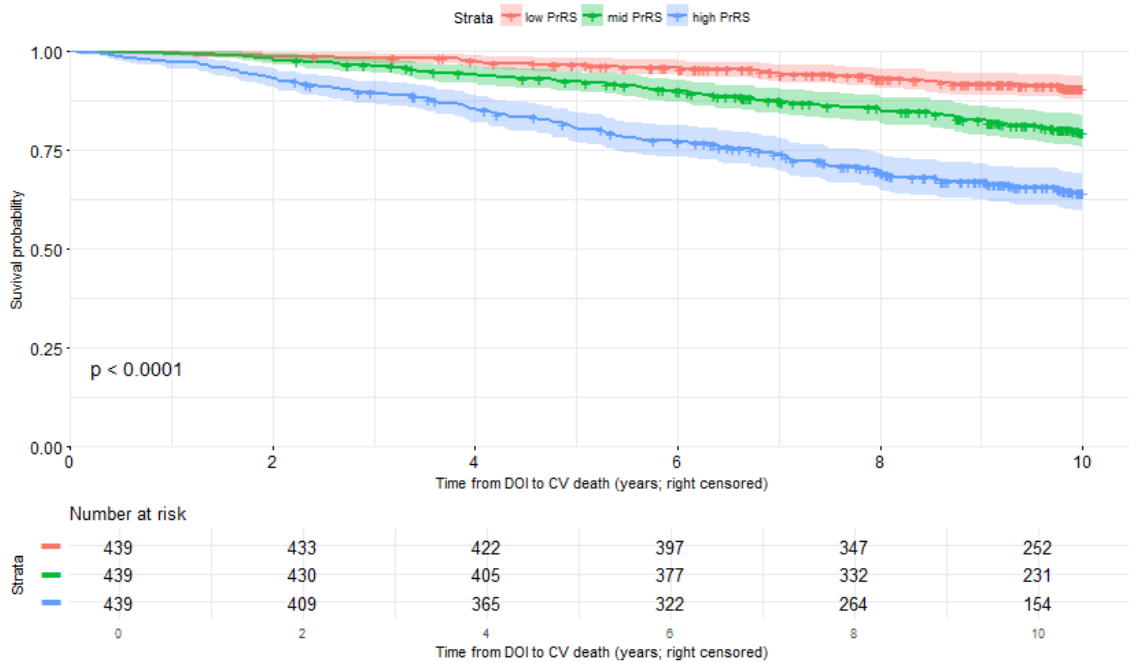
The Coxph regression results with only left-eye and only right-eye risk score predictions are provided in the Appendix A, Table A.11. They show significant results similar to individual-level predictions.

Rate of change of PRS

A follow-up period of 3 years was considered to compute Ω_{rate} . Individuals without at least one follow-up visit after the baseline or if they encountered the MACE or CV death in the follow-up period were excluded. After applying the exclusion criteria, 228 MACE from 1,127 individuals and 199 CV death events from 1,148 individuals were included in the test



(a) KM curves for MACE



(b) KM curves for CV Death

Fig. 6.4 KM curves of right-censored survival data for upper (high PRS), middle and lower (low PRS) tertiles of retinal PRS groups derived from individual-level prediction at baseline.

dataset. PRS for the last available retinal image in the follow-up period was computed using the DL model, trained on the baseline retinal images.

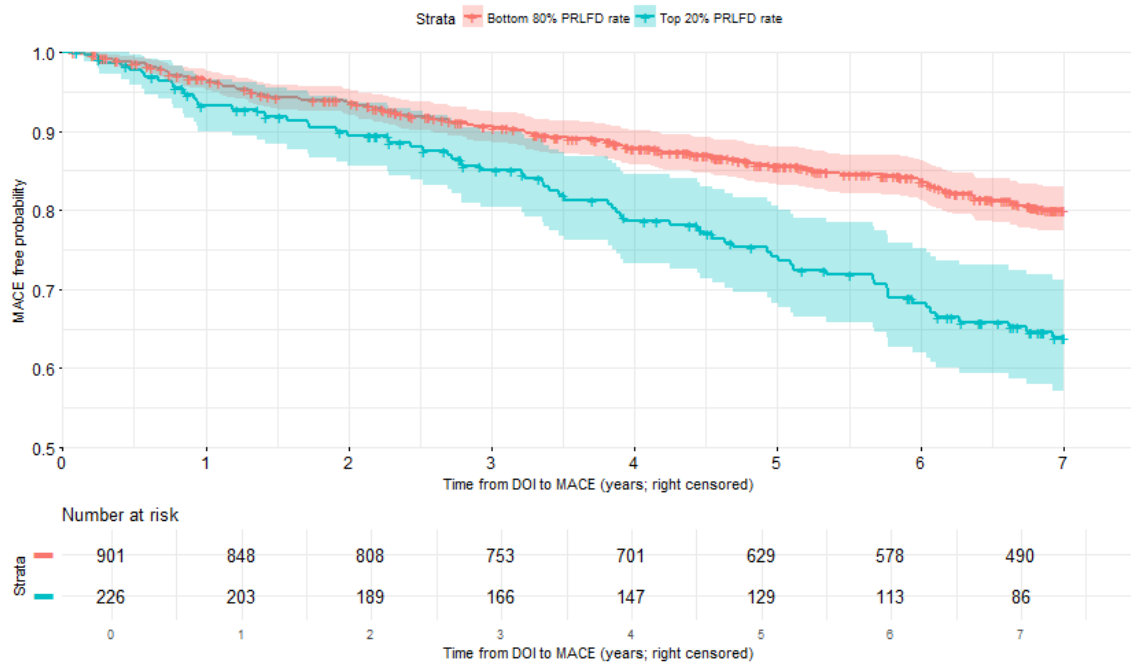
KM curves for MACE and CV death events are shown in Figure 6.5, computed for the top 20% and bottom 80% of the retinal Ω_{rate} derived from the individual-level prediction of PCE risk in the 3-year follow-up period. The range of Ω_{rate} in the top 20% and bottom 80% groups are (0.03 to 0.313) and (-0.337 to 0.03) respectively for both MACE and CV death events. Time from the last available retinal date of imaging (DoI) in the follow-up period to the event in years is considered as the time variable. For KM curves the time considered was up to 7 years follow-up from the last date of imaging, as shown in the x-axis. The KM curves show that the individuals in the top 20% Ω_{rate} group are significantly more associated with early MACE / CV death than those in the bottom 80% Ω_{rate} group.

The Coxph regression analysis adjusted for age at imaging, sex and GWPRS indicates that individuals in the top 20% Ω_{rate} group have a higher risk of developing MACE than those in the bottom 80% Ω_{rate} group by 50.4% (HR = 1.504, 95% CI = 1.122 – 2.016, P = 0.006, n=1,019, events=219). For CV death, individuals in the top 20% Ω_{rate} group have a higher risk of CV death than those in the bottom 80% Ω_{rate} group by 47.5% (HR = 1.475, 95% CI = 1.081 – 2.012, P = 0.014, n=1,112, events=190). The number of events was reduced from 228 to 219 in MACE and 199 to 190 in CV death because of missing values in GWPRS as this was included in the Coxph model.

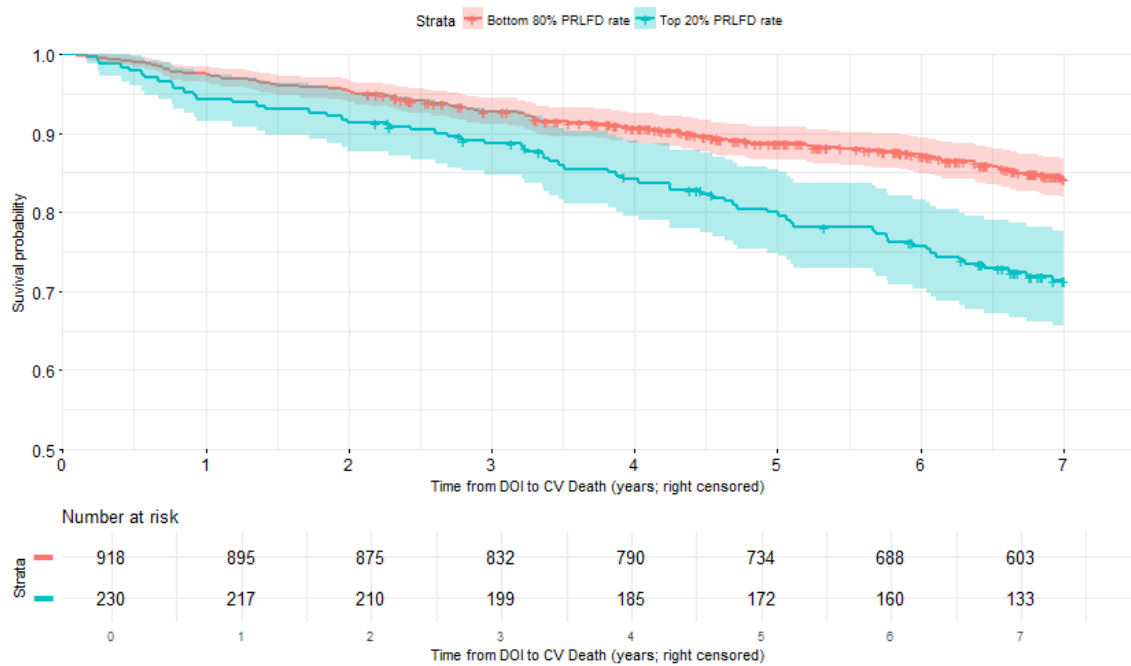
The KM curves and coxph regression results of PRLFD computed from only left-eye and only right-eye risk score predictions are provided in the Appendix A, Figure A.11 and A.12; Table A.12. The results are significant and are similar to individual-level predictions.

6.7 Discussions

This chapter reports several significant findings: the DL model can accurately predict CVD risk from fundus images in diabetic individuals; retinal imaging can provide valuable information in addition to a GWPRS for CVD risk; the DL predicted CVD risk score can predict MACE and CVD mortality independently of clinical CVD risk score and GWPRS



(a) KM curves for MACE



(b) KM curves for CV Death

Fig. 6.5 KM curves of right-censored survival data for top 20% and bottom 80% of the retinal Ω_{rate} groups computed from individual-level predictions for risk at last and first available retinal images.

for CVD; and the rate of change of predicted CVD risk from retinal images over time is independently associated with MACE and CVD mortality.

The use of DL models for predicting cardiovascular (CV) risk factors, disease outcomes such as major adverse cardiovascular events (MACE), chronic kidney disease (Chronic Kidney Disease (CKD)), and mortality has been well-established in the literature. The findings presented in this chapter are consistent with previous studies that have demonstrated the ability of DL to predict CV risk from retinal images. Notably, this chapter is the first to compare the predicted CVD risk from retinal images with that from the PCE ASCVD risk score. Clinical ASCVD risk computation requires multiple clinical measurements such as lipid profiles, which can be time-consuming and difficult to obtain for all patients. The DL predicted CVD risk score from retinal images could serve as a substitute since retinal images are relatively easy to acquire.

As presented in Table 6.4, similar to the age prediction results discussed in Section 5.3.4, individual-level prediction using the DL model improves its performance in predicting CVD risk compared to left- and right-eye image predictions per individual. However, further investigation is needed to determine the reason for this improvement. The coxph regression analysis was performed using only left-eye and only right-eye risk predictions, and the performance in stratifying MACE and CV death was found to be similar to that of individual-level prediction, but with a slightly lesser impact. This finding may be useful in clinical applications as notably large individuals do not have optimal retinal imaging during their eye screening [232].

Our study has some limitations of the experiments conducted. Firstly, the retinal image was downsized from 3500×2500 to 260×260 due to resource limitations of the available GPU. Using higher image sizes could potentially enhance the model's performance [98]. Secondly, the mean age of individuals in the GoDARTS dataset is 66 years, and further experiments could be conducted on individuals in different age groups, both younger and older. Additionally, as discussed in Section 5.6, further validation of the experiments is necessary for populations with existing disease conditions, different ethnicities, and from various geographical backgrounds.

6.8 Conclusions

In this chapter, we presented the results of our investigation on predicting clinical risk score and GWPRS for CVD from retina images using DL and how the predicted clinical risk score stratifies individuals with MACE and CV death within 10 years from the date of retinal imaging.

The results suggest that the retina contains information related to PCE ASCVD risk score but not to GWPRS for CVD. This discovery implies that the retina may offer significant insights into the risk of CVD that are mostly distinct from those provided by a GWPRS.

Using the individual-level PRS in test data, the Coxph regression analysis adjusted for age, sex, and GWPRS, shows statistically significant associations with the risk of developing 10 years MACE and risk of 10 years CV death. The Coxph regression analysis adjusted for age at imaging, sex, and GWPRS indicates that individuals in the top 20% Ω_{rate} (refer to Section 6.6.4, Rate of change of PRS) group have a higher risk of developing MACE and higher risk of CV death than those in the bottom 80% Ω_{rate} group. The associations were also assessed using CV risk score predictions from only right retinal images and only left retinal images and it was observed that the associations still remain significantly similar to individual-level predictions.

We conclude that the retina contains information useful for predicting CV risk score and it could, pending confirmation from further investigation and replication in other cohorts, be used for diagnosing disease conditions associated with CVD. In Chapter 7 we present our analysis for predicting the risk of 5-year MACE from retinal images along with image data converted from tabular data using DL, including whether this could improve the predictive power of the DL models.

Chapter 7

Converting tabular data to images for deep learning

7.1 About this chapter

This chapter introduces Tabular data to Image (T2I) conversion, a recent approach for converting tabular data into grayscale image data. This provides homogeneous data (images) to DL networks instead of heterogeneous inputs (text and images). The ultimate purpose is to use multi-modal data, i.e., retinal images and spreadsheets from Genetics of Diabetes Audit and Research in Tayside Scotland (GoDARTS), to stratify Major Adverse Cardiovascular Event (MACE) events within 5 years from the date of the first retinal images available.

We describe the materials and methods used for this experiment and present Deep Learning (DL) model performance results. The findings on the optimal feature ordering in the tabular data are also discussed.

7.2 Introduction and motivation

In data science, data is broadly categorized into *structured data* (tabular or spreadsheet), *unstructured data* (like image, video, speech, text), and *semi-structured data* (like JSON, XML). As an example, consider the Kaggle [\[233\]](#) online community of data scientists and

machine learning practitioners. Kaggle allows users to publish and use datasets, explore and build models in a web-based data-science environment, and take part in online competitions with well-specified tasks [233]. According to Anthony Goldbloom, founder, and CEO of Kaggle, winning techniques can be divided into two categories depending on the type of data used [234]: Convolutional Neural Network (CNN)s and Recurrent Neural Network (RNN)s perform best with unstructured data [234], but not with structured data, with which models like XGBoost [235] win.

Researchers have tried using One-dimensional (1-D) embedding to address Tabular data Machine Learning (TML) tasks by implementing RNNs or CNNs [236, 237]. 1-D embeddings such as word2vec [238], GLoVe [239], fastText [240], ELMO [241], BERT [242], and Open AI GPT [243] are mainly related to Natural Language Processing (NLP) tasks. Recently, the Super Characters method [244] for text classification and SuperTML [245] for tabular data has shown that the Two-dimensional (2-D) embeddings of the text or characters achieve state-of-the-art results on large datasets. Briefly, both Super Characters [33] and SuperTML [245] first project the text or tabular features to a 2-D embeddings as an image, then train a CNN model for classification.

Work has been reported in genomics on converting non-image or tabular data into images. Lyu et al. [246] embedded the high dimensional RNA-Seq data of tumors into 2-D images and used a convolutional neural network to classify 33 tumor types. DeepInsight [247] converts non-image samples into a structured image, to leverage the potential of CNNs in non-image data. Buturovic et al. [248] developed and evaluated a novel method, TABular Convolution (TAC), for the classification of tabular data using CNNs by transforming tabular data to images and then classifying the images using CNNs. Image Generator from Tabular Data (IGTD) [249] was proposed very recently to transform gene expression profiles of Cancer Cell Lines (CCL) and molecular descriptors of drugs into image-based representations.

Broadly speaking, there are two paradigms to analyzing image data in medical imaging analysis. One is based on a *hand-crafted feature dictionary*. This approach extracts features from images decided *a priori* by experts, then applies statistics or machine learning to analyze

the data, e.g., estimate associations with clinical outcomes [250–253, 115]. The other applies *end-to-end DL*: it trains a DL network to generate the desired output directly from input images [23, 169, 162, 158]. In this approach, the data features relevant to the task at hand are identified automatically by the DL network.

In this chapter we present a third approach, integrating retinal images and images from tabular data, the latter specifying the values of lists of pre-determined features. Our conversion of tabular data to 2-D image data is based on SuperTML[245]. The aim of our work is to show the feasibility and performance of the technique to address a specific question, the stratification of 5-year MACE from the date of retinal imaging. Here, the demographic and clinical measurement features available in GoDARTS are used in the form of spreadsheets as numerical, ordinal, and categorical data and mapped to images according to specific rules, presented in the following. We explore whether the order of features during T2I approach has an impact on the DL model performance by implementing IGTD technique [249].

7.3 Materials

A subset of baseline retinal images were chosen from the individuals in the GoDARTS dataset who did not have a history of hospital admissions for Myocardial Infraction (MI) or stroke, as determined by the absence of International Classification of Diseases (ICD)-10 codes I21-I23 and I60-I63. Individuals who had both left and right eye retinal images available at baseline were included in this experiment. A total of 10,872 retinal images were used from 5,436 individuals. The individuals in the whole dataset were split randomly into three groups: 70% training, 10% validation, and 20% testing. The image distribution is shown in Table 7.1 for the whole cohort and data splits.

7.3.1 Outcome variables

The outcome variable in this experiment is a MACE within 5 years of retinal imaging. It is a binary event (1 for MACE within 5 years, 0 otherwise). For T2I conversion, the Cardiovascular (CV) risk features available in GoDARTS were used, namely age, sex,

Table 7.1 Retinal Image distribution of whole cohort and data splits for predicting 5-year MACE.

	Overall	Train	Validation	Test
Participants	5,436	3,805	543	1,088
Images	10,872	7,610	1,086	2,176
Right eye Images (%)	5,436 (50%)	3,805 (50%)	543 (50%)	1,088 (50%)

Diastolic Blood Pressure (DBP), Systolic Blood Pressure (SBP), Body Mass Index (BMI), Glycated Haemoglobin (GH), High-Density Lipoprotein (HDL), Total Cholesterol (TC), Tryglicerides (Trig). Details on these features are available in Section 3.3.1. The Genome-Wide Polygenic Risk Scores (GWPRS) feature was also included. More details on GWPRS are available in Chapter 6. The dataset characteristics for the whole data as well as data splits are shown in Table 7.2.

Table 7.2 Baseline characteristics of the dataset for 5-years MACE prediction in GoDARTS; n = individuals used for the feature.

Feature (units)	sex	Overall										Train					Validation					Test			
		n		Mean (std)		Range		IQR		n		Mean (std)		Range		IQR		n		Mean (std)		Range		IQR	
Age at imaging (years)	All	5,436	67.77(10.71)	30.97-96.45	60.78-75.7	3,805	67.77(10.73)	30.97-96.45	60.86-75.71	543	67.61(10.87)	37.13-91.09	60.21-75.82	1,088	67.81(10.55)	33.53-92.61	60.92-75.54								
	Male	3,045	67.35(10.47)	33.49-93.59	60.48-75.29	2,141	67.45(10.45)	33.49-93.59	60.7-75.3	306	67.47(9.95)	41.4-91.09	60.44-74.84	598	66.93(10.77)	35.61-92.61	60.11-75.19								
	Female	2,391	68.29(10.99)	30.97-96.45	61.29-76.15	1,664	68.19(11.07)	30.97-96.45	61.19-76.21	237	67.79(11.96)	37.13-90.21	59.13-76.69	490	68.89(10.18)	33.53-92.08	62.29-75.84								
DBP (mmHg)	All	5,436	75.96(7.81)	48.0-108.5	70.5-81.0	3,805	75.95(7.69)	51.0-108.0	71.0-81.5	543	76.38(8.14)	53.0-106.0	71.0-81.0	1,088	75.81(8.08)	48.0-108.5	70.0-81.0								
	Male	3,045	76.52(7.94)	48.0-108.5	71.0-82.0	2,141	76.37(7.81)	52.0-108.0	71.0-81.5	306	77.16(7.99)	54.0-101.0	72.0-82.0	490	76.73(8.34)	48.0-108.5	70.5-82.0								
	Female	2,391	75.26(7.59)	51.0-106.0	70.0-80.0	1,664	75.41(7.49)	51.0-99.0	70.0-80.0	237	75.37(8.23)	53.0-106.0	70.0-80.0	490	74.7(7.62)	51.0-96.0	69.5-80.0								
SBP (mmHg)	All	5,436	138.89(11.99)	93.5-205.0	132.0-145.5	3,805	138.69(11.95)	95.0-205.0	131.0-145.0	543	139.6(12.24)	105.0-190.0	132.0-146.0	1,088	139.25(12.0)	93.5-192.0	132.0-146.0								
	Male	3,045	138.46(12.0)	93.5-205.0	131.0-145.0	2,141	138.19(11.97)	95.0-205.0	131.0-145.0	306	139.5(12.31)	105.0-190.0	132.5-146.0	598	138.9(11.93)	93.5-180.5	132.0-146.0								
	Female	2,391	139.44(11.95)	98.0-198.5	132.0-146.0	1,664	139.33(11.88)	98.0-198.5	132.0-146.0	237	139.72(12.16)	110.0-189.5	132.0-145.5	490	139.68(12.09)	107.0-192.0	132.0-146.0								
BMI (Kg/m ²)	All	5,436	31.4(5.96)	15.2-59.1	27.27-34.61	3,805	31.5(6.03)	15.2-59.1	27.3-34.85	543	30.97(5.72)	17.1-56.7	26.9-34.18	1,088	31.27(5.82)	18.7-58.05	27.28-34.05								
	Male	3,045	30.76(5.31)	15.95-59.1	27.12-33.66	2,141	30.74(5.33)	15.95-59.1	27.1-33.7	306	30.81(4.84)	19.27-46.73	27.45-33.69	598	30.81(5.48)	19.08-58.05	27.11-33.55								
	Female	2,391	32.21(6.61)	15.2-57.7	27.5-36.04	1,664	32.47(6.7)	15.2-57.7	27.61-36.4	237	31.18(6.7)	17.1-56.7	26.67-34.71	490	31.83(6.17)	18.7-54.21	27.63-35.54								
GH (%)	All	5,436	7.5(1.13)	6.08-12.7	6.7-8.05	3,805	7.5(1.13)	6.08-12.7	6.7-8.01	543	7.52(1.16)	6.1-12.55	6.65-8.05	1,088	7.49(1.13)	6.08-12.7	6.7-8.1								
	Male	3,045	7.49(1.1)	6.08-12.7	6.7-8.0	2,141	7.48(1.09)	6.08-12.7	6.7-8.0	306	7.48(1.11)	6.1-12.55	6.7-7.99	598	7.52(1.14)	6.08-12.7	6.7-8.14								
	Female	2,391	7.51(1.16)	6.08-12.7	6.65-8.07	1,664	7.52(1.17)	6.08-12.7	6.65-8.06	237	7.57(1.21)	6.1-12.25	6.65-8.15	490	7.46(1.11)	6.1-12.3	6.65-8.0								
HDL (mmol/L)	All	5,436	1.33(0.34)	0.23-3.71	1.1-1.47	3,805	1.33(0.34)	0.23-3.71	1.1-1.46	543	1.34(0.37)	0.56-3.34	1.11-1.47	1,088	1.32(0.33)	0.23-3.55	1.1-1.46								
	Male	3,045	1.26(0.31)	0.23-3.55	1.05-1.37	2,141	1.26(0.31)	0.39-3.42	1.05-1.37	306	1.26(0.31)	0.56-2.57	1.07-1.34	598	1.27(0.33)	0.23-3.55	1.06-1.39								
	Female	2,391	1.41(0.35)	0.23-3.71	1.2-1.57	1,664	1.42(0.35)	0.23-3.71	1.2-1.58	237	1.45(0.41)	0.58-3.34	1.19-1.62	490	1.38(0.33)	0.63-3.03	1.18-1.53								
TC (mmol/L)	All	5,436	4.34(0.84)	1.76-10.89	3.79-4.77	3,805	4.33(0.83)	1.76-10.73	3.8-4.75	543	4.37(0.9)	2.13-10.89	3.82-4.8	1,088	4.35(0.84)	2.38-9.24	3.76-4.82								
	Male	3,045	4.22(0.83)	1.76-10.89	3.67-4.65	2,141	4.21(0.81)	1.76-10.73	3.68-4.64	306	4.23(0.88)	2.13-10.89	3.72-4.67	598	4.25(0.87)	2.38-9.24	3.61-4.72								
	Female	2,391	4.5(0.83)	2.41-9.59	3.96-4.92	1,664	4.5(0.83)	2.41-9.59	3.98-4.9	237	4.55(0.9)	2.77-7.71	3.96-5.01	490	4.47(0.79)	2.68-7.54	3.91-4.95								
Trig (mmol/L)	All	2,919	2.33(1.84)	0.31-30.61	1.31-2.75	2,053	2.32(1.79)	0.32-30.61	1.34-2.73	298	2.35(1.78)	0.4-14.28	1.3-2.72	568	2.36(2.01)	0.31-21.12	1.21-2.84								
	Male	1,646	2.42(2.08)	0.31-30.61	1.31-2.81	1,157	2.37(2.0)	0.4-30.61	1.33-2.77	175	2.49(1.97)	0.5-14.28	1.4-2.78	314	2.55(2.42)	0.31-21.12	1.24-2.92								
	Female	1,273	2.23(1.45)	0.32-14.51	1.31-2.68	896	2.27(1.49)	0.32-14.51	1.35-2.68	123	2.15(1.46)	0.4-9.7	1.2-2.59	254	2.13(1.28)	0.47-8.82	1.2-2.56								
GWPRS	All	5,436	6.98(0.6)	4.4-9.26	6.66-7.37	3,805	6.98(0.59)	4.4-8.74	6.69-7.37	543	6.94(0.61)	4.75-8.35	6.6-7.35	1,088	6.97(0.63)	4.66-9.26	6.64-7.39								
	Male	3,045	6.97(0.61)	4.4-9.26	6.64-7.37	2,141	6.98(0.6)	4.4-8.74	6.67-7.37	306	6.94(0.61)	5.11-8.35	6.59-7.32	598	6.96(0.67)	4.66-9.26	6.59-7.41								
	Female	2,391	6.99(0.59)	4.65-8.63	6.7-7.37	1,664	6.99(0.58)	4.65-8.61	6.71-7.37	237	6.95(0.61)	4.75-8.31	6.63-7.36	490	6.99(0.58)	4.95-8.63	6.7-7.35								

Categorical features			
	n	events count (% prop.)	events count (% prop.)
sex_male(%)	3,805	2,141 (56.27%)	306 (56.35%)
5-years MACE(%)	3,805	554 (14.66%)	96 (17.68%)
	2,141	335 (15.65%)	58 (18.95%)
	1,664	219 (13.16%)	38 (16.03%)
	3,805	543	1,088
	543	598 (54.96%)	147 (13.51%)
	306	598	97 (16.22%)
	237	50 (10.2%)	

7.4 Methods

This section describes the procedure for T2I conversion, image pre-processing, the deep learning architectures used with left- and right-eye retinal images and T2I data, the machine learning methods using only tabular data and the metrics adopted for evaluation.

The image pre-processing steps applied on the retinal images are described in Section 3.3.2.

7.4.1 Tabular data to image conversion

The conversion of tabular data to the image was initially inspired by SuperTML[245]. Later we developed a novel, independent approach of T2I conversion, composed of two steps.

In the first step, the raw tabular (spreadsheet) numerical data is scaled between 0 and 1 using any scaling technique, here we used min-max normalization, Equation (7.1). Below, x' is the new scaled cell value, x the original cell value from column X , $max(X)$ is the maximum value of the column X , and $min(X)$ is the minimum value of the column X .

$$x' = \frac{x - min(X)}{max(X) - min(X)} \quad (7.1)$$

In the second step, the normalized features are projected into image space. An empty 2-D image with a sample size of width \times height is created and divided into a grid of equal rectangles, one per feature. If it is not possible to evenly distribute the image area among all the available features then dummy features are used; for example, in case of an odd number of features and the image space has to be divided into an even number of equal spaces then one more dummy feature can be used to make the projection of features on to the image space equally distributed. Allocating the same amount of space to all the features makes the method independent of feature importance (i.e. a specific feature that has a larger effect on the model predictions) as opposed to SuperTML_VF algorithm in [245], where, the feature importance in the tabular data is calculated by other Machine Learning (ML) methods prior

to converting tabular data into an image. The rectangle allocated to a feature in the image is given a gray level depending on the feature value. The gray level is computed by multiplying the normalized feature value by 255, where 0 is black and 1 is white. We consider 8-bit images so that 255 is the maximum pixel intensity (white). This method can be extended to categorical and ordinal features too by mapping the feature labels to a set of integers prior to the first step of scaling the features.

The algorithm for converting tabular data to image is described by Algorithm 1. A block diagram describing the steps involved in T2I conversion is shown in Figure 7.1. An example of converting tabular data with 6 columns to a gray-level image is shown in Figure 7.2. The T2I conversion algorithm gives a 2-D gray image and it can be converted to a Three-dimensional (3-D) image by simply copying the pixel values to the third dimension so that the pre-trained weights of ImageNet [85] can be used during training a DL model as described in Section 7.4.3.

Algorithm 1 Procedure for T2I conversion

Input Tabular data and image dimensions (width and height)

Output Grayscale images from tabular data

- 1: Normalize all the tabular data between 0 and 1
 - 2: Create an image of dimensions (width×height) with zero pixel values
 - 3: **for** each row in tabular data **do**
 - 4: **for** each column in the row **do**
 - 5: fill the dedicated feature space in the image area with the value 255*feature value
 - 6: **end for**
 - 7: **end for**
-



Fig. 7.1 Block diagram of T2I conversion.

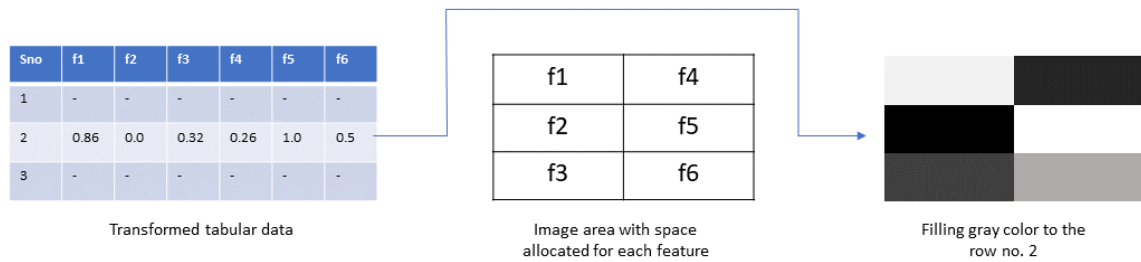


Fig. 7.2 An example showing conversion of a sample tabular data with 6 features to a 2D gray image. Left: Sample tabular data. Middle: Dedicated feature space in the image area. Right: Tabular data to image conversion for row no. 2

The SuperTML algorithm [245], projects features in tabular data into 2-D embeddings by drawing the tabular text or character features directly to image space using different font sizes. We believe that this algorithm can bring more questions on picking the appropriate font and font sizes and also image space can go wasted due to the unused black background region. Our proposed approach of T2I conversion, overcomes the issues of unused image space by equally distributing the image space among the tabular features; and the issue with fonts, and font sizes by assigning the gray level to the features.

7.4.2 IGTD implementation

CNNs are mainly suitable for analyzing the data with spatial or temporal dependencies [254, 255]. But there are no spatial relationships among the features in tabular data like a spreadsheet, as the order of rows and columns is arbitrary. Inspired by the work by Zhu et al. [249], we implemented their IGTD algorithm to investigate the importance of feature ordering in our experiment.

The IGTD algorithm searches for an optimized feature ordering by minimizing the difference between the ranking of distances between features and the ranking of distances between their assigned pixels in the image.

7.4.3 Deep learning architecture and training

We experimented with different deep-learning architectures. The baseline architecture was EfficientNet-B2 [98] (see Section 5.2.2). The architectures were modified so that the retinal image and T2I (also referred to as *feature image*) could be given as input to a single DL model. The different DL models tested are described in the following:

1. *DL model 1*: A plain modified EfficientNet-B2. It has all the convolutional layers from the original EfficientNet-B2 followed by a Global Average Pooling (GAP) layer and a single sigmoid layer, as we want to predict events over 5 years MACE. This model takes a 2-D image as input of dimension 260x260 pixels. The input image can be a retina or a featured image. The DL model 1 architecture is shown in Figure 7.3. This model has 7.7 million trainable parameters in total.
2. *DL model 2*: This model takes both the retinal image and a featured image as input. Two DL model 1s are concatenated after the GAP layer followed by a sigmoid node as output. One of the two DL model 1s takes a retinal image as input, and the other feature image. The DL model 2 architecture is shown in Figure 7.4. This model has 15.4 million trainable parameters in total.
3. *DL model 3*: This model takes both the retinal image and feature image as input. A dropout layer with a dropout rate of 0.2 is introduced between the concatenation layer and the sigmoid layer in DL model 2. The DL model 3 architecture is shown in Figure 7.5. This model has 15.4 million trainable parameters in total.
4. *DL model 4*: This model takes both the retinal image and feature image as input. A Fully Connected (FC) layer with 100 nodes is introduced between the concatenation layer and sigmoid layer in DL model 2. The DL model 4 architecture is shown in Figure 7.6. This model has 15.7 million trainable parameters in total.
5. *DL model 5*: This model takes both the retinal image and feature image as input. A dropout layer with a dropout rate of 0.2 is introduced before and after the FC layer in

DL model 4. The DL model 5 architecture is shown in Figure 7.7. This model has 15.7 million trainable parameters in total.

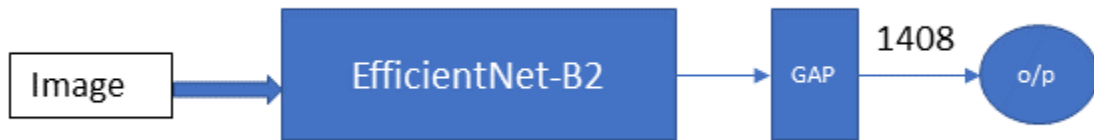


Fig. 7.3 DL model 1: A modified EfficientNet-B2 with image as input and sigmoid output. The GAP layer gives a vector of length 1,408.

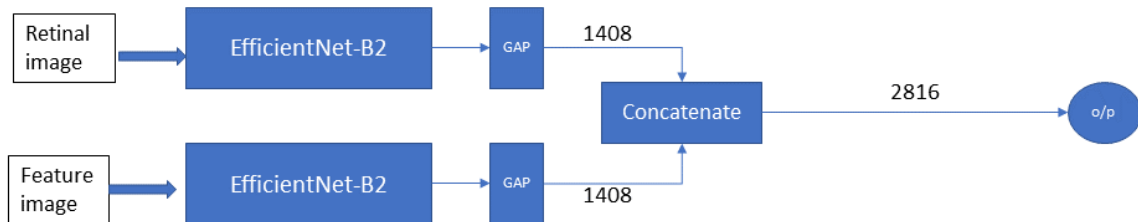


Fig. 7.4 DL model 2: A multi-modal DL architecture with retinal image and T2I image as input and sigmoid output. The GAP layer results in a vector of length 1,408 and the concatenation layer outputs a vector of length 2,816.

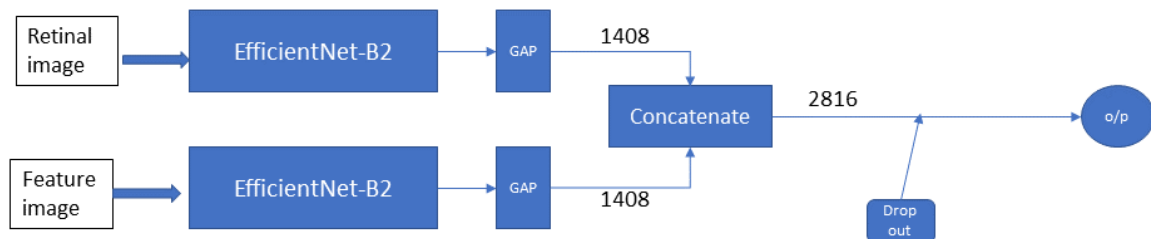


Fig. 7.5 DL model 3: Modified version of DL model 2 with an additional dropout layer before the output layer.

Individuals were split randomly into three groups: 70% training, 10% validation, and 20% testing. The dataset characteristics at baseline in all splits are shown in Table 7.2. The training strategy followed is described in Section 5.2.2. The training specifications for DL models 1 to 5 described above are summarized in Table 7.3.

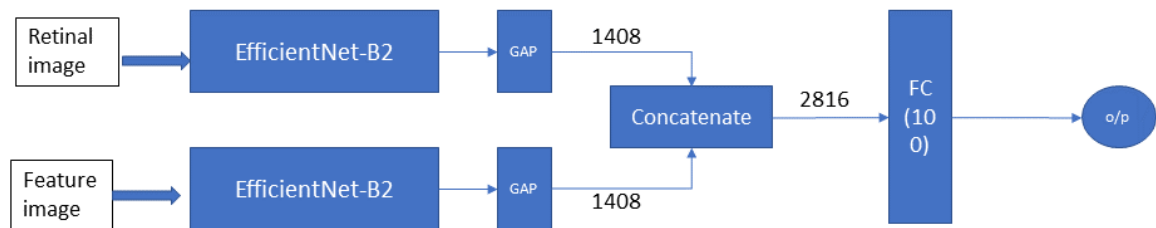


Fig. 7.6 DL model 4: Modified version of DL model 3 by replacing dropout layer with FC layer of 100 nodes.

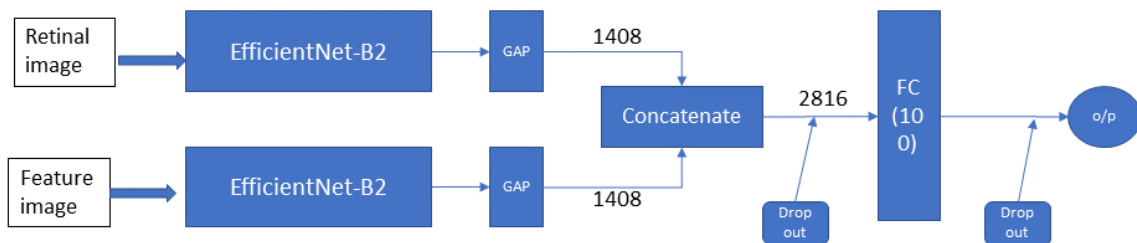


Fig. 7.7 DL model 5: Modified version of DL model 4 by introducing dropout layer before and after FC layer.

7.4.4 Machine learning methods and training

Apart from the DL models, two well-known machine learning models were used, logistic regression [256] and XGBoost [235]. These models were trained on the tabular data described in Section 7.3 for predicting 5-year MACE so that results could be compared with those of DL models. Both machine learning models were implemented in Python using the sklearn package [257] with default parameter values [258, 259]. AUC was used as the evaluation metrics, defined in Section 5.2.2.

7.5 Results

As mentioned above, the features considered for this experiment from tabular data in GoDARTS are age, sex_male, eye (left retina(L) or right retina (R)), DBP, SBP, BMI, GH, HDL, TC and GWPRS.

A total of 5,436 individuals were obtained after applying the following criteria.

1. Include individuals with both left and right retinal images at baseline.

Table 7.3 Summary of training specifications for predicting 5-year MACE using retinal images.

Category	Specification
Input and Output	
Single image input	
DL architecture	modified EfficientNet-B2
Input	Color fundus image or 3-D feature image
Two image input	
DL architecture	Two modified EfficientNet-B2 concatenated
Input	Color fundus image and 3-D feature image
Input dimensions	260×260
Output	Sigmoid activation
Performance metric	Area Under Receiver Operating Characteristic (ROC) Curve (AUC)
Training	
Weight Initialization	ImageNet and random uniform
Epochs	50
Batch size	32 for DL model 1 16 for DL model 2, 3, 4 and 5
Loss function	binary cross-entropy
Optimizer	Nadam
Learning rate	0.001 reduced by a factor of 0.1
Avoid overfitting	
Early stopping	on validation loss
Weights	Best validation loss

2. Exclude individuals with any missing values in any of the features considered.

Therefore there are exactly two retinal images per individual at baseline (left and right eye), totaling 10,872 images. For these individuals, data were split into three groups as usual; 70% for training (7,610 images from 3,805 individuals), 10% validation (1,086 images from 543 individuals), and 20% testing (2,176 images from 1,088 individuals). Dataset characteristics of the whole data and the data splits are provided in Table 7.2.

7.5.1 Tabular data to image conversion

The tabular data with the 10 features (10f) (age, sex_male, eye, DBP, SBP, BMI, GH, HDL, TC and GWPRS) and 9 features (9f) (excluding GWPRS from 10f) from 5,436 individuals were converted to images using the procedure described in Section 7.4.1.

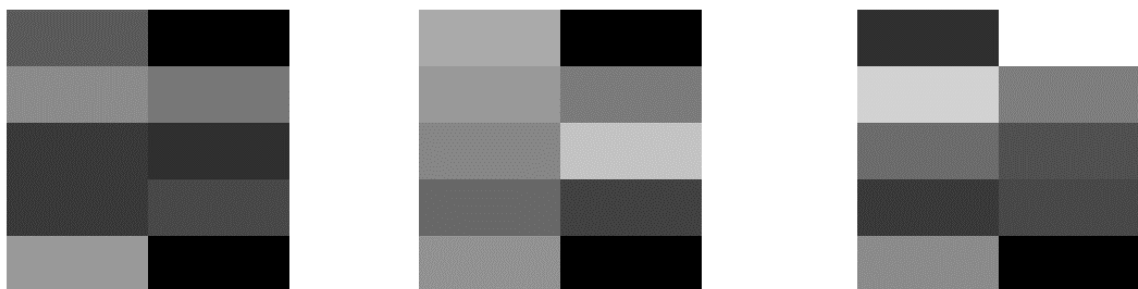
As described in Section 7.4.1, the data was first normalized between 0 and 1 using min-max normalization. Then the normalized tabular data is projected to the image area by equally distributing the features space into 5 rows and 2 columns. The projection of features to the image is done in a row first fashion i.e. to fill the gray color of the respective feature in the first row of all columns and then in the second row of all the columns and so on. The T2I conversion of 10 features and 9 features was done independently to explore the impact of GWPRS in predicting 5-years MACE in combination with the retinal images using DL models. For converting tabular data with 9 features to image data, a dummy feature with a default feature value of 0 was used (see the second step in Section 7.4.1). An example of T2I conversion with a sample subset data of 10 features from the whole data is shown in Figure 7.8.

Age_at_imaging	sex_male	dbp	sbp	bmi	gh	hdl	tc	GWPRS	eye
54.36	0.00	81.00	146.00	25.30	7.35	1.02	4.37	7.30	R
74.70	0.00	84.50	148.00	38.62	11.15	1.64	4.16	7.21	R
43.27	1.00	98.00	149.00	34.22	8.20	1.04	4.41	7.07	R

(a) Sample subset of original data

Age_at_imaging	sex_male	dbp	sbp	bmi	gh	hdl	tc	GWPRS	eye
0.36	0.00	0.55	0.47	0.23	0.19	0.23	0.29	0.60	0
0.67	0.00	0.60	0.49	0.53	0.77	0.41	0.26	0.58	0
0.19	1.00	0.83	0.50	0.43	0.32	0.23	0.29	0.55	0

(b) Sample subset of normalized data



(c) T2I conversion for the above tabular data

Fig. 7.8 An example of T2I conversion with a sample subset data.

7.5.2 Optimal feature ordering

The proposed T2I conversion procedure generated the image data in the same order as the features that appear in the input spreadsheet. The IGTD algorithm is implemented [249] so to identify the optimal feature ordering (i.e., maximizing performance in the target task) thus investigating the importance of the relative positions of feature patches in the grey-level image. As per the suggestions by the authors of IGTD [249], the parameter value is used as described in Table 7.4. Figure 7.9 illustrates the IGTD algorithm applied to the GoDARTS tabular data, which consists of 9 features. The Euclidean distance between all pairs of CV risk factors is represented in the rank matrix shown in Figure 7.9a. The rank matrix of Euclidean distance between all pairs of pixels in a 3 by 3 image is shown in Figure 7.9b. After optimization and rearrangement of features, the feature distance rank matrix is shown in Figure 7.9c, and the change in error during the optimization process is shown in Figure 7.9d. For further details about the IGTD algorithm, refer to [249].

Table 7.4 Training specifications for IGTD.

Category	Specification
Maximum steps	10,000
Validation steps	100
Feature distance measure	Euclidean
Image distance measure	Euclidean
Error function	Absolute or Square

The initial feature ordering with 9 features was age, sex_male, DBP, SBP, BMI, GH, HDL, TC, eye and dummy and for 10 features was age, sex_male, DBP, SBP, BMI, GH, HDL, TC, GWPRS and eye. Table 7.5 describes the optimal feature order after implementing the IGTD algorithm on the tabular data with 10 features and 9 features independently.

Figure 7.10 shows some samples T2I with 10 features data with initial feature ordering and after IGTD optimized feature ordering with Square error function. In Figure 7.10b we observe a pattern: grey levels are sorted from dark to bright when moving from top to bottom. This represents that the IGTD optimized feature ordering transforms T2I data to maintain a spatial relationship between features that are most suitable for CNNs.

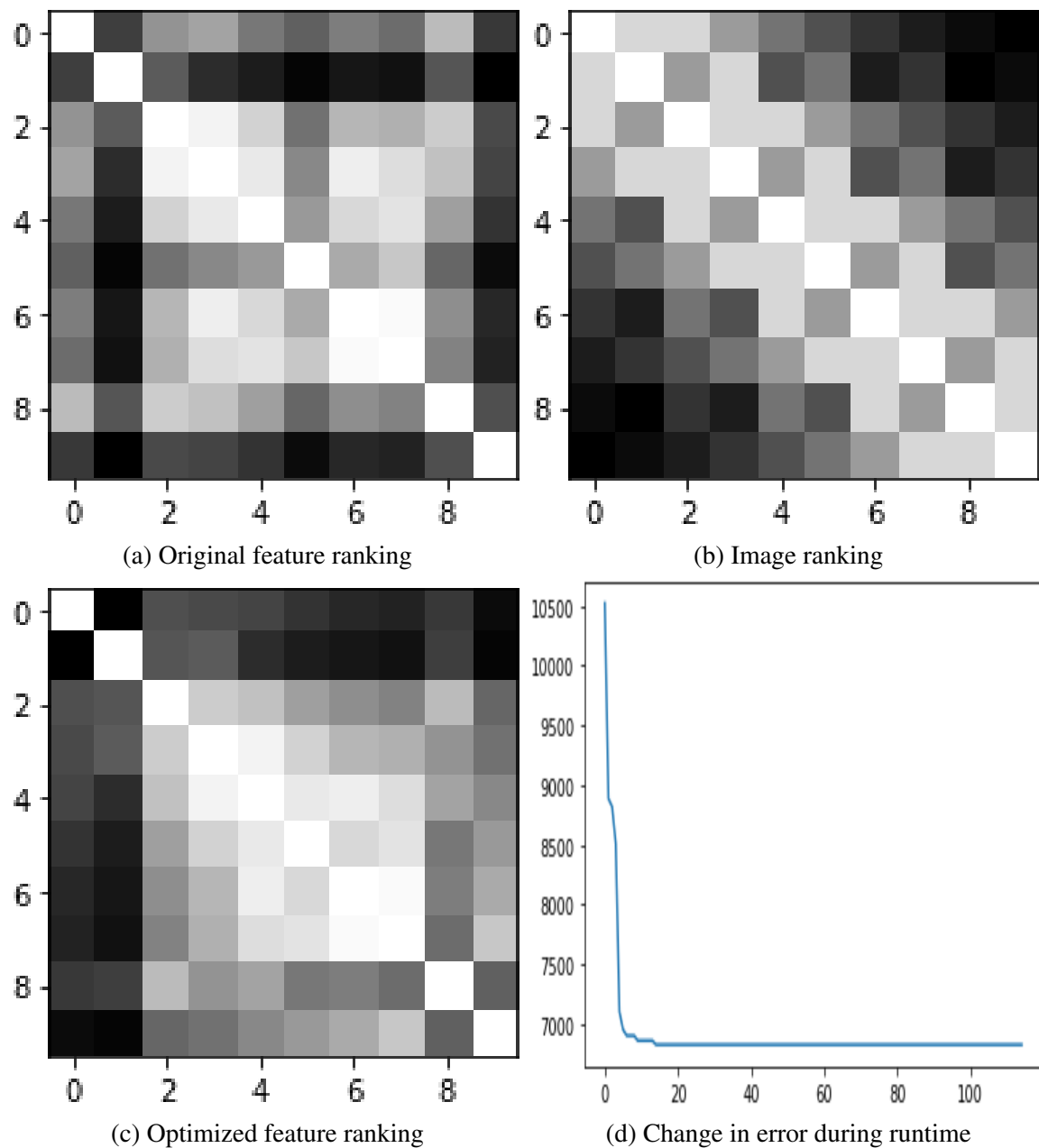


Fig. 7.9 An illustration of IGTD algorithm on the GoDARTS tabular data with square error function.

7.5.3 ML methods for MACE

The two well-known ML models, logistic regression and XGBoost, are used for the classification of 5-years MACE from the baseline. The tabular data from the training split (from 3,805 individuals) was used to train the models and data from the test split (1,088 individuals)

Table 7.5 Optimal feature ordering with IGTD on 9 features and 10 features.

IGTD error function	Category	Features
NA	Initial feature order (9f)	age, sex_male, DBP, SBP, BMI, GH, HDL, TC, eye and dummy
	Initial feature order (10f)	age, sex_male, DBP, SBP, BMI, GH, HDL, TC, GWPRS and eye
Square	feature order (9f)	sex_male, age, DBP, SBP, BMI, HDL, GH, TC, dummy and eye
	feature order (10f)	eye, sex_male, age, GWPRS, SBP, DBP, HDL, BMI, TC and GH
Absolute	feature order (9f)	eye, age, DBP, SBP, BMI, HDL, TC, GH, sex_male and dummy
	feature order (10f)	sex_male, eye, GWPRS, age, DBP, SBP, BMI, HDL, TC and GH

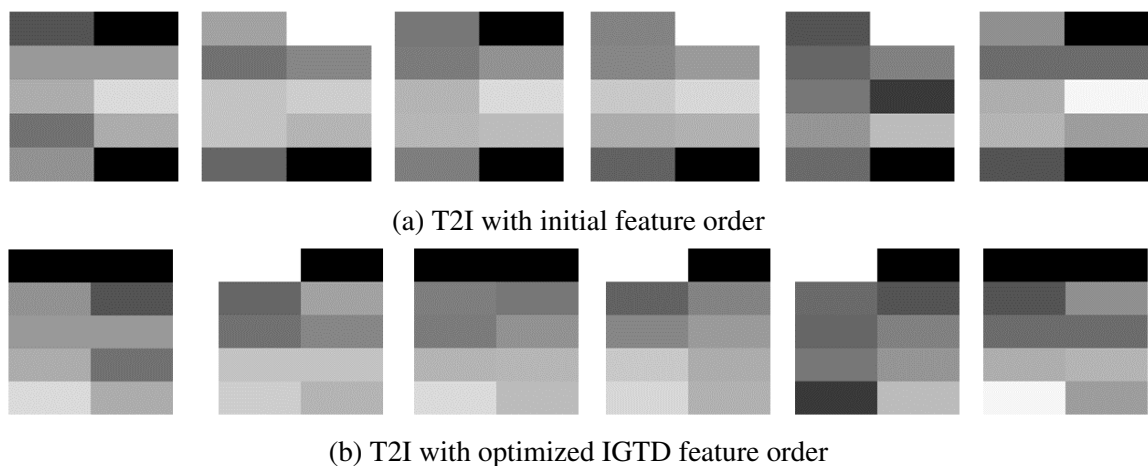


Fig. 7.10 An example of T2I with feature ordering.

was used to evaluate the models. Normalized data were used for both training and testing the models. Table 7.6 shows the AUC performance of the two models and it appears that there is an increase in model's AUC with 10 features when compared to 9 features by $\approx 2.5\%$ in both logistic regression and XGBoost classifier. Comparatively XGBoost classifier gives better performance than the logistic regression for the classification of 5-years MACE. Table 7.6 also signifies that adding GWPRS feature to the data increases the predictive power of both logistic regression and XGBoost classifier.

7.5.4 Retina and T2I for MACE

In total, 10,872 retinal images from 5,436 individuals were available, with one left-eye image and one right-eye image per individual. 10,872 T2Is were generated using the T2I algorithm which is equal to the total number of retinal images from the tabular data of

Table 7.6 ML models performance for classification of 5-years MACE in the test dataset. 9 features = age, sex_male, DBP, SBP, BMI, GH, HDL, TC, eye. 10 features = 9 features + GWPRS.

Input category (normalized)	AUC (%)	
	Logistic regression	XGBoost classifier
9 features	69.14	69.82
10 features	71.78	72.29

5,436 individuals. The DL models described in Section 7.4.3 were trained for predicting 5-years MACE from the date of retinal imaging with 70% of data (7,610 images from 3,805 individuals), validated with 10% of data (1,086 images from 543 individuals) and the model performance is evaluated using 20% of data (2,176 images from 1,088 individuals).

The DL model 1 takes only one image as input and it provides a prediction score between 0 to 1 to the individuals developing MACE within 5 years of retinal imaging. The input image can be a retinal image or a T2I. Table 7.7 shows the AUC for the DL model 1 on the test dataset. The row with IGTD error function as 'NA' represents the T2I data with the initial feature order considered. Not surprisingly the prediction of 5-years MACE is improved by using the T2I with CV risk factors as features when compared to using the retinal images alone. It is noted that by adding additional information through GWPRS feature, the prediction is further improved.

Table 7.7 DL model 1 performance for classification of 5-years MACE in the test dataset. 9f = age, sex_male, DBP, SBP, BMI, GH, HDL, TC, eye. 10f = 9f + GWPRS.

IGTD error function	Input images category	DL model 1 AUC (%)
NA	Only retinal images	66.78
	Only T2I (9f)	68.01
	Only T2I (10f)	70.46
Square	Only T2I (9f)	65.37
	Only T2I (10f)	67.03
Absolute	Only T2I (9f)	67.92
	Only T2I (10f)	69.17

To assess the importance of feature ordering, we shuffled randomly the feature order and generated T2I data for the whole dataset; trained, validated and tested DL model 1. Shuffling is performed 200 times independently on the complete dataset of 5,436 individuals. An example of the representations resulting from shuffling the features of a single row of data in 10 different ways and generating T2I is shown in Figure 7.11. The mean and standard deviation of AUC(%) in the test dataset from training the DL model 1 with 10 features over 200 random shuffles independently was 70.35 and 5.03 respectively. By noting the significant amount of variation in the model performance with the feature ordering, the IGTD algorithm is implemented to investigate optimal feature ordering.

Table 7.7 shows the DL model 1 performance with only T2I generated from the optimized feature order using IGTD algorithm (with square and absolute error function) on tabular data with 9 features and 10 features. It does not show any significant improvement in model performance with the IGTD optimized feature ordering but performance is improved when adding GWPRS information (10 features) to the T2I compared to 9 features.

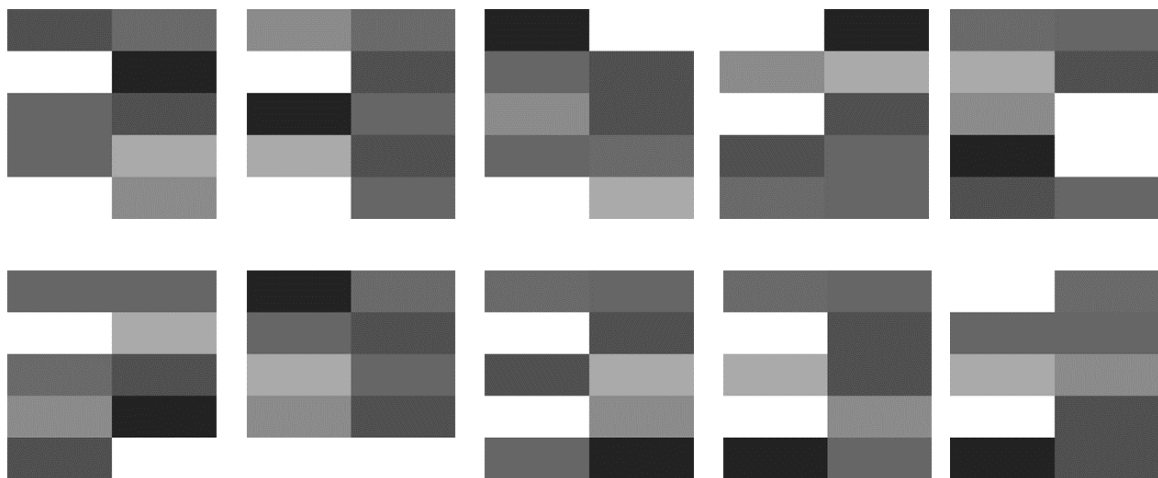


Fig. 7.11 A T2I example representation of a single row data with randomly shuffling the features in 10 different ways.

We further investigated with different DL models, which take both retinal image and T2I as input for predicting 5-years MACE. Table 7.8 shows the performance of DL models 2, 3, 4, and 5. Results suggest that there is no significant improvement with the IGTD optimized feature ordering when compared to the initial feature ordering. The performance is improved

with 10 features (GWPRS added) when compared to 9 features in all the categories of feature ordering and in all the different DL models. This shows that GWPRS contains additional information for 5-years MACE prediction compared to other 9 features of Cardiovascular Disease (CVD) risk factors. Table 7.8 also shows that the DL model with dropout layer i.e. DL model 3 and DL model 5 gave slightly higher performance but it is not consistent.

Table 7.8 DL model 2, 3, 4, and 5 performance for classification of 5-years MACE in the test dataset. 9f = age, sex_male, DBP, SBP, BMI, GH, HDL, TC, eye. 10f = 9f + GWPRS.

IGTD Error function	Input images	AUC (%)			
		DL Model 2	DL Model 3 (dropout)	DL Model 4	DL Model 5 (dropout)
NA	Retina + T2I (9f)	68.75	70.6	70.57	70.26
	Retina + T2I (10f)	70.55	72.54	71.95	71.43
Square	Retina + T2I (9f)	68.46	67.81	68.4	69.14
	Retina + T2I (10f)	71.01	71.12	70.48	71.21
Absolute	Retina + T2I (9f)	66.4	68.07	67.63	68.18
	Retina + T2I (10f)	71.2	69.78	69.48	70.18

7.6 Discussions

Dealing with large amounts of tabular data, such as genetic data or SNPs (Single Nucleotide Polymorphisms) [260], can be challenging due to limitations in data analysis algorithms, as in GWAS (Genome-wide association study). While GWAS employs simple regression analysis, ignoring correlations among SNPs [261], CNNs excel at analyzing image datasets (unstructured data) by learning features from spatial dependencies in the images.

The purpose of this chapter was to propose a method for using CNNs to learn features for disease stratification by converting tabular data into images. Although genetic data was not available, existing CV risk factors were used in GoDARTS. DL architectures were applied for risk stratification of 5-year MACE using retinal images alone and with T2I converted image from CV risk factor features. The results of DL models presented in Table 7.7 and 7.8 demonstrate that incorporating GWPRS feature data enhances the model performance for the stratification of 5-year MACE, confirming our previous findings reported in Chapter 6 that the information from GWPRS is complementary to that from the retina [262]. However, it

is noticed that the boost in performance varies across different models, possibly due to the random initialization of hyper-parameters in DL models, which requires further investigation.

Zhu et al. [249] have highlighted the significance of optimal feature ordering for the effective application of CNNs on image data. They have introduced the IGTD algorithm as a means of achieving optimal feature ordering, using genetic data as an example. In our study, we employed the IGTD algorithm on CV risk factors data but found that the use of ordered or unordered features did not make much difference in the results. This might be attributed to the relatively small number of features used in our experiment.

There are certain drawbacks associated with this approach. Firstly, generating image data from a smaller tabular dataset may produce redundant image features. Secondly, when new tabular data is added, all the images have to be regenerated because normalization is necessary. Thirdly, data imputation can be difficult, and special techniques such as reserving pixel values have to be utilized. Lastly, converting tabular data to grayscale values may result in some loss of data granularity.

7.7 Conclusions

In previous chapters, only retinal images for the prediction of MACE were used. Here, we experimented with multi-modal input. The CV risk factors, available as tabular (spreadsheet) data, were converted to images with our proposed T2I algorithm and used in conjunction with retinal images for predicting 5-years MACE from the date of imaging. The influence of feature ordering on performance was explored by implementing IGTD [249]. The machine learning models, logistic regression and XGBoost classifier were compared for predicting 5-years MACE using the risk factors from tabular data. Different DL models were trained with slight variations in their architectures, i.e., introducing dropout layers and FC layers. The FC layers are responsible for taking the output from convolutional and pooling layers and producing the final prediction. However, they can lead to overfitting due to a large number of trainable parameters. To prevent overfitting, the dropout regularization technique is often used in conjunction with FC layers. The dropout layer randomly drops out some of

the neurons in the FC layer during training to prevent them from becoming too dependent on the input data and memorizing it.

The maximum AUC obtained was 72.54 % using DL model 3 (with dropout layer) with multi-image input of retinal image and T2I with 10 features. XGBoost classifier gave a similar performance of AUC 72.29 with 10 feature tabular data. From the results, it is noted that adding GWPRS feature data to the model boosts the performance in predicting 5-years MACE by a 2% increase in AUC. There was no boosting in performance with the IGTD optimized feature ordering. The reason might be that the number of features is too small for IGTD. Results also suggest that multi-modal image data has more predictive power than using retinal image alone for 5-years MACE prediction. With only retinal images AUC was 66.78% using DL model 1 and with multi-modal images (retina + T2I (10 features)), AUCs were 70.55%, 72.54%, 71.95% and 71.43 % using DL Model 2, DL Model 3, DL Model 4 and DL Model 5 respectively. T2I with 10 features gave better performance than using retinal images alone in the DL models.

We also observe that the DL model performance for predicting 5 years MACE using only retinal images is higher with an AUC of 66.78% compared to predicting MACE within 12 years of retinal imaging from only retinal images using DL where, AUC was 64.2%, described in Section 5.5.3 (Prediction of MACE). This shows that the MACE prediction closer to the retinal imaging date gives better predictive performance than farther MACE from fundus images using DL.

The next and final chapter summarises the work presented in this thesis, discusses the main contributions, and outlines topics for future work.

Chapter 8

Conclusions and future work

8.1 Summary of work

This thesis has focused on analyzing retinal color fundus images to find associations with systemic conditions, especially Cardiovascular Disease (CVD) and its associated risk factors, using the Genetics of Diabetes Audit and Research in Tayside Scotland (GoDARTS) bio-resource and deep learning algorithms.

The first step was to investigate the best-performing deep learning model for classification or regression tasks with retinal images, using synthetic images and controlled levels of task difficulty. This was presented in Chapter 4.

The chapter reports our general framework for generating synthetic datasets parameterized by difficulty level to test a Deep Learning (DL) model. The steps involved in the proposed framework are (1) identify the target problem (e.g., binary classification), (2) identify the set of parameters characterizing difficulty and their ranges, (3) define a discrete grid by sampling the parameter space, (4) generate synthetic images, (5) run the classifier on the whole grid and analyze the results. These steps were described in detail in Section 4.2.

For the synthetic data generation, 547 healthy retinal images from Methods to Evaluate Segmentation and Indexing Techniques in the field of Retinal Ophthalmology (MESSIDOR) were augmented using manually cropped lesion patches from the 640 Diabetic Retinopathy (DR) 1 to 3 graded retinal images of MESSIDOR, described in Section 4.3.1. Three

dimensions, namely size of the manually cropped lesion patch, the number of patches, and the transparency level of the patch, were chosen to build a parametric space. Based on the parametric space, a synthetic dataset was generated. Difficulty levels are clearly organized in the space; the most difficult problem (the fewest and smallest patches with high transparency) is situated opposite the easiest one (the most numerous and largest patches with zero transparency).

Multiple Convolutional Neural Network (CNN) models were tried by varying different hyper-parameters starting from a simple 2-layered CNN to the initial classification task of healthy vs unhealthy retinal images that were synthetically generated. Several well-known CNN models, namely VGG16 [92], ResNet50 [93], InceptionV3 [218], DenseNet201 [94] and EfficientNet-B2 [98], were tried by modifying their architectures described in Section 4.3.5. A comparative performance evaluation showed that EfficientNet-B2 provided better performance, based on the 10-fold cross-validation analysis at each data point in the parametric space. We also tested performance on an independent dataset, Indian Diabetic Retinopathy Image Dataset (IDRID), built for the classification of DR, and still found that EfficientNet-B2 was performing better than the other CNN models. Hence, the modified EfficientNet-B2 was chosen for our subsequent work using real retinal fundus images from GoDARTS. The higher performance of EfficientNet-B2 compared to other DL architectures might be due to their new compound model scaling approach, where all the dimensions of depth, width, and resolution are uniformly scaled using compound coefficients as described in [98].

The clinical measurements (Systolic Blood Pressure (SBP), Diastolic Blood Pressure (DBP), High-Density Lipoprotein (HDL), Total Cholesterol (TC), Glycated Haemoglobin (GH), Body Mass Index (BMI) and Tryglicerides (Trig)), disease outcomes (Major Adverse Cardiovascular Event (MACE), All Cause Death (ACD), Chronic Kidney Disease (CKD), DR, Diabetic Peripheral Neuropathy (DPN)) and Genome-Wide Polygenic Risk Scores (GWPRS) for the individuals at baseline (earliest retinal image available) in GoDARTS were provided by a team of clinicians and researchers as a part of our collaboration (Section 1.1.1). This was a fixed dataset that had been provided for undertaking the analysis to investigate

retinal associations with systemic conditions in GoDARTS using DL. This data was mapped to retinal images of GoDARTS using proCommunity Health Index (CHI), an anonymized numerical identifier unique to each patient.

In Chapter 5, retinal images for GoDARTS were used for predicting demographic features, clinical measurements, and disease outcomes. 102,082 images were used from 8,570 individuals available in GoDARTS for predicting demographic features, age, and sex using the modified EfficientNet-B2 model. 71,434 retinal images were used for training, 9,954 images for validation, and 20,694 for testing. The mean age of the whole population was 66.11 ± 11.77 years. For age prediction from only retinal fundus image as input, the model achieved Mean Absolute Error (MAE) of 3.951 (95% CI 3.908, 3.995) and R^2 of 0.809 (0.804, 0.814) on the complete test data. For predicting gender from only the retinal fundus image, the model achieved Area Under Receiver Operating Characteristic (ROC) Curve (AUC) of 0.899 (0.895, 0.903), an accuracy of 0.811 (0.806, 0.817), the sensitivity of 0.886 (0.88, 0.891) and specificity of 0.717 (0.708, 0.727) on the complete test data. Similar model performance was observed when the performance metrics were computed using only left-eye retinal images and only right-eye retinal images for both age and sex predictions. A little higher MAE for age predictions and a little lower AUC for sex prediction was observed in our experiments compared to [23, 158, 155, 162], possible reason might be due to the saturation of ageing changes in the retinal fundus of the older age groups.

The Grad-CAM heatmap generated shows that Optic Disc (OD) and macula are the critical regions for age prediction in all age groups and the retinal vasculature is additionally important in younger and middle-age groups. For sex prediction, from Grad-CAM heatmaps it appears that OD and the region around the macula are the important features. Further, the heatmaps suggest that, for the majority of male and female predictions, the OD and the temporal vascular arcade region are activated respectively. To confirm this observation, a systematic analysis was performed to check the consistency of the heatmap activations in all the male and female predictions separately. With a common color bar representing the same pixel intensity levels across all the mean heatmaps, Figure 5.4 shows that the OD region is highly activated for all the correct male predictions but not for the correct female

predictions. More details were presented in Section 5.2.3 (Grad-CAM heatmap consistency). The exact visualizations of the important regions are not very clear from Gradient-based Class Activation Mapping (Grad-CAM) heatmaps. This might be because of substantial upscaling (a process involved in generating Grad-CAM heatmap) that can generate artifacts that may introduce serious errors like masking out of important regions in the image or creating false positives. Further studies are required in the direction of visualizing the CNN predictions.

Section 5.3 reports our investigation on the difference between chronological and retinal vascular age predicted by DL and associations with MACE and ACD of Type 2 Diabetes (T2D) population in GoDARTS. For this experiment, all T2D individuals above 30 years of age at the time of the first available retinal image and with no previous history of hospitalization for myocardial infarction (MI) or stroke were considered. A total of 81,260 retinal images were obtained from 6,464 individuals. A modified EfficientNet-B2 model was trained and validated for age predictions from the train and validation splits. The model achieved MAE of 4.088 (95% CI 4.037, 4.135) and R^2 of 0.77 (0.764, 0.776) on the complete test data with 16,125 retinal images. The individual-level performance (average of the individual's left- and right-eye age prediction) at baseline in test data (from 1,316 individuals) is MAE of 3.692 (3.543, 3.844) and R^2 of 0.823 (0.806, 0.839). The average of the left eye and right predictions improve the model's performance.

The Predicted Age Difference (PAD) (difference between the age predicted by the DL model from retinal image and the chronological age) was computed at an individual level for 1,316 individuals at baseline in test data. Using Coxph regression analysis adjusted for age and sex, a 1-year increase in retinal PAD score increases the risk of mortality by 5.9% (Hazard Ratio (HR) = 1.0597, 95% CI = 1.034 – 1.085, P = 1.62e-06) and risk of developing MACE increases by 5.8% (HR = 1.0587, 95% CI = 1.028 - 1.089, P = 1.06e-4). Coxph results after adjusting for Atherosclerotic Cardiovascular Disease (ASCVD) risk score alone show a risk of mortality increases by 3.3% (hazard ratio (HR) = 1.033, 95% CI = 1.01 – 1.058, P = 0.0043) and risk of developing MACE increases by 4% (HR = 1.04, 95% CI = 1.011 - 1.069, P = 0.006) with a 1-year increase in retinal PAD score. More details are available

in Section 5.3.4 (Early Mortality and MACE in older retina group). Also, a longitudinal analysis was performed by computing τ_{rate} (see Section 5.3.4 (rate of change of PAD)) over a period of 5 years. Coxph analysis with 326 mortality events from 1,172 individuals adjusted for predicted age at first available retinal image and sex indicates that individuals in the high τ_{rate} group have a higher risk of mortality than those in the low τ_{rate} group by 57% (HR = 1.575, 95% CI = 1.196 – 2.074, P = 0.0011). There were no significant associations observed between τ_{rate} and MACE. One of the possible reasons might be because of the less number of MACE in the dataset the predictive power of the model might be less. For more details refer to Section 5.3.4 (rate of change of PAD).

DL model was used to predict Cardiovascular (CV) risk factors namely SBP, DBP, HDL, TC, GH, BMI and Trig from retinal fundus images in a complete diabetic population of GoDARTS, discussed in Section 5.4. A total of 13,964 retinal images from 6,656 individuals were used in this experiment. Each DL model was trained independently to predict these CV risk factors. The model achieved MAE of 5.88 (95% CI 5.71, 6.07) and R^2 of 0.14 (0.1, 0.17) for the prediction of DBP on the complete test dataset with 2,786 images. R^2 values for the rest of the CV risk factors is approximately equal to zero specifying that there is no information available in the retina for the corresponding feature predictions using this dataset. More details were presented in Section 5.4.3.

Section 5.5 presented an investigation on predicting systemic disease outcomes (with binary labels) namely MACE, ACD, and microvasculature complications (CKD, DPN, DR) within 12 years from the date of retinal imaging using DL. For the classification of MACE and ACD events, 13,964 retinal images from 6,656 individuals were used and for microvascular complications, 17,139 retinal images from 8,222 individuals were used. Each DL model was trained independently for the stratification of disease outcomes. The model achieved an AUC of 0.642 (95% CI 0.619, 0.665), 0.711 (0.693, 0.728), 0.741 (0.722, 0.759), 0.633 (0.614, 0.653), 0.57 (0.55, 0.591) for the classification of MACE, ACD, CKD, DR, DPN respectively. The AUC value is poor for the prediction of DPN. The results show that the retinal image contains some signal for the prediction of CKD, ACD, MACE, DR but no

signal for DPN predictions using the cross-sectional image analysis in GoDARTS. Detailed results were given in Section 5.5.3.

Chapter 6 presented results of our investigation on predicting Pooled Cohort Equations (PCE) ASCVD clinical risk score and GWPRS for CVD from retinal images using DL, and how the predicted clinical risk score stratifies individuals with MACE and Cardiovascular death (CV death) within 10 years from the date of retinal imaging. 13,964 retinal images from 6,656 individuals were used for this experiment. The modified EfficientNet-B2 model achieved an R^2 of 0.554 (95% CI 0.528, 0.579) and MAE of 0.107 (0.104, 0.11) for estimating the PCE ASCVD risk score in the test dataset. For GWPRS, the model achieved an R^2 of -0.005 (-0.019 , 0.009) and a MAE was 0.484 (0.467, 0.5). These results suggest that the retina does not contain information related to GWPRS for CVD. Moreover, the findings indicate that the retina may provide valuable information regarding CVD risk, which complements a GWPRS.

The individual-level Predicted Risk Score (PRS) (the DL predicted PCE ASCVD risk score from a retinal image) was computed for 1,317 individuals at baseline in test data. Using Coxph regression analysis adjusted for age, sex and GWPRS, the risk of developing MACE increases by 2.9% (HR = 1.029, 95% CI = 1.015 - 1.042, P = 3.4e-5) and the risk of CV death increases by 1.9% (hazard ratio (HR) = 1.019, 95% CI = 1.004 – 1.034, P = 0.009) with 1 percent increase in retinal PRS. More details are available in Section 6.6.4. Ω_{rate} (see Section 6.6.2 (rate of change of PRS)) was computed as part of a longitudinal analysis over a period of 3 years. The Coxph regression analysis adjusted for age at imaging, sex, and GWPRS indicates that individuals in the top 20% Ω_{rate} group have a higher risk of developing MACE than those in the bottom 80% Ω_{rate} group by 50.4% (HR = 1.504, 95% CI = 1.122 – 2.016, P = 0.006) and 47.5% (HR = 1.475, 95% CI = 1.081 – 2.012, P = 0.014) for CV death. More details were provided in Section 6.6.4, Rate of change of PRS. The results show that the retina contains information useful for predicting CV risk score and it could be used for diagnosing disease conditions associated with CVD. Further investigation and replication in other cohorts are required.

Finally, we experimented with an approach for converting tabular data of clinical measurements and data from GoDARTS into images, allowing multi-modal data analysis with a single data type (images). The retinal images and tabular data covered the stratification of 5 years MACE. Experimental results were presented in Chapter 7. For the conversion of tabular data to the image, CV risk features available in GoDARTS namely age, sex, DBP, SBP, BMI, GH, HDL, TC; along with GWPRS were used. The process for the conversion of tabular data to the image was described in Section 7.4.1. We further investigated the effect of feature ordering in Tabular data to Image (T2I), as CNNs are mainly suitable for analyzing the data with spatial or temporal dependencies between the components. For this, the Image Generator from Tabular Data (IGTD) algorithm was implemented (Section 7.4.2). Machine learning models, logistic regression and XGBoost classifier were explored for predicting 5-years MACE using the risk factors from tabular data. Five different DL models (Section 7.4.3) were trained with slight variations in their architectures. The maximum AUC obtained was 72.54 % using a DL model with a dropout layer and input consisting of the retinal image and T2I with 10 features. T2I with 10 features gave better performance than using retinal images alone in the DL models. We observed that multi-modal image data has more predictive power than a retinal image alone for 5-year MACE prediction. The addition of GWPRS feature data to the DL model along with retinal image and T2I from CV risk factors, improves the performance in predicting 5-years MACE by 2% increase in AUC. Experimental results were presented in Section 7.5.

8.2 Contributions

1. We proposed a framework for generating synthetic datasets *parameterized by difficulty level* to test classifiers for medical image analysis. The main contribution is a protocol to build structured datasets for systematic testing *by quantifying the difficulty of the data for the target task*. This framework is used for hyper-parameter tuning of the DL model and identifies the robust model for our retinal image analysis. To our best knowledge, this was not been addressed in the literature of synthetic medical data.

2. We investigated DL model for predicting CV risk factors, namely age, sex, SBP, DBP, HDL, TC, GH, BMI and Trig using only retinal images in the elderly diabetic cohort of GoDARTS with mean age of 66.11 ± 11.77 years. We are the first to apply DL on retinal images of GoDARTS for predicting CV risk factors. The results show that the DL algorithm using retina images: can perform very well in predicting age and sex; achieved relatively less performance for DBP prediction; and could not predict well for SBP, HDL, TC, GH, BMI and Trig.
3. We investigated the difference between predicted age from retinal image using DL and chronological age as a biomarker to find associations with MACE and ACD and strong statistically significant associations was observed. There is very limited literature available who worked in a similar direction and the majority of them used retinal images from the UK biobank [151]. To our best knowledge, we are the first to try this analysis in a complete diabetic population. Also the DL predicted age from the retina is investigated in longitudinal image data and observed a statistical significant association with ACD but not with MACE.
4. We investigated the ability of a deep learning approach applied to retinal images to predict the clinical risk score for PCE ASCVD and a GWPRS for CVD. Our results demonstrate that the retina contains information that can indicate clinical risk score, but no indication that it contains information related to GWPRS. These findings suggest that the retina may provide valuable information complementary to GWPRS for CVD risk. To the best of our knowledge, this is the first investigation of the complementarity of retinal and genetic information for CVD risk using DL.
5. We investigated whether retinal images and a DL algorithm can predict MACE and CV death within 10 years from the date of retinal imaging and how predictions from the images associated with those from a clinical risk score used in clinical practice. The coxph regression results show that there is a strong, statistically significant association with the predicted PCE ASCVD clinical risk score and MACE, CV death in 10 years.

The longitudinal retinal image data analysis also shows that predicted clinical risk score can stratify the risk of MACE and CV death.

6. A method to convert T2I was proposed and multi-modal image analysis performed using DL. The inputs were retinal images and converted T2I from CV risk factors and GWPRS for predicting MACE within 5 years of retinal imaging. Also a IGTD algorithm was implemented to find the optimal feature ordering for improving the DL model performance. It was observed that multi-modal image data has more predictive power than using retinal image alone for predicting MACE in 5 years.

8.3 Limitations and future work

This section discusses the limitations of our work and proposes possible solutions. It also presents topics that can be extended in the future.

8.3.1 Synthetic data generation

In the synthetic data generation, the primary aim was to build a DL architecture that can differentiate normal vs abnormal fundus images. During the synthetic data generation, manually cropped lesion patches were randomly added to healthy retinal images and this does not follow any pathological explanation for the lesions appearing on the retina. This synthetic data does not represent the real fundus images with the lesion of DR. The task of adding a cropped lesion patch to the retinal images by following the pathology required a manual segmentation of retinal images to extract lesions. Segmentation and annotations of retinal images are in themselves a large domain of research [119, 263, 204]. Currently, Generative Adversarial Network (GAN)s are a promising approach for synthetic dataset generation [206, 264–266] but generating synthetic data in a controlled fashion by difficulty, as we have done, has not yet been attempted. In the future, GANs can be explored on generating synthetic datasets as per the framework proposed in Section 4.2.

8.3.2 DL architecture

Designing a best-performing DL architecture seems to be a never-ending task. It is extremely difficult to control and explore values for all meta-parameters involved, such as the number of layers, kernel size, residual connections, dense connections, dropout, pooling, normalization, etc., all of which might improve performance. Extensive training and validation were performed to fine-tune the CNN models and find the best performing DL architecture using the synthetic dataset generated.

As per the recommendations from EfficientNet-B2 authors [98], for optimal performance the original retinal images were downsized to 260×260 pixels. The other variants of EfficientNet family DL architectures take different sizes of input images and have shown improved performance on ImageNet Large Scale Visual Recognition Challenge (ILSVRC) dataset. Given the timeline and computation resources available to us, we stick to the EfficientNet-B2 DL model. The recent DL literature on medical image analysis for example [267] shows that higher input image dimensions might improve the model performance further. With higher computational resources, DL models with higher input image dimensions can be tried as the next step. Because using higher dimension images can lead to a large number of trainable parameters and computations, increases computational complexity, and require more powerful hardware for efficient training and inference.

It is practically impossible to try all the DL models and variations reported in the literature to find the absolute best-performing one for a specific problem. It is also challenging to test fully new variants of a DL model, as this might need heavy computational resources and large amounts of data. Therefore we selected a few contemporary DL architectures already proven to be robust.

8.3.3 Explainability of DL

Explainability of DL models is the most challenging task in the DL research community. Each neuron receives a set of inputs from its previous layer, performs multiplication with weights learned during training through back-propagation, computes the output, and passes

to the neuron in the next layer. It remains unclear in which way exactly neurons cooperate to arrive at the final output so that the processing leading to the network's answer remains largely **black box** [268]. Geoffrey Hinton, one of the "godfathers of deep learning", has been reported to say that he is now "deeply suspicious of back-propagation" and "my view is thrown it all away and start again," [269].

The Grad-CAM [99] algorithm that we adopted allows one to visualize the critical regions in the input image that are mainly responsible for the predictions made by the trained DL model. Grad-CAM heatmaps are usually generated at the last convolutional layer of the DL model, where the dimensions of the feature maps become very small after the cascade of convolutional and pooling operations (Section 1.4.1). To visualize the important regions, the feature maps are enlarged to match the input image size. In the EfficientNet-B2 experiments, the 9×9 feature maps with weights from the last convolutional layer are resized to match the input image dimensions (260×260), generating heatmaps. However, this significant enlargement can introduce errors such as obscuring crucial areas or creating false positives. The Grad-CAM heatmaps we generated did not precisely identify the critical regions and lacked detailed resolution. Although the heatmap showed the primary branching point in the lower half of the retinal vasculature, as displayed in Figure 6.2, the heatmap's finer details were missing.

There is therefore a need for a robust visualization mechanism in DL [270–272]. Recently, ProtoPNet [273] has been proposed to address this problem.

8.3.4 GoDARTS

The following are some thoughts for future research from retinal images using DL involving GoDARTS.

1. *Image quality check.* All retinal images from GoDARTS were utilized in the experiments without any quality check mechanism, but image pre-processing was applied to all. However, as these images were captured during real-world diabetic eye screening programs, there were several images with poor quality, such as blur and shakiness,

which might be due to factors like the individual's age and mental health. Incorporating quality check measures, as suggested by Coyner et al. [274], could potentially enhance the DL model performance in the findings reported in this thesis.

2. *Image centering*. The experiments carried out in this thesis utilized fundus images from GoDARTS that are centered on the macula. However, Mookiah et al. [108] reported in their study that they observed substantial variation and insufficient concordance in the retinal measurement obtained from fundus images captured at two different centers, OD centered and macula centered. It would be intriguing to explore the outcomes of this thesis further using OD centered images.

3. *Longitudinal data analysis*

- GoDARTS contains longitudinal data collected for over 20 years. However, most of the research work in this thesis was done on cross-sectional data. Along with predictions for age and ASCVD risk score, CV risk factors can also be generated for all the retinal images in the longitudinal data using DL and with different follow-up window associations with MACE and CV death. This would take forward the work discussed in Section 5.3.4 (rate of change of PAD), and 6.6.4 (rate of change of PRS).
- It is practically impossible to collect real clinical data at regular points in time. Consequently, GoDARTS data are frequently unevenly distributed. For instance, over a span of 5 years, some individuals might have an average of 10 retinal images collected through eye examination while some others might have only 2. Consequently, applying DL, specifically Recurrent Neural Network (RNN), for predicting disease progression can be difficult as models often require data equally spaced in time. Recently, Bridge et al. [275] proposed a DL method to predict the disease progression in longitudinal image data with uneven time intervals with no prior feature extraction needed. Similar methods could be applied to retinal images in GoDARTS for the prediction of systemic disease progression, e.g., CVD, Alzheimer's, diabetes, and CKD.

4. *Microvascular complications.* For the classification of DR, a binary classification problem of no DR vs any DR was solved (see Section 5.5). This work can be extended to multi-class classification with 5 classes of DR, grade 0 to grade 4. Similarly, detecting biomarkers for CKD was analyzed as a binary classification problem pivoting on stage 3, i.e., CKD/higher or not. This work could be extended to multi-class classification by creating multiple classes based on estimated Glomerular Filtration Rate (eGFR) values.

The literature on predicting CV risk factors [23, 158, 155, 162] from retinal images using DL reports modest performance improvements over the results obtained in this thesis with GoDARTS data. A possible reason is that in GoDARTS the mean age of individuals at the time of the first available retinal image is nearly 66 years and the whole cohort is diabetic, whereas the mean age of participants in the literature is frequently below 55 years [23, 158, 155, 162]. Consistently with this, Kim et al. [155] reported that the performance of their DL model for age prediction deteriorated significantly in participants with age over 60 years. Further studies should focus on understanding the mechanism behind age prediction from fundus images.

8.3.5 Generalizability of results

Most of the experimental work in this thesis was carried out using retinal images from GoDARTS. Their retinal images might differ based on camera, Field of View (FoV), lightning conditions, image centering (macula or OD), etc. There is a need to assess the generalizability of the investigation results obtained from GoDARTS with retinal images from other sources.

To this purpose, retinal images from MESSIDOR [12] were used for fine-tuning different CNN models. IDRID [13] is an independent dataset used for further validation of the selection of the best performing DL model for carrying our experiments with retinal images in GoDARTS (refer Chapters 5, 6 and 7).

References

- [1] Topcon fundus camera. <https://www.mulamoottileyehospital.com/topcon-trc50-ex-fundus-camera.php>, 28th Sep. 2022.
- [2] Wikipedia. https://en.wikipedia.org/wiki/Fundus_photography, 23rd Aug. 2022.
- [3] CS231n Convolutional Neural Networks for Visual Recognition. <https://cs231n.github.io/neural-networks-1/>, 21th Aug. 2022.
- [4] Typical CNN architecture. https://commons.wikimedia.org/wiki/File:Typical_cnn.png, 21th Aug. 2022.
- [5] A Practical Introduction to Deep Learning with Caffe and Python. <http://adilmoujahid.com/posts/2016/06/introduction-deep-learning-python-caffe/>, 21th Aug. 2022.
- [6] Yann LeCun, Koray Kavukcuoglu, and Clément Farabet. Convolutional networks and applications in vision. In *Proceedings of 2010 IEEE international symposium on circuits and systems*, pages 253–256. IEEE, 2010.
- [7] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2921–2929, 2016.
- [8] Tao Li, Wang Bo, Chunyu Hu, Hong Kang, Hanruo Liu, Kai Wang, and Huazhu Fu. Applications of deep learning in fundus images: A review. *Medical Image Analysis*, 69:101971, 2021.
- [9] TJ MacGillivray, Emanuele Trucco, JR Cameron, Baljean Dhillon, JG Houston, and EJR Van Beek. Retinal imaging as a source of biomarkers for diagnosis, characterization and prognosis of chronic illness or long-term conditions. *The British journal of radiology*, 87(1040):20130832, 2014.
- [10] Sarah McGrory, Lucia Ballerini, Fergus N Doubal, Julie Staals, Mike Allerhand, Maria del C Valdes-Hernandez, Xin Wang, Tom MacGillivray, Alex SF Doney, Baljean Dhillon, et al. Retinal microvasculature and cerebral small vessel disease in the lothian birth cohort 1936 and mild stroke study. *Scientific reports*, 9(1):1–11, 2019.
- [11] Christian Janiesch, Patrick Zschech, and Kai Heinrich. Machine learning and deep learning. *Electronic Markets*, 31(3):685–695, 2021.

- [12] Etienne Decencière, Xiwei Zhang, Guy Cazuguel, Bruno Lay, Béatrice Cochener, Caroline Trone, Philippe Gain, Richard Ordonez, Pascale Massin, Ali Erginay, Béatrice Charton, and Jean-Claude Klein. Feedback on a publicly distributed database: the messidor database. *Image Analysis & Stereology*, 33(3):231–234, August 2014.
- [13] Prasanna Porwal, Samiksha Pachade, Ravi Kamble, Manesh Kokare, Girish Deshmukh, Vivek Sahasrabuddhe, and Fabrice Meriaudeau. Indian diabetic retinopathy image dataset (idrid): a database for diabetic retinopathy screening research. *Data*, 3(3):25, 2018.
- [14] The top 10 causes of death. <https://www.who.int/news-room/fact-sheets/detail/the-top-10-causes-of-death>, 21st Sep. 2022.
- [15] Cardiovascular diseases (CVDs). [https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-\(cvds\)#:~:text=Key%20facts,to%20heart%20attack%20and%20stroke.](https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds)#:~:text=Key%20facts,to%20heart%20attack%20and%20stroke.), 21st Sep. 2022.
- [16] Elizabeth Wilkins, L Wilson, Kremlin Wickramasinghe, Prachi Bhatnagar, Jose Leal, Ramon Luengo-Fernandez, R Burns, Mike Rayner, and Nick Townsend. European cardiovascular disease statistics 2017. 2017.
- [17] Ensuring early and accurate diagnosis to improve access to healthcare. <https://www.roche.com/about/strategy/access-to-healthcare/diagnosis/#:~:text=Effective%20screening%20and%20accurate%2C%20early,health%20strategy%20in%20all%20settings>, 27th Jan. 2023.
- [18] Why early diagnosis of dementia is important. <https://www.scie.org.uk/dementia/symptoms/diagnosis/early-diagnosis.asp>, 27th Jan. 2023.
- [19] Michael D Abràmoff, Mona K Garvin, and Milan Sonka. Retinal imaging and image analysis. *IEEE reviews in biomedical engineering*, 3:169–208, 2010.
- [20] Anat London, Inbal Benhar, and Michal Schwartz. The retina as a window to the brain—from eye research to cns disorders. *Nature Reviews Neurology*, 9(1):44–53, 2013.
- [21] Shervin Minaee, Yuri Y Boykov, Fatih Porikli, Antonio J Plaza, Nasser Kehtarnavaz, and Demetri Terzopoulos. Image segmentation using deep learning: A survey. *IEEE transactions on pattern analysis and machine intelligence*, 2021.
- [22] Maryam Badar, Muhammad Haris, and Anam Fatima. Application of deep learning for retinal image analysis: A review. *Computer Science Review*, 35:100203, 2020.
- [23] Ryan Poplin, Avinash V Varadarajan, Katy Blumer, Yun Liu, Michael V McConnell, Greg S Corrado, Lily Peng, and Dale R Webster. Prediction of cardiovascular risk factors from retinal fundus photographs via deep learning. *Nature Biomedical Engineering*, 2(3):158–164, 2018.
- [24] INSPIRED-NIHR. <https://inspired-nihr.com/>, 28th Sep. 2022.
- [25] Health Informatics Center. <https://www.dundee.ac.uk/hic>, 12th August 2021.

-
- [26] William Alexander Newman Dorland. *Dorland's illustrated medical dictionary*. WB Saunders, 1925.
- [27] Yeong Yeh Lee, Mohammad Majharul Haque, Rona Marie Lawenko, and Amol Sharma. Systemic disorders that affect gastrointestinal motility. In *Clinical and Basic Neurogastroenterology and Motility*, pages 601–618. Elsevier, 2020.
- [28] Cardiovascular disease. <https://www.nhs.uk/conditions/cardiovascular-disease/>, 10th Sep. 2022.
- [29] Ralph B D'Agostino Sr, Ramachandran S Vasan, Michael J Pencina, Philip A Wolf, Mark Cobain, Joseph M Massaro, and William B Kannel. General cardiovascular risk profile for use in primary care: the framingham heart study. *Circulation*, 117(6):743–753, 2008.
- [30] David C Goff, Donald M Lloyd-Jones, Glen Bennett, Sean Coady, Ralph B D'agostino, Raymond Gibbons, Philip Greenland, Daniel T Lackland, Daniel Levy, Christopher J O'donnell, et al. 2013 acc/aha guideline on the assessment of cardiovascular risk: a report of the american college of cardiology/american heart association task force on practice guidelines. *Journal of the American College of Cardiology*, 63(25 Part B):2935–2959, 2014.
- [31] What is Diabetes? <https://www.niddk.nih.gov/health-information/diabetes/overview/what-is-diabetes>, 10th Sep. 2022.
- [32] Diabetes. <https://www.nhs.uk/conditions/diabetes/>, 10th Sep. 2022.
- [33] About diabetes. https://web.archive.org/web/20140331094533/http://www.who.int/diabetes/action_online/basics/en/, 10th Sep. 2022.
- [34] Diabetes. https://en.wikipedia.org/wiki/Diabetes#cite_note-WHO2013-2, 10th Sep. 2022.
- [35] Elham Saedi, Mohammad Reza Gheini, Firoozeh Faiz, and Mohammad Ali Arami. Diabetes mellitus and cognitive impairments. *World journal of diabetes*, 7(17):412, 2016.
- [36] Chronic kidney disease. <https://www.nhs.uk/conditions/kidney-disease/>, 10th Sep. 2022.
- [37] Chronic kidney disease. <https://www.hopkinsmedicine.org/health/conditions-and-diseases/chronic-kidney-disease>, 10th Sep. 2022.
- [38] Your Eyes Could Be the Windows to Your Health. <https://www.aaopt.org/eye-health/tips-prevention/diagnosing-systemic-diseases-eye-exams>, 18th Sep. 2022.
- [39] Ophthalmology. <https://chicago.medicine.uic.edu/departments/academic-departments/ophthalmology-visual-sciences/>, 18th Sep. 2022.
- [40] John E Dowling. *Neurons and networks: an introduction to behavioral neuroscience*. Harvard University Press, 2001.

-
- [41] Vilayanur S Ramachandran. *Encyclopedia of the Human Brain: Col-Mem*, volume 2. Academic Press, 2002.
- [42] Nikolaus F Troje and A Basbaum. Biological motion perception. *The senses: A comprehensive reference*, 2:231–238, 2008.
- [43] Larry Benowitz and Yuqin Yin. Rewiring the injured cns: lessons from the optic nerve. *Experimental neurology*, 209(2):389–398, 2008.
- [44] Paul Lingor, Nicole Teusch, Katrin Schwarz, Reinhold Mueller, Helmut Mack, Mathias Bähr, and Bernhard K Mueller. Inhibition of rho kinase (rock) increases neurite outgrowth on chondroitin sulphate proteoglycan in vitro and axonal regeneration in the adult optic nerve in vivo. *Journal of neurochemistry*, 103(1):181–189, 2007.
- [45] J Wayne Streilein. Ocular immune privilege: therapeutic opportunities from an experiment of nature. *Nature Reviews Immunology*, 3(11):879–889, 2003.
- [46] C Kaur, WS Foulds, and EA Ling. Blood–retinal barrier in hypoxic ischaemic conditions: basic concepts, clinical features and management. *Progress in retinal and eye research*, 27(6):622–647, 2008.
- [47] Kenneth W Wright. Alphabet patterns and oblique muscle dysfunctions. In *Pediatric ophthalmology and strabismus*, pages 232–249. Springer, 2003.
- [48] Charumathi Sabanayagam, Anoop Shankar, David Koh, Kee Seng Chia, Seang Mei Saw, Su Chi Lim, E Shyong Tai, and Tien Yin Wong. Retinal microvascular caliber and chronic kidney disease in an asian population. *American journal of epidemiology*, 169(5):625–632, 2009.
- [49] Laurence Shen Lim, Carol Yim-lui Cheung, Charumathi Sabanayagam, Su Chi Lim, E Shyong Tai, Lei Huang, and Tien Yin Wong. Structural changes in the retinal microvasculature and renal function. *Investigative ophthalmology & visual science*, 54(4):2970–2976, 2013.
- [50] Barbara Cassin, Sheila Solomon, and Melvin L Rubin. *Dictionary of eye terminology*. Triad Publishing Company Gainesville, 1990.
- [51] Patrick J Saine and Marshall E Tyler. *Ophthalmic photography: retinal photography, angiography, and electronic imaging*, volume 132. Butterworth-Heinemann Boston, 2002.
- [52] MD Abramoff, RH Kardon, KA Vermeer, and MH Mensink. A portable, patient friendly scanning laser ophthalmoscope for diabetic retinopathy imaging: Exudates and hemorrhages. *Investigative Ophthalmology & Visual Science*, 48(13):2592–2592, 2007.
- [53] Daniel X Hammer, R Daniel Ferguson, Teoman E Ustun, Chad E Bigelow, Nicusor V Iftimia, and Robert H Webb. Line-scanning laser ophthalmoscope. *Journal of biomedical optics*, 11(4):041126, 2006.
- [54] Jan Van de Kraats and Dirk van Norren. Directional and nondirectional spectral reflection from the human fovea. *Journal of biomedical optics*, 13(2):024010, 2008.

-
- [55] François C Delori and Kent P Pflibsen. Spectral reflectance of the human ocular fundus. *Applied optics*, 28(6):1061–1077, 1989.
- [56] Michael D Abramoff, Young H Kwon, Dan Ts'o, Peter Soliz, Bridget Zimmerman, Joel Pokorny, and Randy Kardon. Visual stimulus–induced changes in human near-infrared fundus reflectance. *Investigative ophthalmology & visual science*, 47(2):715–721, 2006.
- [57] Mircea Mujat, R Daniel Ferguson, Nicusor Iftimia, and Daniel X Hammer. Compact adaptive optics line scanning ophthalmoscope. *Optics express*, 17(12):10242–10258, 2009.
- [58] Wolfgang Drexler, Mengyang Liu, Abhishek Kumar, Tschackad Kamali, Angelika Unterhuber, and Rainer A Leitgeb. Optical coherence tomography today: speed, contrast, and multimodality. *Journal of biomedical optics*, 19(7):071412, 2014.
- [59] Richard F Spaide, James M Klancnik, and Michael J Cooney. Retinal vascular layers imaged by fluorescein angiography and optical coherence tomography angiography. *JAMA ophthalmology*, 133(1):45–50, 2015.
- [60] Noemi Lois and John V Forrester. *Fundus autofluorescence*. Lippincott Williams & Wilkins, 2009.
- [61] Matthew T Witmer and Szilárd Kiss. Wide-field imaging of the retina. *Survey of ophthalmology*, 58(2):143–154, 2013.
- [62] Aaron Nagiel, Robert A Lalane, Srinivas R Sadda, and Steven D Schwartz. Ultra-widefield fundus imaging: a review of clinical applications and future trends. *Retina*, 36(4):660–678, 2016.
- [63] Xavier Hadoux, Flora Hui, Jeremiah KH Lim, Colin L Masters, Alice Pébay, Sophie Chevalier, Jason Ha, Samantha Loi, Christopher J Fowler, Christopher Rowe, et al. Non-invasive in vivo hyperspectral imaging of the retina for potential biomarker use in alzheimer's disease. *Nature communications*, 10(1):1–12, 2019.
- [64] Hyperspectral Retinal Imaging. <https://www.photonetc.com/applications/hyperspectral-retinal-imaging>, 29th Sep. 2022.
- [65] Anil K Jain, Jianchang Mao, and K Moidin Mohiuddin. Artificial neural networks: A tutorial. *Computer*, 29(3):31–44, 1996.
- [66] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *nature*, 521(7553):436–444, 2015.
- [67] Yoshua Bengio et al. Learning deep architectures for ai. *Foundations and trends® in Machine Learning*, 2(1):1–127, 2009.
- [68] Jürgen Schmidhuber. Deep learning in neural networks: An overview. *Neural networks*, 61:85–117, 2015.
- [69] Frank Rosenblatt. Principles of neurodynamics. perceptrons and the theory of brain mechanisms. Technical report, Cornell Aeronautical Lab Inc Buffalo NY, 1961.

-
- [70] Y Le Cun. A learning procedure for asymmetric network. *Proceedings of Cognitive*, 85:599–604, 1985.
- [71] David E Rumelhart, James L McClelland, PDP Research Group, et al. *Parallel distributed processing: Explorations in the microstructure of cognition two voll*, 1986.
- [72] Yann LeCun, D Touresky, G Hinton, and T Sejnowski. A theoretical framework for back-propagation. In *Proceedings of the 1988 connectionist models summer school*, volume 1, pages 21–28, 1988.
- [73] Marvin Minsky and Seymour A Papert. *Perceptrons, Reissue of the 1988 Expanded Edition with a new foreword by Léon Bottou: An Introduction to Computational Geometry*. MIT press, 2017.
- [74] David E Rumelhart, Geoffrey E Hinton, and Ronald J Williams. Learning representations by back-propagating errors. *nature*, 323(6088):533–536, 1986.
- [75] Ruslan Salakhutdinov and Geoffrey Hinton. Semantic hashing. *International Journal of Approximate Reasoning*, 50(7):969–978, 2009.
- [76] Nathan Hubens. Deep inside: Autoencoders-towards data science. *Google Scholar*, 2018.
- [77] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014.
- [78] Quoc V Le et al. A tutorial on deep learning part 2: Autoencoders, convolutional neural networks and recurrent neural networks. *Google Brain*, 20:1–20, 2015.
- [79] Filippo Maria Bianchi, Enrico Maiorino, Michael C Kampffmeyer, Antonello Rizzi, and Robert Jenssen. An overview and comparative analysis of recurrent neural networks for short term load forecasting. *arXiv preprint arXiv:1705.04378*, 2017.
- [80] Yann LeCun, Bernhard Boser, John Denker, Donnie Henderson, Richard Howard, Wayne Hubbard, and Lawrence Jackel. Handwritten digit recognition with a back-propagation network. *Advances in neural information processing systems*, 2, 1989.
- [81] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.
- [82] David H Hubel and Torsten N Wiesel. Receptive fields, binocular interaction and functional architecture in the cat’s visual cortex. *The Journal of physiology*, 160(1):106, 1962.
- [83] Daniel J Felleman and David C Van Essen. Distributed hierarchical processing in the primate cerebral cortex. *Cerebral cortex (New York, NY: 1991)*, 1(1):1–47, 1991.
- [84] Neena Aloysius and M Geetha. A review on deep convolutional neural networks. In *2017 international conference on communication and signal processing (ICCSP)*, pages 0588–0592. IEEE, 2017.

-
- [85] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115(3):211–252, 2015.
- [86] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25:1097–1105, 2012.
- [87] Matthew D Zeiler, Graham W Taylor, and Rob Fergus. Adaptive deconvolutional networks for mid and high level feature learning. In *2011 international conference on computer vision*, pages 2018–2025. IEEE, 2011.
- [88] Matthew D Zeiler and Rob Fergus. Visualizing and understanding convolutional networks. In *European conference on computer vision*, pages 818–833. Springer, 2014.
- [89] Min Lin, Qiang Chen, and Shuicheng Yan. Network in network. *arXiv preprint arXiv:1312.4400*, 2013.
- [90] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1–9, 2015.
- [91] Dan Claudiu Ciresan, Ueli Meier, Jonathan Masci, Luca Maria Gambardella, and Jürgen Schmidhuber. Flexible, high performance convolutional neural networks for image classification. In *Twenty-second international joint conference on artificial intelligence*, 2011.
- [92] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.
- [93] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.
- [94] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4700–4708, 2017.
- [95] Saining Xie, Ross Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He. Aggregated residual transformations for deep neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1492–1500, 2017.
- [96] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7132–7141, 2018.

-
- [97] Chenxi Liu, Barret Zoph, Maxim Neumann, Jonathon Shlens, Wei Hua, Li-Jia Li, Li Fei-Fei, Alan Yuille, Jonathan Huang, and Kevin Murphy. Progressive neural architecture search. In *Proceedings of the European conference on computer vision (ECCV)*, pages 19–34, 2018.
- [98] Mingxing Tan and Quoc Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International Conference on Machine Learning*, pages 6105–6114. PMLR, 2019.
- [99] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pages 618–626, 2017.
- [100] Nicholas Witt, Tien Y Wong, Alun D Hughes, Nish Chaturvedi, Barbara E Klein, Richard Evans, Mary McNamara, Simon A McG Thom, and Ronald Klein. Abnormalities of retinal microvascular structure and risk of mortality from ischemic heart disease and stroke. *Hypertension*, 47(5):975–981, 2006.
- [101] Marta Favali, Samaneh Abbasi-Sureshjani, Bart ter Haar Romeny, and Alessandro Sarti. Analysis of vessel connectivities in retinal images by cortically inspired spectral clustering. *Journal of Mathematical Imaging and Vision*, 56(1):158–172, 2016.
- [102] Jeffrey Wigdahl, Pedro Guimaraes, and Alfredo Ruggeri. A shortest path approach to optic disc detection in retinal fundus images. *Modeling and Artificial Intelligence in Ophthalmology*, 1(4):29–42, 2017.
- [103] Lorenza Bonaldi, Elisa Menti, Lucia Ballerini, Alfredo Ruggeri, and Emanuele Trucco. Automatic generation of synthetic retinal fundus images: vascular network. *Procedia Computer Science*, 90:54–60, 2016.
- [104] Sara B Seidemann, Brian Claggett, Paco E Bravo, Ankur Gupta, Hoshang Farhad, Barbara E Klein, Ronald Klein, Marcelo Di Carli, and Scott D Solomon. Retinal vessel calibers in predicting long-term cardiovascular outcomes: the atherosclerosis risk in communities study. *Circulation*, 134(18):1328–1338, 2016.
- [105] Clarissa Shu Ming Cheng, Yi Fang Lee, Charles Ong, Zhu Li Yap, Andrew Tsai, Aditi Mohla, Monisha E Nongpiur, Tin Aung, and Shamira A Perera. Inter-eye comparison of retinal oximetry and vessel caliber between eyes with asymmetrical glaucoma severity in different glaucoma subtypes. *Clinical Ophthalmology (Auckland, NZ)*, 10:1315, 2016.
- [106] ML Rasmussen, R Broe, U Frydkjaer-Olsen, BS Olsen, HB Mortensen, T Peto, and J Grauslund. Retinal vascular geometry and its association to microvascular complications in patients with type 1 diabetes: the danish cohort of pediatric diabetes 1987 (dcpd1987). *Graefe's Archive for Clinical and Experimental Ophthalmology*, 255(2):293–299, 2017.
- [107] Kuryati Kipli, Mohammed Enamul Hoque, Lik Thai Lim, Muhammad Hamdi Mahmood, Siti Kudnie Sahari, Rohana Sapawi, Nordiana Rajae, and Annie Joseph. A review on the extraction of quantitative retinal microvascular image feature. *Computational and mathematical methods in medicine*, 2018, 2018.

-
- [108] Muthu Rama Krishnan Mookiah, Sarah McGrory, Stephen Hogg, Jackie Price, Rachel Forster, Thomas J MacGillivray, and Emanuele Trucco. Towards standardization of retinal vascular measurements: on the effect of image centering. In *Computational pathology and ophthalmic medical image analysis*, pages 294–302. Springer, 2018.
- [109] Carol Yim-lui Cheung, ShinYeu Ong, M Kamran Ikram, Yi Ting Ong, Christopher P Chen, Narayanaswamy Venketasubramanian, and Tien Yin Wong. Retinal vascular fractal dimension is associated with cognitive dysfunction. *Journal of Stroke and Cerebrovascular Diseases*, 23(1):43–50, 2014.
- [110] Sophie Lemmens, Astrid Devulder, Karel Van Keer, Johan Bierkens, Patrick De Boever, and Ingeborg Stalmans. Systematic review on fractal dimension of the retinal vasculature in neurodegeneration and stroke: assessment of a potential biomarker. *Frontiers in neuroscience*, 14:16, 2020.
- [111] Ce Shi, Yihong Chen, William Robert Kwapong, Qiaowen Tong, Senxiang Wu, Yuheng Zhou, Hanpei Miao, Meixiao Shen, and Hua Ye. Characterization by fractal dimension analysis of the retinal capillary network in parkinson disease. *Retina*, 40(8):1483–1491, 2020.
- [112] Sam Yu and Vasudevan Lakshminarayanan. Fractal dimension and retinal pathology: a meta-analysis. *Applied Sciences*, 11(5):2376, 2021.
- [113] Tien Yin Wong, Michael D Knudtson, Ronald Klein, Barbara EK Klein, Stacy M Meuer, and Larry D Hubbard. Computer-assisted measurement of retinal vessel diameters in the beaver dam eye study: methodology, correlation between eyes, and effect of refractive errors. *Ophthalmology*, 111(6):1183–1190, 2004.
- [114] CAROL YIM-LUI CHEUNG, Wynne Hsu, Mong Li Lee, Jie Jin Wang, Paul Mitchell, Qiangfeng Peter Lau, Haslina Hamzah, Maisie Ho, and Tien Yin Wong. A new method to measure peripheral retinal vascular caliber over an extended area. *Microcirculation*, 17(7):495–503, 2010.
- [115] Adria Perez-Rovira, T MacGillivray, Emanuele Trucco, KS Chin, K Zutis, C Lupascu, Domenico Tegolo, Andrea Giachetti, Peter J Wilson, A Doney, et al. Vampire: vessel assessment and measurement platform for images of the retina. In *2011 Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, pages 3391–3394. IEEE, 2011.
- [116] Muhammad Moazam Fraz, RA Welikala, Alicja R Rudnicka, Christopher G Owen, DP Strachan, and Sarah A Barman. Quartz: Quantitative analysis of retinal vessel topology and size—an automated system for quantification of retinal vessels morphology. *Expert Systems with Applications*, 42(20):7221–7234, 2015.
- [117] FN Doubal, TJ MacGillivray, PE Hokke, B Dhillon, MS Dennis, and JM Wardlaw. Differences in retinal vessels support a distinct vasculopathy causing lacunar stroke. *Neurology*, 72(20):1773–1778, 2009.
- [118] Harry Leung, Jie Jin Wang, Elena Rochtchina, Ava G Tan, Tien Y Wong, Ronald Klein, Larry D Hubbard, and Paul Mitchell. Relationships between age, blood pressure, and retinal vessel diameters in an older population. *Investigative ophthalmology & visual science*, 44(7):2900–2904, 2003.

-
- [119] Muthu Rama Krishnan Mookiah, Stephen Hogg, Tom J MacGillivray, Vijayaraghavan Prathiba, Rajendra Pradeepa, Viswanathan Mohan, Ranjit Mohan Anjana, Alexander S Doney, Colin NA Palmer, and Emanuele Trucco. A review of machine learning methods for retinal blood vessel segmentation and artery/vein classification. *Medical Image Analysis*, 68:101905, 2021.
- [120] Sebastian Dinesen, Pia S Jensen, Maria Bloksgaard, Søren Leer Blindbæk, Jo De Mey, Lars M Rasmussen, Jes S Lindholt, and Jakob Grauslund. Retinal vascular fractal dimensions and their association with macrovascular cardiac disease. *Ophthalmic Research*, 64(4):561–566, 2021.
- [121] Thomas Lee Torp, Ryo Kawasaki, Tien Yin Wong, Tunde Peto, and Jakob Grauslund. Retinal arteriolar calibre and venular fractal dimension predict progression of proliferative diabetic retinopathy 6 months after panretinal photocoagulation: a prospective, clinical interventional study. *BMJ open ophthalmology*, 6(1):e000661, 2021.
- [122] Emmanuel Sandoval-Garcia, Stela McLachlan, Anna H Price, Thomas J MacGillivray, Mark WJ Strachan, James F Wilson, and Jackie F Price. Retinal arteriolar tortuosity and fractal dimension are associated with long-term cardiovascular outcomes in people with type 2 diabetes. *Diabetologia*, 64(10):2215–2227, 2021.
- [123] Yevgeniya Atiskova, Jan Wildner, Martin Stephan Spitzer, Charlotte Aries, Nicole Muschol, and Simon Dulz. Retinal vessel tortuosity as a prognostic marker for disease severity in fabry disease. *Orphanet Journal of Rare Diseases*, 16(1):1–9, 2021.
- [124] Caixia Sun, Tingli Chen, Jing Cong, Xinyuan Wu, Jing Wang, and Yuanzhi Yuan. Changes in retinal vascular bifurcation in eyes with myopia. 2022.
- [125] Cong Sun, Gerald Liew, Jie Jin Wang, Paul Mitchell, Seang Mei Saw, Tin Aung, E Shyong Tai, and Tien Y Wong. Retinal vascular caliber, blood pressure, and cardiovascular risk factors in an asian population: the singapore malay eye study. *Investigative ophthalmology & visual science*, 49(5):1784–1790, 2008.
- [126] Hiroshi Yatsuya, Aaron R Folsom, Tien Y Wong, Ronald Klein, Barbara EK Klein, and A Richey Sharrett. Retinal microvascular abnormalities and risk of lacunar stroke: Atherosclerosis risk in communities study. *Stroke*, 41(7):1349–1355, 2010.
- [127] Matthias P Naegele, Jens Barthelmes, Valeria Ludovici, Silviya Cantatore, Arnold von Eckardstein, Frank Enseleit, Thomas F Lüscher, Frank Ruschitzka, Isabella Sudano, and Andreas J Flammer. Retinal microvascular dysfunction in heart failure. *European heart journal*, 39(1):47–56, 2018.
- [128] Wanfen Yip, Peng Guan Ong, Boon Wee Teo, Carol Yim-lui Cheung, E Shyong Tai, Ching-Yu Cheng, Ecosse Lamoureux, Tien Yin Wong, and Charumathi Sabanayagam. Retinal vascular imaging markers and incident chronic kidney disease: a prospective cohort study. *Scientific reports*, 7(1):1–9, 2017.
- [129] Shaohua Guo, Songtao Yin, Gary Tse, Guangping Li, Long Su, and Tong Liu. Association between caliber of retinal vessels and cardiovascular disease: a systematic review and meta-analysis. *Current Atherosclerosis Reports*, 22(4):1–13, 2020.

-
- [130] Karen KW Chan, Fangyao Tang, Clement CY Tham, Alvin L Young, and Carol Y Cheung. Retinal vasculature in glaucoma: a review. *BMJ open ophthalmology*, 1(1):e000032, 2017.
- [131] Zhuo Zhang, Ruchir Srivastava, Huiying Liu, Xiangyu Chen, Lixin Duan, Damon Wing Kee Wong, Chee Keong Kwoh, Tien Yin Wong, and Jiang Liu. A survey on computer aided diagnosis for ocular diseases. *BMC medical informatics and decision making*, 14(1):1–29, 2014.
- [132] Ronald Klein, Barbara EK Klein, Scot E Moss, Tien Y Wong, and A Richey Sharrett. Retinal vascular caliber in persons with type 2 diabetes: the wisconsin epidemiological study of diabetic retinopathy: Xx. *Ophthalmology*, 113(9):1488–1498, 2006.
- [133] Annette Kifley, Jie Jin Wang, Sudha Cugati, Tien Y Wong, and Paul Mitchell. Retinal vascular caliber, diabetes, and retinopathy. *American journal of ophthalmology*, 143(6):1024–1026, 2007.
- [134] Shawn Frost, Yogi Kanagasingam, Hamid Sohrabi, Janardhan Vignarajan, P Bourgeat, Oliver Salvado, Victor Villemagne, Christopher C Rowe, S Lance Macaulay, Cassandra Szoeki, et al. Retinal vascular biomarkers for early detection and monitoring of alzheimer’s disease. *Translational psychiatry*, 3(2):e233–e233, 2013.
- [135] Sourya Sengupta, Amitojdeep Singh, Henry A Leopold, Tanmay Gulati, and Vasudevan Lakshminarayanan. Application of deep learning in fundus image processing for ophthalmic diagnosis—a review. *arXiv preprint arXiv:1812.07101*, 2018.
- [136] Traci E Clemons, Emily Y Chew, Susan B Bressler, Wendy McBee, AREDS Research Group, et al. National eye institute visual function questionnaire in the age-related eye disease study (areds): Areds report no. 10. *Archives of Ophthalmology*, 121(2):211–217, 2003.
- [137] Philippe Burlina, David E Freund, Neil Joshi, Y Wolfson, and Neil M Bressler. Detection of age-related macular degeneration via deep learning. In *2016 IEEE 13th International Symposium on Biomedical Imaging (ISBI)*, pages 184–188. IEEE, 2016.
- [138] Srihari Kankanahalli, Philippe M Burlina, Yulia Wolfson, David E Freund, and Neil M Bressler. Automated classification of severity of age-related macular degeneration from fundus photographs. *Investigative ophthalmology & visual science*, 54(3):1789–1796, 2013.
- [139] Francisco Fumero, Silvia Alayón, José L Sanchez, Jose Sigut, and M Gonzalez-Hernandez. Rim-one: An open retinal image database for optic nerve evaluation. In *2011 24th international symposium on computer-based medical systems (CBMS)*, pages 1–6. IEEE, 2011.
- [140] Oscar Perdomo, Vincent Andrearczyk, Fabrice Meriaudeau, Henning Müller, and Fabio A González. Glaucoma diagnosis from eye fundus images based on deep morphometric feature estimation. In *Computational pathology and ophthalmic medical image analysis*, pages 319–327. Springer, 2018.

-
- [141] Gaurav O Gajbhiye and Ashok N Kamthane. Automatic classification of glaucomatous images using wavelet and moment feature. In *2015 annual IEEE India conference (INDICON)*, pages 1–5. IEEE, 2015.
- [142] Etienne Decencière, Xiwei Zhang, Guy Cazuguel, Bruno Lay, Béatrice Cochener, Caroline Trone, Philippe Gain, Richard Ordonez, Pascale Massin, Ali Erginay, et al. Feedback on a publicly distributed image database: the messidor database. *Image Analysis & Stereology*, 33(3):231–234, 2014.
- [143] Rishab Gargeya and Theodore Leng. Automated identification of diabetic retinopathy using deep learning. *Ophthalmology*, 124(7):962–969, 2017.
- [144] Sohini Roychowdhury, Dara D Koozekanani, and Keshab K Parhi. Dream: diabetic retinopathy analysis using machine learning. *IEEE journal of biomedical and health informatics*, 18(5):1717–1728, 2013.
- [145] Joes Staal, Michael D Abràmoff, Meindert Niemeijer, Max A Viergever, and Bram Van Ginneken. Ridge-based vessel segmentation in color images of the retina. *IEEE transactions on medical imaging*, 23(4):501–509, 2004.
- [146] João VB Soares, Jorge JG Leandro, Roberto M Cesar, Herbert F Jelinek, and Michael J Cree. Retinal vessel segmentation using the 2-d gabor wavelet and supervised classification. *IEEE Transactions on medical Imaging*, 25(9):1214–1222, 2006.
- [147] Bálint Antal and András Hajdu. An ensemble-based system for automatic screening of diabetic retinopathy. *Knowledge-based systems*, 60:20–27, 2014.
- [148] Mrinal Haloi. Improved microaneurysm detection using deep neural networks. *arXiv preprint arXiv:1505.04424*, 2015.
- [149] Kevis-Kokitsi Maninis, Jordi Pont-Tuset, Pablo Arbeláez, and Luc Van Gool. Deep retinal image understanding. In *International conference on medical image computing and computer-assisted intervention*, pages 140–148. Springer, 2016.
- [150] Henry A Leopold, Jeff Orchard, John Zelek, and Vasudevan Lakshminarayanan. Segmentation and feature extraction of retinal vascular morphology. In *Medical Imaging 2017: Image Processing*, volume 10133, pages 251–257. SPIE, 2017.
- [151] UK Biobank. <https://www.ukbiobank.ac.uk/learn-more-about-uk-biobank>, 6th Sep. 2022.
- [152] EyePACS. <https://www.eyepacs.org/home>, 6th Sep. 2022.
- [153] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*, 2014.
- [154] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *International conference on machine learning*, pages 2048–2057. PMLR, 2015.

- [155] Yong Dae Kim, Kyoung Jin Noh, Seong Jun Byun, Soochahn Lee, Tackeun Kim, Leonard Sunwoo, Kyong Joon Lee, Si-Hyuck Kang, Kyu Hyung Park, and Sang Jun Park. Effects of hypertension, diabetes, and smoking on age and sex prediction from retinal fundus images. *Scientific reports*, 10(1):1–14, 2020.
- [156] Sang Jun Park, Joo Young Shin, Sangkeun Kim, Jaemin Son, Kyu-Hwan Jung, and Kyu Hyung Park. A novel fundus image reading tool for efficient generation of a multi-dimensional categorical image database for machine learning algorithm training. *Journal of Korean medical science*, 33(43), 2018.
- [157] Jaemin Son, Joo Young Shin, Hoon Dong Kim, Kyu-Hwan Jung, Kyu Hyung Park, and Sang Jun Park. Development and validation of deep learning models for screening multiple abnormal findings in retinal fundus images. *Ophthalmology*, 127(1):85–94, 2020.
- [158] Tyler Hyungtaek Rim, Geunyoung Lee, Youngnam Kim, Yih-Chung Tham, Chan Joo Lee, Su Jung Baik, Young Ah Kim, Marco Yu, Mihir Deshmukh, Byoung Kwon Lee, et al. Prediction of systemic biomarkers from retinal photographs: development and validation of deep-learning algorithms. *The Lancet Digital Health*, 2(10):e526–e536, 2020.
- [159] Jost B Jonas, Liang Xu, and Ya Xing Wang. The beijing eye study. *Acta ophthalmologica*, 87(3):247–261, 2009.
- [160] Athena WP Foong, Seang-Mei Saw, Jing-Liang Loo, Sunny Shen, Seng-Chee Loon, Mohamad Rosman, Tin Aung, Donald TH Tan, E Shyong Tai, and Tien Y Wong. Rationale and methodology for a population-based study of eye diseases in malay people: The singapore malay eye study (simes). *Ophthalmic epidemiology*, 14(1):25–35, 2007.
- [161] Raghavan Lavanya, V Swetha E Jeganathan, Yingfeng Zheng, Prema Raju, Ning Cheung, E Shyong Tai, Jie Jin Wang, Ecosse Lamoureux, Paul Mitchell, Terri L Young, et al. Methodology of the singapore indian chinese cohort (sicc) eye study: quantifying ethnic variations in the epidemiology of eye diseases in asians. *Ophthalmic epidemiology*, 16(6):325–336, 2009.
- [162] Nele Gerrits, Bart Elen, Toon Van Craenendonck, Danai Triantafyllidou, Ioannis N Petropoulos, Rayaz A Malik, and Patrick De Boever. Age and sex affect deep learning prediction of cardiometabolic risk factors from retinal images. *Scientific reports*, 10(1):1–9, 2020.
- [163] Hanan Al Kuwari, Asma Al Thani, Ajayeb Al Marri, Abdulla Al Kaabi, Hadi Abderahim, Nahla Afifi, Fatima Qafoud, Queenie Chan, Ioanna Tzoulaki, Paul Downey, et al. The qatar biobank: background and methods. *BMC public health*, 15(1):1–9, 2015.
- [164] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. Mobilenetv2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4510–4520, 2018.

-
- [165] Shanchen Pang, Shuo Wang, Alfonso Rodríguez-Patón, Pibao Li, and Xun Wang. An artificial intelligent diagnostic system on mobile android terminals for cholelithiasis by lightweight convolutional neural network. *PLoS One*, 14(9):e0221720, 2019.
- [166] Ehsan Vaghefi, Song Yang, Sophie Hill, Gayl Humphrey, Natalie Walker, and David Squirrell. Detection of smoking status from retinal images; a convolutional neural network study. *Scientific reports*, 9(1):1–9, 2019.
- [167] Li Zhang, Mengya Yuan, Zhen An, Xiangmei Zhao, Hui Wu, Haibin Li, Ya Wang, Beibei Sun, Huijun Li, Shibin Ding, et al. 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*, 15(5):e0233166, 2020.
- [168] Takehiro Yamashita, Ryo Asaoka, Hiroto Terasaki, Hiroshi Murata, Minoru Tanaka, Kumiko Nakao, and Taiji Sakamoto. Factors in color fundus photographs that can be used by humans to determine sex of individuals. *Translational Vision Science & Technology*, 9(2):4–4, 2020.
- [169] Jooyoung Chang, Ahryoung Ko, Sang Min Park, Seulgie Choi, Kyuwoong Kim, Sung Min Kim, Jae Moon Yun, Uk Kang, Il Hyung Shin, Joo Young Shin, et al. Association of cardiovascular mortality and deep learning-funduscopy atherosclerosis score derived from retinal fundus images. *American Journal of Ophthalmology*, 217:121–130, 2020.
- [170] Zhuoting Zhu, Danli Shi, Peng Guankai, Zachary Tan, Xianwen Shang, Wenyi Hu, Huan Liao, Xueli Zhang, Yu Huang, Honghua Yu, et al. Retinal age gap as a predictive biomarker for mortality risk. *British Journal of Ophthalmology*, 2022.
- [171] Sharon Yu Lin Chua, Dhanes Thomas, Naomi Allen, Andrew Lotery, Parul Desai, Praveen Patel, Zaynah Muthy, Cathie Sudlow, Tunde Peto, Peng Tee Khaw, et al. Cohort profile: design and methods in the eye and vision consortium of uk biobank. *BMJ open*, 9(2):e025077, 2019.
- [172] François Chollet. Xception: Deep learning with depthwise separable convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1251–1258, 2017.
- [173] Charumathi Sabanayagam, Dejiang Xu, Daniel SW Ting, Simon Nusinovici, Riswana Banu, Haslina Hamzah, Cynthia Lim, Yih-Chung Tham, Carol Y Cheung, E Shyong Tai, et al. A deep learning algorithm to detect chronic kidney disease from retinal photographs in community-based populations. *The Lancet Digital Health*, 2(6):e295–e302, 2020.
- [174] Charumathi Sabanayagam, Wanfen Yip, Preeti Gupta, Riswana BB Mohd Abdul, Ecosse Lamoureux, Neelam Kumari, Gemmy CM Cheung, Carol Y Cheung, Jie Jin Wang, Ching-Yu Cheng, et al. Singapore indian eye study-2: methodology and impact of migration on systemic and eye outcomes. *Clinical & experimental ophthalmology*, 45(8):779–789, 2017.
- [175] Jie Xu, Liang Xu, Ya Xing Wang, Qi Sheng You, Jost B Jonas, and Wen Bin Wei. Ten-year cumulative incidence of diabetic retinopathy. the beijing eye study 2001/2011. *PLoS One*, 9(10):e111320, 2014.

-
- [176] Dejiang Xu, Mong Li Lee, and Wynne Hsu. Propagation mechanism for deep and wide neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9220–9228, 2019.
- [177] Kang Zhang, Xiaohong Liu, Jie Xu, Jin Yuan, Wenjia Cai, Ting Chen, Kai Wang, Yuanxu Gao, Sheng Nie, Xiaodong Xu, et al. Deep-learning models for the detection and incidence prediction of chronic kidney disease and type 2 diabetes from retinal fundus images. *Nature Biomedical Engineering*, 5(6):533–545, 2021.
- [178] Diego R Cervera, Luke Smith, Luis Diaz-Santana, Meenakshi Kumar, Rajiv Raman, and Sobha Sivaprasad. Identifying peripheral neuropathy in colour fundus photographs based on deep learning. *Diagnostics*, 11(11):1943, 2021.
- [179] Rajiv Raman, Padmaja Kumari Rani, Sudhir Reddi Racheppalle, Perumal Gnanamoorthy, Satagopan Uthra, Govindasamy Kumaramanickavel, and Tarun Sharma. Prevalence of diabetic retinopathy in india: Sankara nethralaya diabetic retinopathy epidemiology and molecular genetics study report 2. *Ophthalmology*, 116(2):311–318, 2009.
- [180] Forrest N Iandola, Song Han, Matthew W Moskewicz, Khalid Ashraf, William J Dally, and Kurt Keutzer. Squeezenet: Alexnet-level accuracy with 50x fewer parameters and < 0.5 mb model size. *arXiv preprint arXiv:1602.07360*, 2016.
- [181] Md Mohaimenul Islam, Hsuan-Chia Yang, Tahmina Nasrin Poly, Wen-Shan Jian, and Yu-Chuan Jack Li. Deep learning algorithms for detection of diabetic retinopathy in retinal fundus photographs: A systematic review and meta-analysis. *Computer Methods and Programs in Biomedicine*, 191:105320, 2020.
- [182] Varun Gulshan, Lily Peng, Marc Coram, Martin C Stumpe, Derek Wu, Arunachalam Narayanaswamy, Subhashini Venugopalan, Kasumi Widner, Tom Madams, Jorge Cuadros, et al. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. *Jama*, 316(22):2402–2410, 2016.
- [183] Chan Yoon, Eurah Goh, Sang Min Park, and Belong Cho. Effects of smoking cessation and weight gain on cardiovascular disease risk factors in asian male population. *Atherosclerosis*, 208(1):275–279, 2010.
- [184] Alexander Kurz, Katja Hauser, Hendrik Alexander Mehrrens, Eva Krieghoff-Henning, Achim Hekler, Jakob Nikolas Kather, Stefan Fröhling, Christof von Kalle, Titus Josef Brinker, et al. Uncertainty estimation in medical image classification: systematic review. *JMIR Medical Informatics*, 10(8):e36427, 2022.
- [185] A.D. Hoover, V. Kouznetsova, and M. Goldbaum. Locating blood vessels in retinal images by piecewise threshold probing of a matched filter response. *IEEE Transactions on Medical Imaging*, 19(3):203–210, 2000.
- [186] J. Staal, M.D. Abramoff, M. Niemeijer, M.A. Viergever, and B. van Ginneken. Ridge-based vessel segmentation in color images of the retina. *IEEE Transactions on Medical Imaging*, 23(4):501–509, 2004.

- [187] T. Kauppi, V. Kalesnykiene, J.-K. Kamarainen, L. Lensu, I. Sorri, A. Raninen, R. Voutilainen, H. Uusitalo, H. Kalviainen, and J. Pietila. the diaretdb1 diabetic retinopathy database and evaluation protocol. In *Proc. BMVC*, pages 15.1–15.10, 2007. doi:10.5244/C.21.15.
- [188] Meindert Niemeijer, Bram van Ginneken, Michael J. Cree, Atsushi Mizutani, GwÉnolÉ Quellec, Clara I. Sanchez, Bob Zhang, Roberto Hornero, Mathieu Lamard, Chisako Muramatsu, Xiangqian Wu, Guy Cazuguel, Jane You, AgustÍn Mayo, Qin Li, Yuji Hatanaka, BÉatrice Cochener, Christian Roux, Fakhri Karray, MarÍa Garcia, Hiroshi Fujita, and Michael D. Abramoff. Retinopathy online challenge: Automatic detection of microaneurysms in digital color fundus photographs. *IEEE Transactions on Medical Imaging*, 29(1):185–195, 2010.
- [189] Luca Giancardo, Fabrice Meriaudeau, Thomas P. Karnowski, Yaqin Li, Seema Garg, Kenneth W. Tobin, and Edward Chaum. Exudate-based diabetic macular edema detection in fundus images using publicly available datasets. *Medical Image Analysis*, 16(1):216–226, 2012.
- [190] E. DecenciÈre, G. Cazuguel, X. Zhang, G. Thibault, J.-C. Klein, F. Meyer, B. Marcotegui, G. Quellec, M. Lamard, R. Danno, D. Elie, P. Massin, Z. Viktor, A. Erginay, B. Laÿ, and A. Chabouis. Teleophta: Machine learning and image processing methods for teleophthalmology. *IRBM*, 34(2):196–203, 2013. Special issue : ANR TECSAN : Technologies for Health and Autonomy.
- [191] Kaggle and EyePacs. <https://www.kaggle.com/c/diabetic-retinopathy-detection/data>, 12th Aug. 2021.
- [192] MESSIDOR. <https://www.adcis.net/en/third-party/messidor/>, 12th Aug. 2021.
- [193] Harry L Hébert, Bridget Shepherd, Keith Milburn, Abirami Veluchamy, Weihua Meng, Fiona Carr, Louise A Donnelly, Roger Tavendale, Graham Leese, Helen M Colhoun, et al. Cohort profile: genetics of diabetes audit and research in tayside scotland (godarts). *International journal of epidemiology*, 47(2):380–381j, 2018.
- [194] SHARE. <https://www.registerforshare.org/>, 12th August 2021.
- [195] Scottish Diabetic Eye Screening. <https://www.ndrs.scot.nhs.uk/>, 12th August 2021.
- [196] Mustafa Adnan Malki, Adem Y Dawed, Caroline Hayward, Alex Doney, and Ewan R Pearson. Utilizing large electronic medical record data sets to identify novel drug–gene interactions for commonly used drugs. *Clinical Pharmacology & Therapeutics*, 110(3):816–825, 2021.
- [197] Scottish Diabetic Retinopathy Screening Programme ANNUAL REPORT 2017. <https://www.ndrs.scot.nhs.uk/wp-content/uploads/2019/11/DRS-Annual-Report-2017.pdf>, 29th Sep. 2022.
- [198] G. Bradski. The OpenCV Library. *Dr. Dobb’s Journal of Software Tools*, 2000.

-
- [199] Stephen M Pizer, E Philip Amburn, John D Austin, Robert Cromartie, Ari Geselowitz, Trey Greer, Bart ter Haar Romeny, John B Zimmerman, and Karel Zuiderveld. Adaptive histogram equalization and its variations. *Computer vision, graphics, and image processing*, 39(3):355–368, 1987.
- [200] Emanuele Trucco, Alfredo Ruggeri, Thomas Karnowski, Luca Giancardo, Edward Chaum, Jean Pierre Hubschman, Bashir Al-Diri, Carol Y Cheung, Damon Wong, Michael Abramoff, et al. Validating retinal fundus image analysis algorithms: issues and a proposal. *Investigative ophthalmology & visual science*, 54(5):3546–3559, 2013.
- [201] Lena Maier-Hein, Anja Groch, Adrien Bartoli, Sebastian Bodenstedt, G Boissonnat, P-L Chang, NT Clancy, Daniel S Elson, Sven Haase, Eric Heim, et al. Comparative validation of single-shot optical techniques for laparoscopic 3-d surface reconstruction. *IEEE transactions on medical imaging*, 33(10):1913–1930, 2014.
- [202] Leo Joskowicz, D Cohen, N Caplan, and Jacob Sosna. Automatic segmentation variability estimation with segmentation priors. *Medical image analysis*, 50:54–64, 2018.
- [203] Jeffrey De Fauw, Joseph R Ledsam, Bernardino Romera-Paredes, Stanislav Nikolov, Nenad Tomasev, Sam Blackwell, Harry Askham, Xavier Glorot, Brendan O’Donoghue, Daniel Visentin, et al. Clinically applicable deep learning for diagnosis and referral in retinal disease. *Nature medicine*, 24(9):1342–1350, 2018.
- [204] Vanya V Valindria, Ioannis Lavdas, Wenjia Bai, Konstantinos Kamnitsas, Eric O Aboagye, Andrea G Rockall, Daniel Rueckert, and Ben Glocker. Reverse classification accuracy: predicting segmentation performance in the absence of ground truth. *IEEE transactions on medical imaging*, 36(8):1597–1606, 2017.
- [205] Jamie Shotton, Ross Girshick, Andrew Fitzgibbon, Toby Sharp, Mat Cook, Mark Finocchio, Richard Moore, Pushmeet Kohli, Antonio Criminisi, Alex Kipman, et al. Efficient human pose estimation from single depth images. *IEEE transactions on pattern analysis and machine intelligence*, 35(12):2821–2840, 2012.
- [206] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *Communications of the ACM*, 63(11):139–144, 2020.
- [207] Andy Kitchen and Jarrel Seah. Deep generative adversarial neural networks for realistic prostate lesion mri synthesis. *arXiv preprint arXiv:1708.00129*, 2017.
- [208] Guang Li, Lu Bai, Chuanwei Zhu, Enhe Wu, and Ruibing Ma. A novel method of synthetic ct generation from mr images based on convolutional neural networks. In *2018 11th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI)*, pages 1–5. IEEE, 2018.
- [209] Faisal Mahmood, Richard Chen, and Nicholas J Durr. Unsupervised reverse domain adaptation for synthetic medical images via adversarial training. *IEEE transactions on medical imaging*, 37(12):2572–2581, 2018.

-
- [210] Pedro Costa, Adrian Galdran, Maria Ines Meyer, Meindert Niemeijer, Michael Abràmoff, Ana Maria Mendonça, and Aurélio Campilho. End-to-end adversarial retinal image synthesis. *IEEE transactions on medical imaging*, 37(3):781–791, 2017.
- [211] Talha Iqbal and Hazrat Ali. Generative adversarial network for medical images (mi-gan). *Journal of medical systems*, 42(11):1–11, 2018.
- [212] Philippe M Burlina, Neil Joshi, Katia D Pacheco, TY Alvin Liu, and Neil M Bressler. Assessment of deep generative models for high-resolution synthetic retinal image generation of age-related macular degeneration. *JAMA ophthalmology*, 137(3):258–264, 2019.
- [213] Yi-Chieh Liu, Hao-Hsiang Yang, C-H Huck Yang, Jia-Hong Huang, Meng Tian, Hiromasa Morikawa, Yi-Chang James Tsai, and Jesper Tegner. Synthesizing new retinal symptom images by multiple generative models. In *Asian Conference on Computer Vision*, pages 235–250. Springer, 2018.
- [214] Stéfan van der Walt, Johannes L. Schönberger, Juan Nunez-Iglesias, François Boulogne, Joshua D. Warner, Neil Yager, Emmanuelle Gouillart, Tony Yu, and the scikit-image contributors. scikit-image: image processing in Python. *PeerJ*, 2:e453, 6 2014.
- [215] Pauli Virtanen, Ralf Gommers, Travis E. Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni Burovski, Pearu Peterson, Warren Weckesser, Jonathan Bright, Stéfan J. van der Walt, Matthew Brett, Joshua Wilson, K. Jarrod Millman, Nikolay Mayorov, Andrew R. J. Nelson, Eric Jones, Robert Kern, Eric Larson, C J Carey, İlhan Polat, Yu Feng, Eric W. Moore, Jake VanderPlas, Denis Laxalde, Josef Perktold, Robert Cimrman, Ian Henriksen, E. A. Quintero, Charles R. Harris, Anne M. Archibald, Antônio H. Ribeiro, Fabian Pedregosa, Paul van Mulbregt, and SciPy 1.0 Contributors. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods*, 17:261–272, 2020.
- [216] GIMP. <https://www.gimp.org/>, 6th Aug. 2021.
- [217] Patrick Pérez, Michel Gangnet, and Andrew Blake. Poisson image editing. In *ACM SIGGRAPH 2003 Papers*, pages 313–318. 2003.
- [218] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2818–2826, 2016.
- [219] François Chollet. keras, 2015.
- [220] Muhammad Naseer Bajwa, Yoshinobu Taniguchi, Muhammad Imran Malik, Wolfgang Neumeier, Andreas Dengel, and Sheraz Ahmed. Combining fine-and coarse-grained classifiers for diabetic retinopathy detection. In *Annual Conference on Medical Image Understanding and Analysis*, pages 242–253. Springer, 2019.
- [221] IDRID Leaderboard. <https://idrid.grand-challenge.org/Leaderboard/>, 12th Aug. 2021.

- [222] Mingxing Tan, Bo Chen, Ruoming Pang, Vijay Vasudevan, Mark Sandler, Andrew Howard, and Quoc V Le. Mnasnet: Platform-aware neural architecture search for mobile. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2820–2828, 2019.
- [223] Sergey Zagoruyko and Nikos Komodakis. Wide residual networks. *arXiv preprint arXiv:1605.07146*, 2016.
- [224] Magda R Hamczyk, Rosa M Nevado, Ana Baretino, Valentín Fuster, and Vicente Andrés. Biological versus chronological aging: Jacc focus seminar. *Journal of the American College of Cardiology*, 75(8):919–930, 2020.
- [225] Anders M Fjell and Kristine B Walhovd. Structural brain changes in aging: courses, causes and cognitive consequences. *Reviews in the Neurosciences*, 21(3):187–222, 2010.
- [226] Faith M Gunning-Dixon, Adam M Brickman, Janice C Cheng, and George S Alexopoulos. Aging of cerebral white matter: a review of mri findings. *International Journal of Geriatric Psychiatry: A journal of the psychiatry of late life and allied sciences*, 24(2):109–117, 2009.
- [227] James H Cole, Stuart J Ritchie, Mark E Bastin, MC Valdés Hernández, S Muñoz Maniega, Natalie Royle, Janie Corley, Alison Pattie, Sarah E Harris, Qian Zhang, et al. Brain age predicts mortality. *Molecular psychiatry*, 23(5):1385–1392, 2018.
- [228] Vineet K Raghunath, Jakob Weiss, Udo Hoffmann, Hugo JWL Aerts, and Michael T Lu. Deep learning to estimate biological age from chest radiographs. *JACC: Cardiovascular Imaging*, 14(11):2226–2236, 2021.
- [229] A Le Goallec, S Diai, S Collin, T Vincent, and CJ Patel. Identifying the genetic and non-genetic factors associated with accelerated eye aging by using deep learning to predict age from fundus and optical coherence tomography images. 2021.
- [230] Christopher A Girkin, Gerald McGwin Jr, Micheal J Sinai, G Chandra Sekhar, Murrey Fingeret, Gadi Wollstein, Rohit Varma, David Greenfield, Jeffery Liebmann, Makoto Araie, et al. Variation in optic nerve and macular structure with age and race with spectral-domain optical coherence tomography. *Ophthalmology*, 118(12):2403–2408, 2011.
- [231] Amit V Khera, Mark Chaffin, Krishna G Aragam, Mary E Haas, Carolina Roselli, Seung Hoan Choi, Pradeep Natarajan, Eric S Lander, Steven A Lubitz, Patrick T Ellinor, et al. Genome-wide polygenic scores for common diseases identify individuals with risk equivalent to monogenic mutations. *Nature genetics*, 50(9):1219–1224, 2018.
- [232] Peter H Scanlon, R Malhotra, G Thomas, C Foy, JN Kirkpatrick, N Lewis-Barned, B Harney, and SJ Aldington. The effectiveness of screening for diabetic retinopathy by digital imaging photography and technician ophthalmoscopy. *Diabetic medicine*, 20(6):467–474, 2003.
- [233] Kaggle. <https://en.wikipedia.org/wiki/Kaggle>, 16th Aug. 2022.

-
- [234] William Vorhies. <https://www.analytikus.com/post/2017/04/03/has-deep-learning-made-traditional-machine-learning-irrelevant>, 16th Aug. 2022.
- [235] Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pages 785–794, 2016.
- [236] Hoang Thanh Lam, Tran Ngoc Minh, Mathieu Sinn, Beat Buesser, and Martin Wistuba. Neural feature learning from relational database. *arXiv preprint arXiv:1801.05372*, 2018.
- [237] An Introduction to Deep Learning for Tabular Data. <https://www.fast.ai/2018/04/29/categorical-embeddings/>, 16th Aug. 2022.
- [238] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. *Advances in neural information processing systems*, 26, 2013.
- [239] Jeffrey Pennington, Richard Socher, and Christopher D Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543, 2014.
- [240] Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. Bag of tricks for efficient text classification. *arXiv preprint arXiv:1607.01759*, 2016.
- [241] Matthew E Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. Deep contextualized word representations. *corr abs/1802.05365* (2018). *arXiv preprint arXiv:1802.05365*, 1802.
- [242] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.
- [243] Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. Improving language understanding by generative pre-training. 2018.
- [244] Baohua Sun, Lin Yang, Patrick Dong, Wenhan Zhang, Jason Dong, and Charles Young. Super characters: A conversion from sentiment classification to image classification. *arXiv preprint arXiv:1810.07653*, 2018.
- [245] Baohua Sun, Lin Yang, Wenhan Zhang, Michael Lin, Patrick Dong, Charles Young, and Jason Dong. Supertml: Two-dimensional word embedding for the precognition on structured tabular data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 0–0, 2019.
- [246] Boyu Lyu and Anamul Haque. Deep learning based tumor type classification using gene expression data. In *Proceedings of the 2018 ACM international conference on bioinformatics, computational biology, and health informatics*, pages 89–96, 2018.

- [247] Alok Sharma, Edwin Vans, Daichi Shigemizu, Keith A Boroevich, and Tatsuhiko Tsunoda. Deepinsight: A methodology to transform a non-image data to an image for convolution neural network architecture. *Scientific reports*, 9(1):1–7, 2019.
- [248] Ljubomir Buturović and Dejan Miljković. A novel method for classification of tabular data using convolutional neural networks. *bioRxiv*, 2020.
- [249] Yitan Zhu, Thomas Brettin, Fangfang Xia, Alexander Partin, Maulik Shukla, Hyunseung Yoo, Yvonne A Evrard, James H Doroshov, and Rick L Stevens. Converting tabular data into images for deep learning with convolutional neural networks. *Scientific reports*, 11(1):1–11, 2021.
- [250] Ify R Mordi, Emanuele Trucco, Mohammad Ghouse Syed, Tom MacGillivray, Adi Nar, Yu Huang, Gittu George, Stephen Hogg, Venkatesan Radha, Vijayaraghavan Prathiba, et al. Prediction of major adverse cardiovascular events from retinal, clinical, and genomic data in individuals with type 2 diabetes: A population cohort study. *Diabetes Care*, 45(3):710–716, 2022.
- [251] Mirna Kirin, Reka Nagy, Thomas J MacGillivray, Ozren Polašek, Caroline Hayward, Igor Rudan, Harry Campbell, Sarah Wild, Alan F Wright, James F Wilson, et al. Determinants of retinal microvascular features and their relationships in two european populations. *Journal of hypertension*, 35(8):1646, 2017.
- [252] Devanjali Relan, T MacGillivray, Lucia Ballerini, and Emanuele Trucco. Automatic retinal vessel classification using a least square-support vector machine in vampire. In *2014 36th annual international conference of the IEEE Engineering in medicine and biology society*, pages 142–145. IEEE, 2014.
- [253] E Trucco, L Ballerini, D Relan, A Giachetti, T MacGillivray, K Zutis, C Lupascu, D Tegolo, E Pellegrini, G Robertson, et al. Novel vampire algorithms for quantitative analysis of the retinal vasculature. In *2013 ISSNIP Biosignals and Biorobotics Conference: Biosignals and Robotics for Better and Safer Living (BRC)*, pages 1–4. IEEE, 2013.
- [254] Itamar Arel, Derek C Rose, and Thomas P Karnowski. Deep machine learning—a new frontier in artificial intelligence research [research frontier]. *IEEE computational intelligence magazine*, 5(4):13–18, 2010.
- [255] Hassan Ismail Fawaz, Germain Forestier, Jonathan Weber, Lhassane Idoumghar, and Pierre-Alain Muller. Deep learning for time series classification: a review. *Data mining and knowledge discovery*, 33(4):917–963, 2019.
- [256] David R Cox. The regression analysis of binary sequences. *Journal of the Royal Statistical Society: Series B (Methodological)*, 20(2):215–232, 1958.
- [257] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. Scikit-learn: Machine learning in python. *the Journal of machine Learning research*, 12:2825–2830, 2011.
- [258] LogisticRegression in sklearn. https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LogisticRegression.html, 19th Aug. 2022.

-
- [259] GradientBoostingClassifier in sklearn. <https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.GradientBoostingClassifier.html>, 19th Aug. 2022.
- [260] Ann-Christine Syvänen. Accessing genetic variation: genotyping single nucleotide polymorphisms. *Nature Reviews Genetics*, 2(12):930–942, 2001.
- [261] Vivian Tam, Nikunj Patel, Michelle Turcotte, Yohan Bossé, Guillaume Paré, and David Meyre. Benefits and limitations of genome-wide association studies. *Nature Reviews Genetics*, 20(8):467–484, 2019.
- [262] Mohammad Ghouse Syed, Alexander Doney, Gittu George, Ify Mordi, and Emanuele Trucco. Are cardiovascular risk scores from genome and retinal image complementary? a deep learning investigation in a diabetic cohort. In *International Workshop on Ophthalmic Medical Image Analysis*, pages 109–118. Springer, 2021.
- [263] Chetan L Srinidhi, P Aparna, and Jeny Rajan. Recent advancements in retinal vessel segmentation. *Journal of medical systems*, 41(4):1–22, 2017.
- [264] Salome Kazemina, Christoph Baur, Arjan Kuijper, Bram van Ginneken, Nassir Navab, Shadi Albarqouni, and Anirban Mukhopadhyay. Gans for medical image analysis. *Artificial Intelligence in Medicine*, 109:101938, 2020.
- [265] Fabio Henrique Kiyoyiti dos Santos Tanaka and Claus Aranha. Data augmentation using gans. *arXiv preprint arXiv:1904.09135*, 2019.
- [266] Valentina Bellemo, Philippe Burlina, Liu Yong, Tien Yin Wong, and Daniel Shu Wei Ting. Generative adversarial networks (gans) for retinal fundus image synthesis. In *Asian Conference on Computer Vision*, pages 289–302. Springer, 2018.
- [267] Linwei Wang, Qi Dou, P Thomas Fletcher, Stefanie Speidel, and Shuo Li. *Medical Image Computing and Computer Assisted Intervention—MICCAI 2022: 25th International Conference, Singapore, September 18–22, 2022, Proceedings, Part VIII*, volume 13438. Springer Nature, 2022.
- [268] Why We Will Never Open Deep Learning’s Black Box. <https://towardsdatascience.com/why-we-will-never-open-deep-learnings-black-box-4c27cd335118>, 21st Sep. 2022.
- [269] Artificial intelligence pioneer says we need to start over. <https://www.axios.com/2017/12/15/artificial-intelligence-pioneer-says-we-need-to-start-over-1513305524>, 21st Sep. 2022.
- [270] Vishal Monga, Yuelong Li, and Yonina C Eldar. Algorithm unrolling: Interpretable, efficient deep learning for signal and image processing. *IEEE Signal Processing Magazine*, 38(2):18–44, 2021.
- [271] Alejandro Barredo Arrieta, Natalia Díaz-Rodríguez, Javier Del Ser, Adrien Bennetot, Siham Tabik, Alberto Barbado, Salvador García, Sergio Gil-López, Daniel Molina, Richard Benjamins, et al. Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai. *Information fusion*, 58:82–115, 2020.

- [272] Cynthia Rudin. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1(5):206–215, 2019.
- [273] Chaofan Chen, Oscar Li, Daniel Tao, Alina Barnett, Cynthia Rudin, and Jonathan K Su. This looks like that: deep learning for interpretable image recognition. *Advances in neural information processing systems*, 32, 2019.
- [274] Aaron S Coyner, Ryan Swan, J Peter Campbell, Susan Ostmo, James M Brown, Jayashree Kalpathy-Cramer, Sang Jin Kim, Karyn E Jonas, RV Paul Chan, Michael F Chiang, et al. Automated fundus image quality assessment in retinopathy of prematurity using deep convolutional neural networks. *Ophthalmology Retina*, 3(5):444–450, 2019.
- [275] Joshua Bridge, Simon Harding, and Yalin Zheng. Development and validation of a novel prognostic model for predicting amd progression using longitudinal fundus images. *BMJ open ophthalmology*, 5(1):e000569, 2020.

Appendix A

Additional figures and results

This appendix presents some additional tables, Gradient-based Class Activation Mapping (Grad-CAM) heatmaps, Kaplan-Meier (KM) plots, and coxph regression results from only left-eye or right-eye retinal images.

Table A.1 Age sub-group-wise model performance on the complete test data using only left eye images.

Age group	# Images	Mean actual age	Mean predicted age	MAE (95% CI)
0-10	0	-	-	-
10-20	0	-	-	-
20-30	27	27.457	33.202	5.925(4.76,7.006)
30-40	182	36.671	40.239	4.328(3.74,4.941)
40-50	667	45.881	48.638	4.201(3.946,4.473)
50-60	1578	55.671	58.524	3.998(3.83,4.166)
60-70	3160	65.157	66.275	3.538(3.444,3.635)
70-80	3487	74.807	73.603	3.604(3.509,3.692)
80-90	1210	83.326	78.951	4.99(4.788,5.186)
90-100	94	92.393	83.524	8.886(8.137,9.676)

This section consists of KM plots and coxph regression results for predicting 10-years Major Adverse Cardiovascular Event (MACE) and All Cause Death (ACD) using the Predicted Age Difference (PAD) computed from the retinal images of only left eyes or only right eyes.

Table A.2 Age sub-group-wise model performance on the complete test data using only right eye images.

Age group	# Images	Mean actual age	Mean predicted age	MAE (95% CI)
0-10	0	-	-	-
10-20	0	-	-	-
20-30	27	27.457	32.596	5.173(4.051,6.397)
30-40	180	36.605	40.507	4.599(3.986,5.24)
40-50	658	45.886	49.334	4.532(4.236,4.84)
50-60	1557	55.687	58.609	4.308(4.139,4.485)
60-70	3137	65.159	66.172	3.594(3.496,3.699)
70-80	3436	74.794	73.701	3.548(3.463,3.635)
80-90	1200	83.307	78.842	5.102(4.899,5.315)
90-100	94	92.471	83.626	8.839(8.195,9.468)

Table A.3 Age sub-group-wise model performance on the test data of T2D individuals using both left and right eye image predictions at baseline.

Age group	# Images	Mean actual age	Mean predicted age	MAE (95% CI)
0-10	0	-	-	-
10-20	0	-	-	-
20-30	0	-	-	-
30-40	30	36.719	42.672	5.943(4.791,7.191)
40-50	168	45.761	49.932	4.822(4.272,5.394)
50-60	466	55.769	59.129	4.409(4.091,4.75)
60-70	788	64.988	66.242	3.641(3.443,3.851)
70-80	860	74.724	73.912	3.618(3.434,3.804)
80-90	306	83.282	79.737	4.336(3.972,4.702)
90-100	14	92.134	84.122	7.998(6.775,9.058)

Table A.4 Age sub-group-wise model performance on the test data of T2D individuals using only left eye image predictions at baseline.

Age group	# Images	Mean actual age	Mean predicted age	MAE (95% CI)
0-10	0	-	-	-
10-20	0	-	-	-
20-30	0	-	-	-
30-40	15	36.719	42.608	5.905(4.31,7.675)
40-50	84	45.761	49.755	4.733(3.963,5.518)
50-60	233	55.769	59.540	4.736(4.286,5.207)
60-70	394	64.988	66.466	3.694(3.428,3.986)
70-80	430	74.724	73.881	3.779(3.52,4.048)
80-90	153	83.282	79.671	4.347(3.824,4.904)
90-100	7	92.134	84.310	7.817(6.913,8.582)

Table A.5 Age sub-group-wise model performance on the test data of T2D individuals using only right eye image predictions at baseline.

Age group	# Images	Mean actual age	Mean predicted age	MAE (95% CI)
0-10	0	-	-	-
10-20	0	-	-	-
20-30	0	-	-	-
30-40	15	36.719	42.736	6.041(4.355,7.996)
40-50	84	45.761	50.108	4.911(4.184,5.677)
50-60	233	55.769	58.719	4.094(3.671,4.549)
60-70	394	64.988	66.018	3.591(3.319,3.87)
70-80	430	74.724	73.944	3.453(3.204,3.717)
80-90	153	83.282	79.803	4.307(3.84,4.785)
90-100	7	92.134	83.934	8.196(5.917,10.226)

Table A.6 Age sub-group-wise model performance on the test data of T2D individuals using the average of left and right eye image predictions at baseline.

Age group	# Images	Mean actual age	Mean predicted age	MAE (95% CI)
0-10	0	-	-	-
10-20	0	-	-	-
20-30	0	-	-	-
30-40	15	36.719	42.672	5.943(4.393,7.693)
40-50	84	45.761	49.932	4.588(3.888,5.337)
50-60	233	55.769	59.129	4.168(3.763,4.583)
60-70	394	64.988	66.242	3.275(3.031,3.548)
70-80	430	74.724	73.912	3.338(3.105,3.579)
80-90	153	83.282	79.737	4.111(3.646,4.6)
90-100	7	92.134	84.122	8.027(6.603,9.34)

Table A.7 Coxph regression results adjusted for age, and sex for predicting MACE and ACD using PAD from only left- and right-eye retinal images. PAD score is a continuous variable. The results of PAD Middle and high tertile are with respect to PAD low tertile group.

Eye	Outcome	# individuals	# events	feature	HR	conf.low	conf.high	p.value
Left	Mortality	1316	466	PAD score	1.04	1.02	1.06	0.0004
				PAD Middle tertile	1.38	1.11	1.71	0.004
				PAD high tertile	1.49	1.15	1.94	0.003
	MACE	1316	305	PAD score	1.04	1.02	1.07	0.0009
				PAD Middle tertile	1.17	0.9	1.53	0.3
				PAD high tertile	1.52	1.11	2.08	0.009
Right	Mortality	1316	466	PAD score	1.06	1.04	1.08	2.68e-07
				PAD Middle tertile	1.24	0.1	1.55	0.05
				PAD high tertile	1.63	1.27	2.1	0.0001
	MACE	1316	305	PAD score	1.05	1.02	1.08	0.0002
				PAD Middle tertile	1.15	0.87	1.51	0.3
				PAD high tertile	1.55	1.14	2.11	0.005

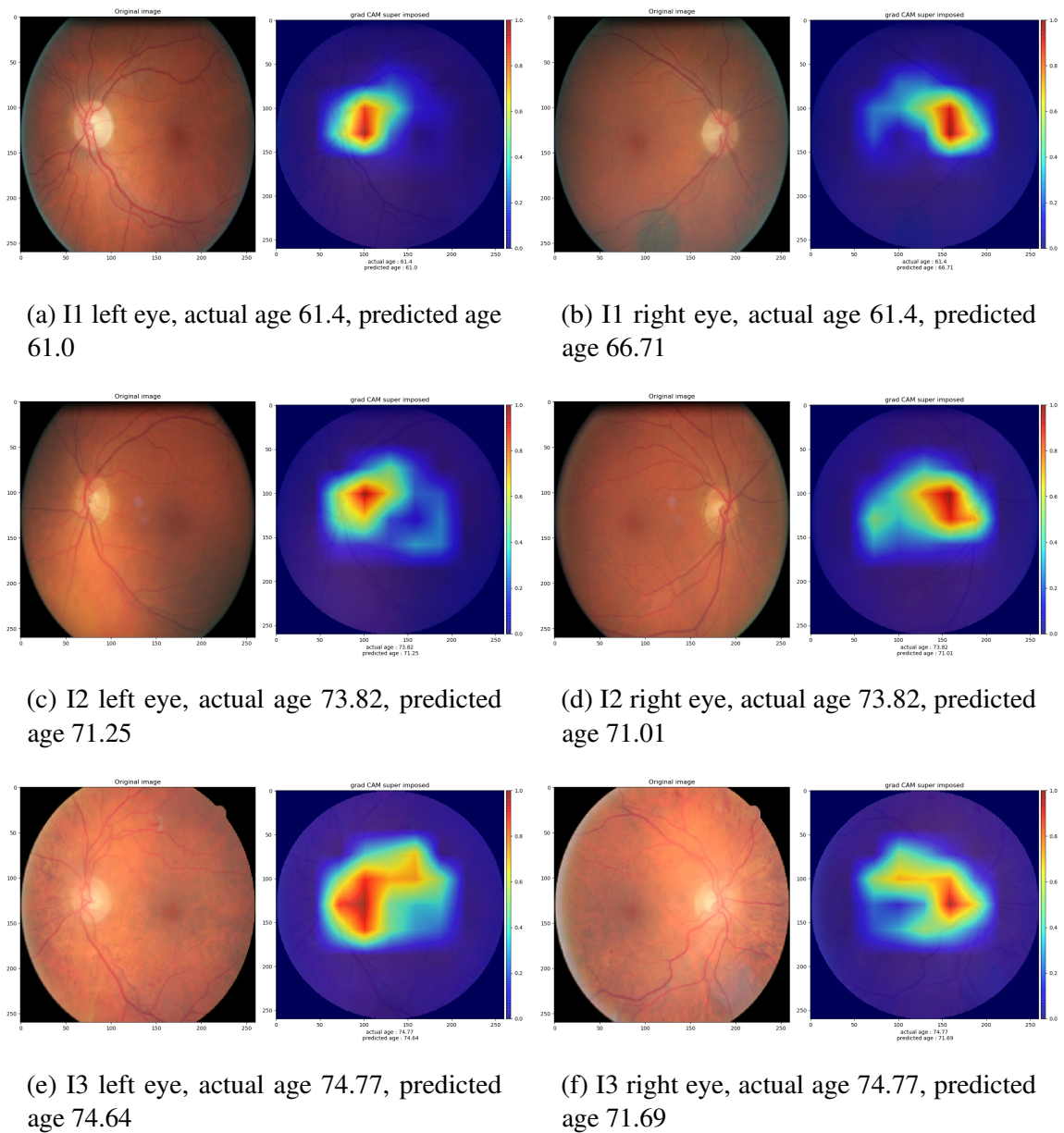
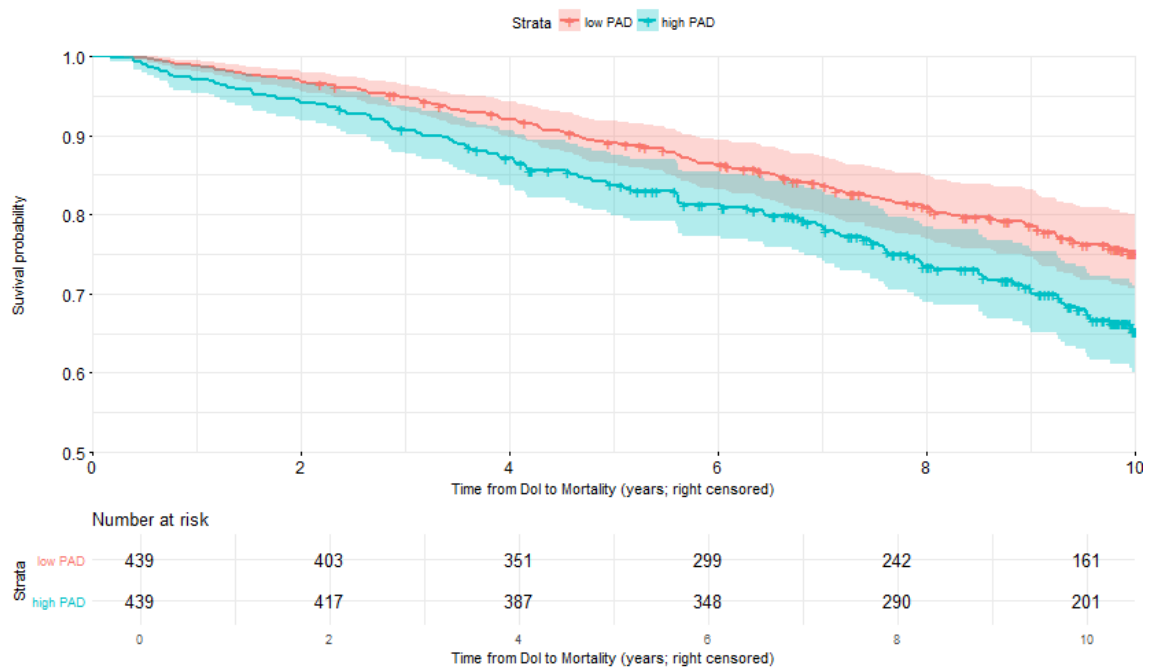
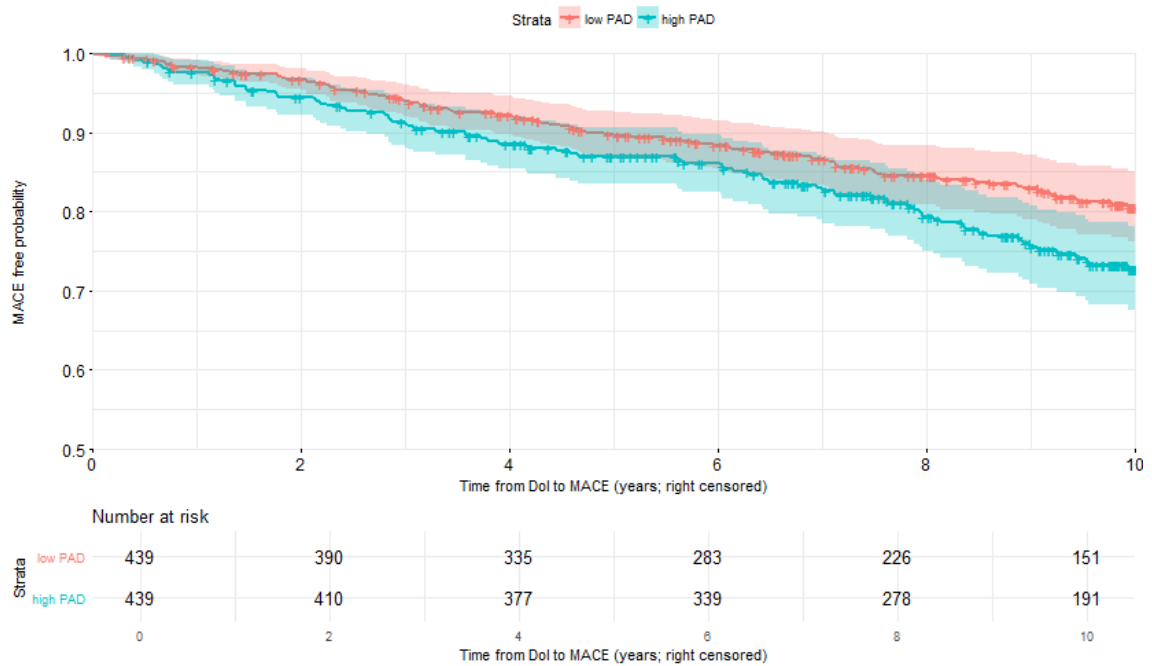


Fig. A.1 Sample grad-CAM heatmaps of three individuals from model trained for age prediction in T2D individuals. I1 = Individual 1, I2 = Individual 2, I3 = Individual 3.

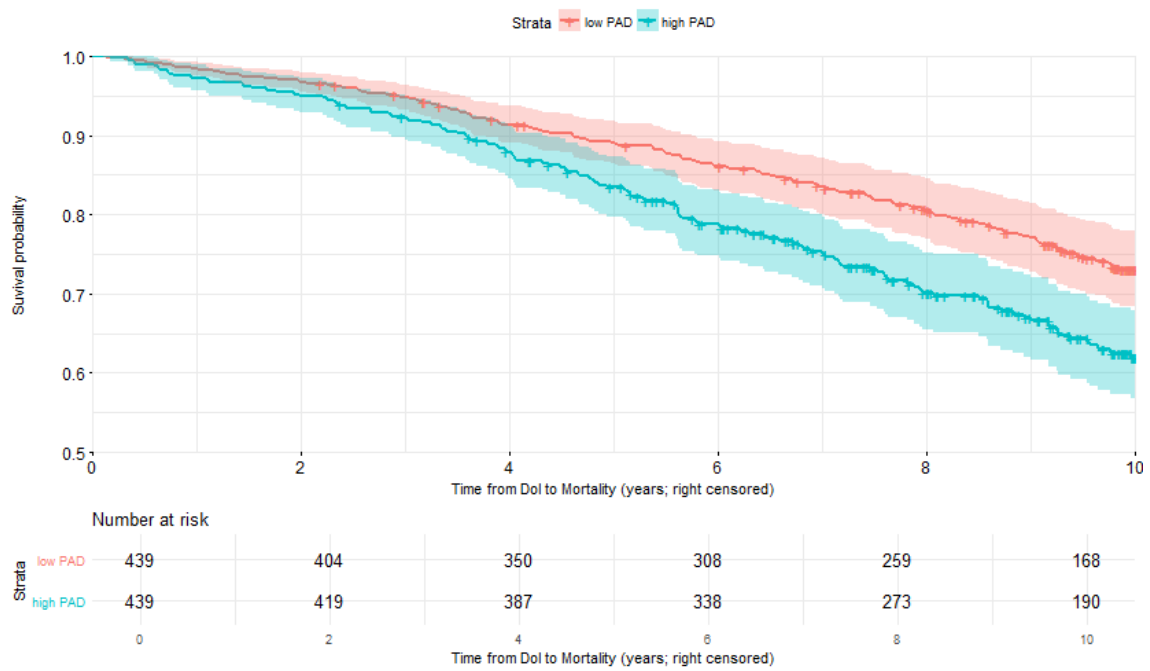


(a) KM curves for mortality

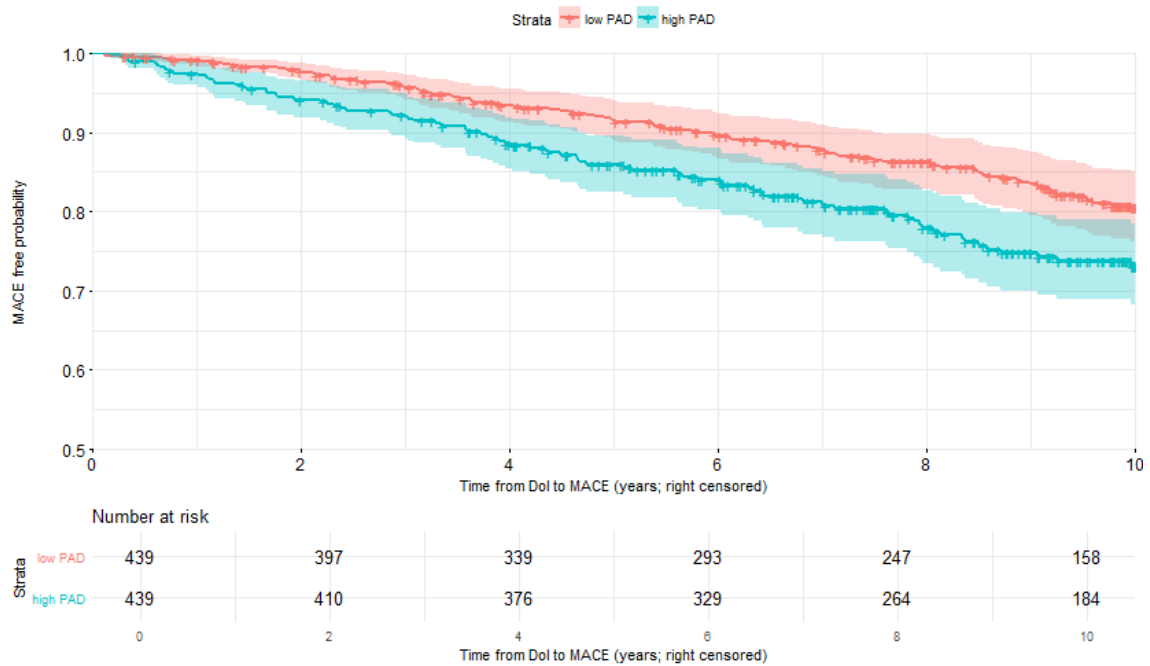


(b) KM curves for MACE

Fig. A.2 KM curves of right-censored survival data for upper (high PAD) and lower (low PAD) tertiles of retinal PAD groups using only left eye image predictions for age.



(a) KM curves for mortality



(b) KM curves for MACE

Fig. A.3 KM curves of right-censored survival data for upper (high PAD) and lower (low PAD) tertiles of retinal PAD groups using only right eye image predictions for age.

Table A.8 Coxph regression results adjusted for CV risk score for predicting MACE and ACD using PAD from only left- and right-eye retinal images. PAD score is a continuous variable. The results of PAD Middle and high tertile are with respect to PAD low tertile group.

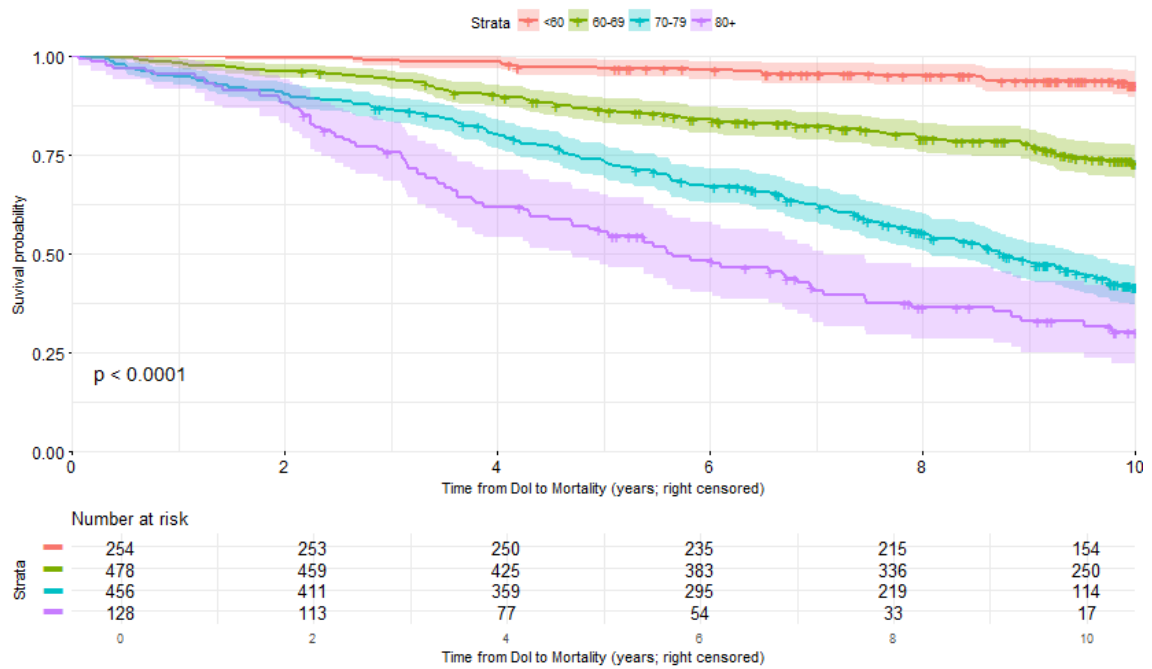
Eye	Outcome	# individuals	# events	feature	HR	conf.low	conf.high	p.value
Left	Mortality	1316	466	PAD score	1.02	0.1	1.04	0.08
				PAD Middle tertile	1.34	1.08	1.67	0.008
				PAD high tertile	1.24	0.95	1.61	0.1
	MACE	1316	305	PAD score	1.03	1.003	1.05	0.02
				PAD Middle tertile	1.15	0.88	1.52	0.3
				PAD high tertile	1.32	0.97	1.81	0.08
Right	Mortality	1316	466	PAD score	1.04	1.02	1.06	0.0005
				PAD Middle tertile	1.1	0.89	1.37	0.3
				PAD high tertile	1.37	1.07	1.76	0.01
	MACE	1316	305	PAD score	1.04	1.01	1.06	0.006
				PAD Middle tertile	1.05	0.8	1.37	0.7
				PAD high tertile	1.36	1.004	1.84	0.05

Table A.9 Coxph regression results adjusted for sex for predicting MACE and ACD using DLPA and CA from only left- and right-eye retinal images. DLPA = DL predicted age, CA = chronological age. DLPA and CA are continuous variables.

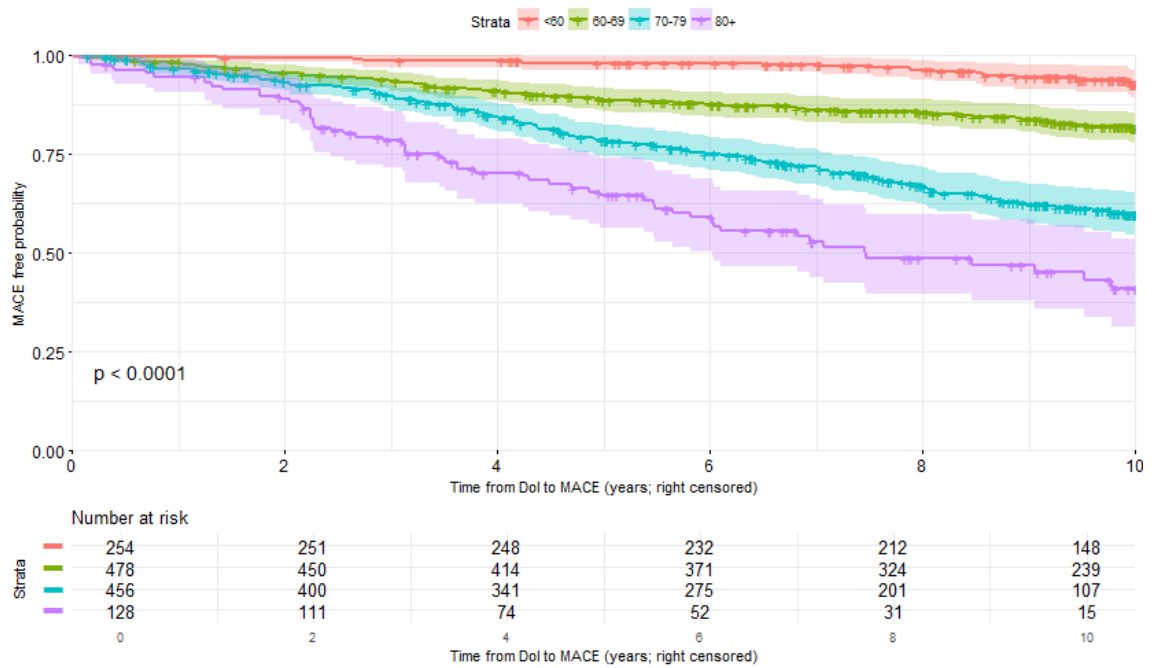
Eye	Outcome	# individuals	# events	Feature	HR	conf.low	conf.high	p.value
Left	Mortality	1316	466	DLPA	1.104	1.09	1.117	<2e-16
				CA	1.097	1.084	1.109	<2e-16
	MACE	1316	305	DLPA	1.09	1.08	1.113	<2e-16
				CA	1.088	1.07	1.101	<2e-16
Right	Mortality	1316	466	DLPA	1.111	1.098	1.12	<2e-16
				CA	1.097	1.085	1.108	<2e-16
	MACE	1316	305	DLPA	1.09	1.08	1.116	<2e-16
				CA	1.088	1.074	1.102	<2e-16

Table A.10 Coxph regression results adjusted for predicted age and sex for predicting ACD using τ_{rate} from only left- and right-eye retinal images. $\tau_{rate_{lt}} = \tau_{rate}$ low tertile, $\tau_{rate_{mt}} = \tau_{rate}$ middle tertile, $\tau_{rate_{ht}} = \tau_{rate}$ high tertile. The results of $\tau_{rate_{mt}}$ group and $\tau_{rate_{ht}}$ group are with respect to $\tau_{rate_{lt}}$ group.

Eye	Outcome	# individuals	# events	Feature	HR	conf.low	conf.high	p.value
Left	Mortality	1172	326	$\tau_{rate_{mt}}$	1.31	0.995	1.71	0.0535
				$\tau_{rate_{ht}}$	1.62	1.23	2.14	0.000607
Right	Mortality	1172	326	$\tau_{rate_{mt}}$	0.8	0.608	1.05	0.108
				$\tau_{rate_{ht}}$	1.26	0.967	1.65	8.70e-02

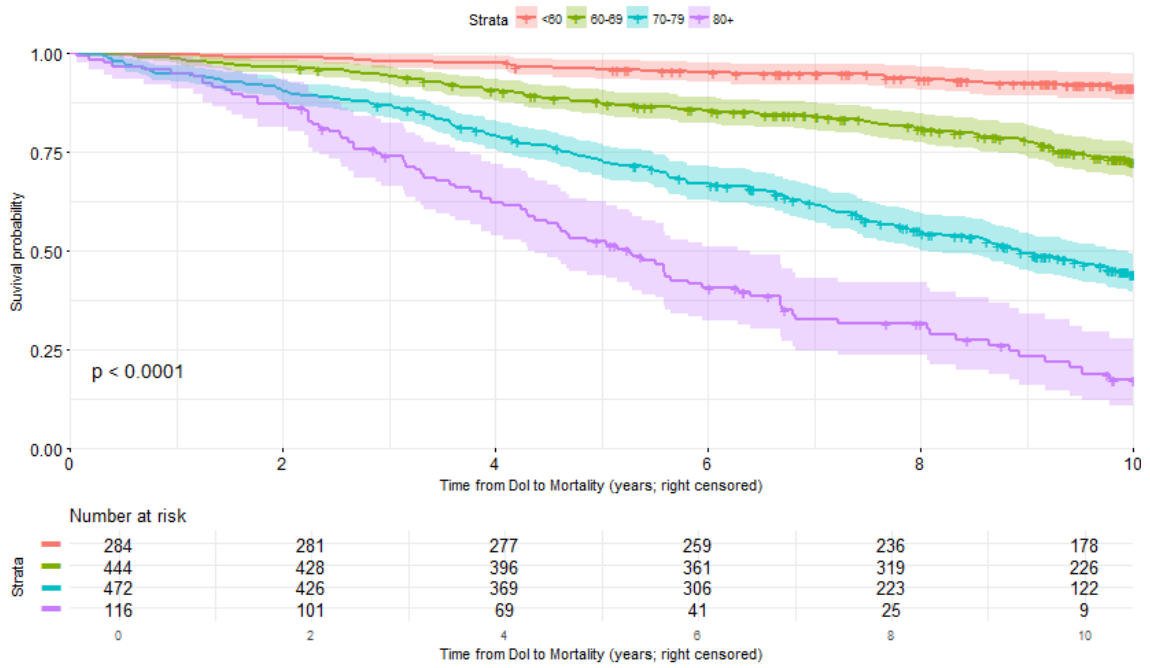


(a) KM curves for mortality

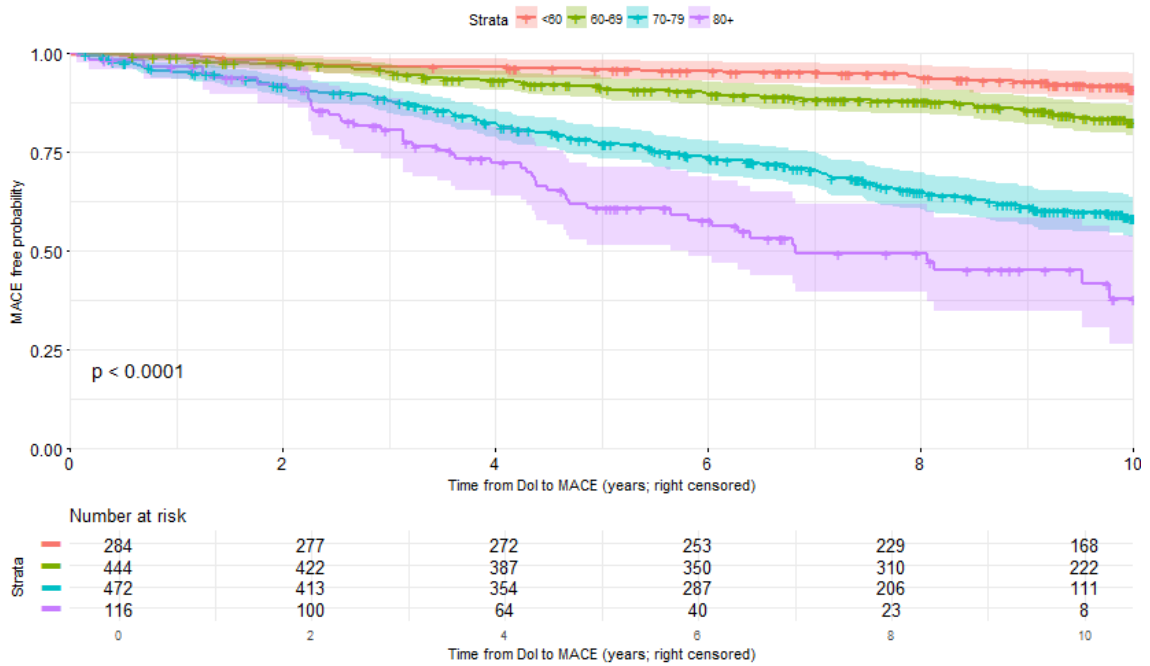


(b) KM curves for MACE

Fig. A.4 KM curves of right-censored survival data for groups of predicted age computed from only left eye retinal images



(a) KM curves for mortality



(b) KM curves for MACE

Fig. A.5 KM curves of right-censored survival data for groups of predicted age computed from only right eye retinal images

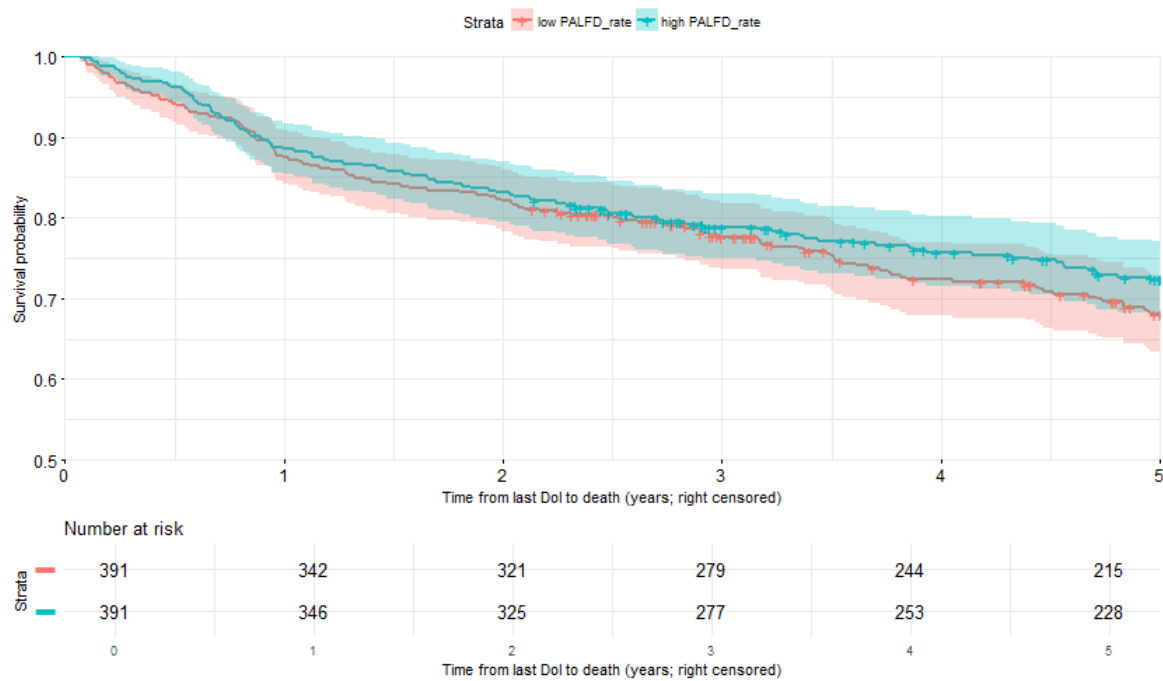


Fig. A.6 KM curves for mortality from right censored survival data for upper (high τ_{rate}) and lower (low τ_{rate}) tertiles of τ_{rate} groups computed from only left eye predictions. τ_{rate} = PALFD rate.

Table A.11 Coxph regression results adjusted for age, sex and GWPRS for predicting MACE and CV Death using PRS from only left- and right-eye retinal images. PRS score is a continuous variable. The results of PRS middle and high tertile are with respect to PRS low tertile group.

Eye	Outcome	# individuals	# events	feature	HR	conf.low	conf.high	p.value
Left	MACE	1273	317	PRS score	1.02	1.01	1.03	4.40E-04
				PRS Middle tertile	1.19	0.82	1.72	3.62E-01
				PRS high tertile	1.7	1.1	2.62	1.71E-02
	CV Death	1273	255	PRS score	1.02	1.003	1.03	1.73E-02
				PRS Middle tertile	1.15	0.74	1.78	5.34E-01
				PRS high tertile	1.49	0.9	2.46	1.24E-01
Right	MACE	1273	317	PRS score	1.02	1.01	1.03	1.36E-04
				PRS Middle tertile	1.37	0.94	2.006	9.89E-02
				PRS high tertile	1.88	1.22	2.89	4.18E-03
	CV Death	1273	255	PRS score	1.01	1.001	1.03	2.85E-02
				PRS Middle tertile	1.61	1.02	2.54	4.02E-02
				PRS high tertile	2.05	1.22	3.44	6.46E-03

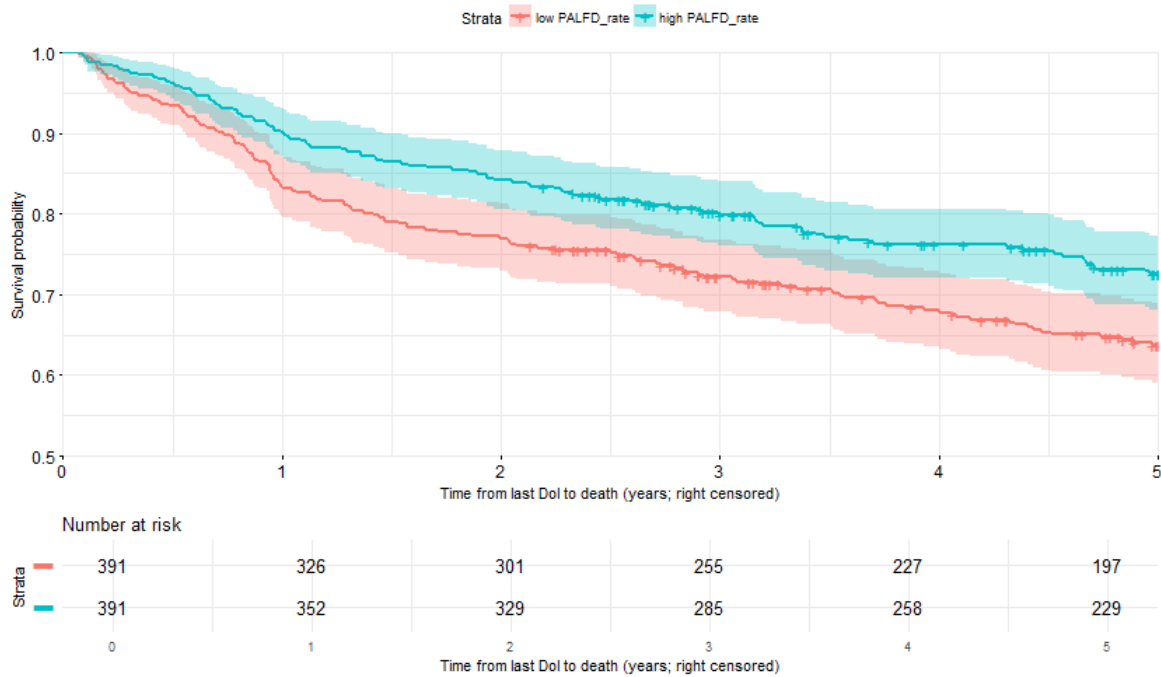
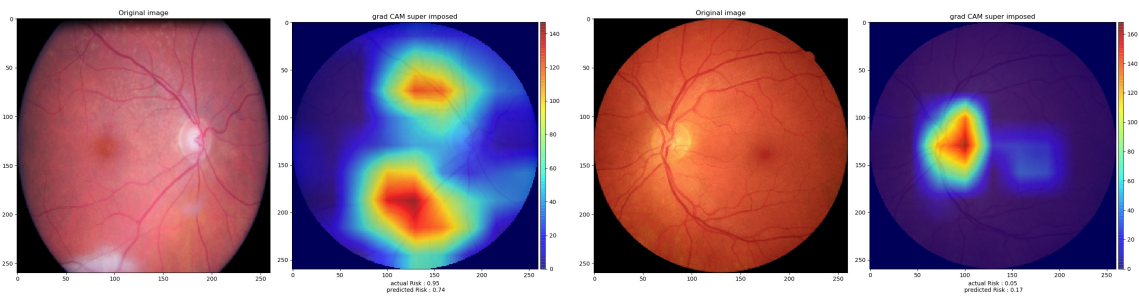


Fig. A.7 KM curves for mortality from right censored survival data for upper (high τ_{rate}) and lower (low τ_{rate}) tertiles of τ_{rate} groups computed from only right eye predictions. τ_{rate} = PALFD rate.

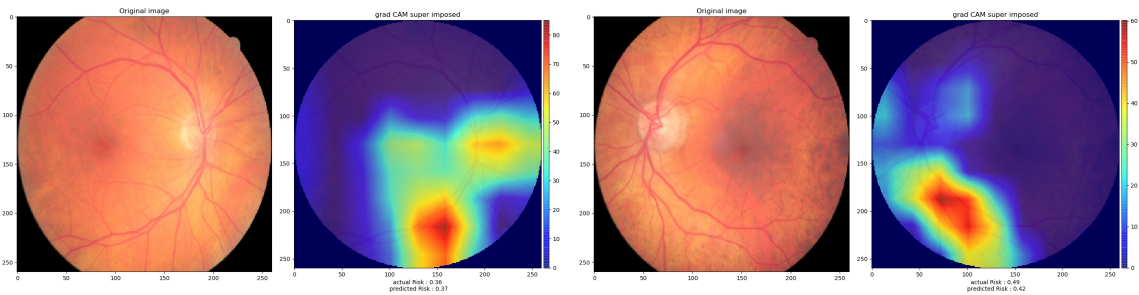
Table A.12 Coxph regression results adjusted for predicted age, sex and GWPRS for predicting MACE and CV Death using Ω_{rate} from only left- and right-eye retinal images. $\Omega_{rate_{t20}} = \Omega_{rate}$ Top 20%, $\Omega_{rate_{b80}} = \Omega_{rate}$ Bottom 20% . The results of $\Omega_{rate_{t20}}$ are with respect to $\Omega_{rate_{b80}}$ group.

Eye	Outcome	# individuals	# events	Feature	HR	conf.low	conf.high	p.value
Left	MACE	1091	219	$\Omega_{rate_{t20}}$	1.33	0.98	1.792	6.39E-02
	CV Death	1112	190	$\Omega_{rate_{t20}}$	1.44	1.05	1.98	2.26E-02
Right	MACE	1091	219	$\Omega_{rate_{t20}}$	1.295	0.96	1.76	9.54E-02
	CV Death	1112	190	$\Omega_{rate_{t20}}$	1.41	1.03	1.93	3.24E-02



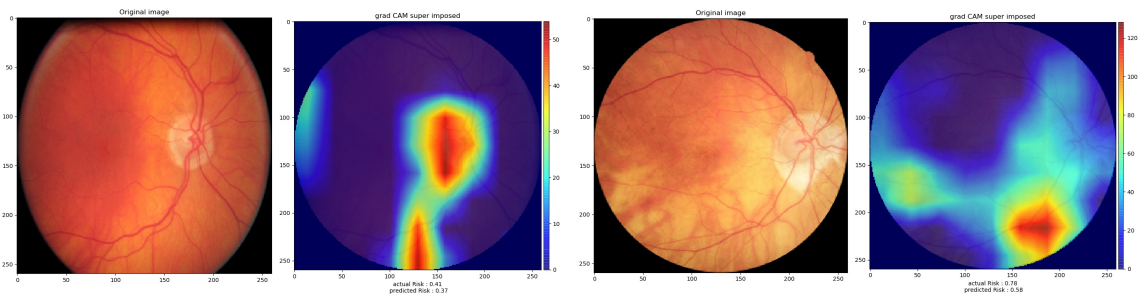
(a) Right eye, actual risk 0.95, predicted risk 0.74

(b) Left eye, actual risk 0.05, predicted risk 0.17



(c) Right eye, actual risk 0.36, predicted risk 0.37

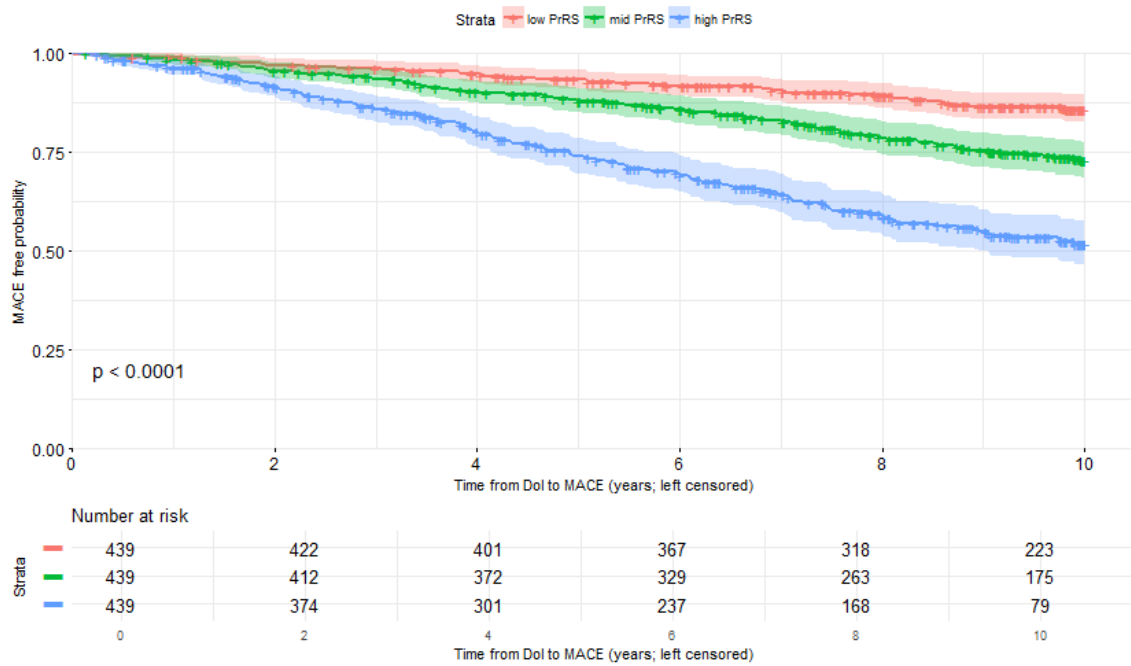
(d) Left eye, actual risk 0.49, predicted risk 0.42



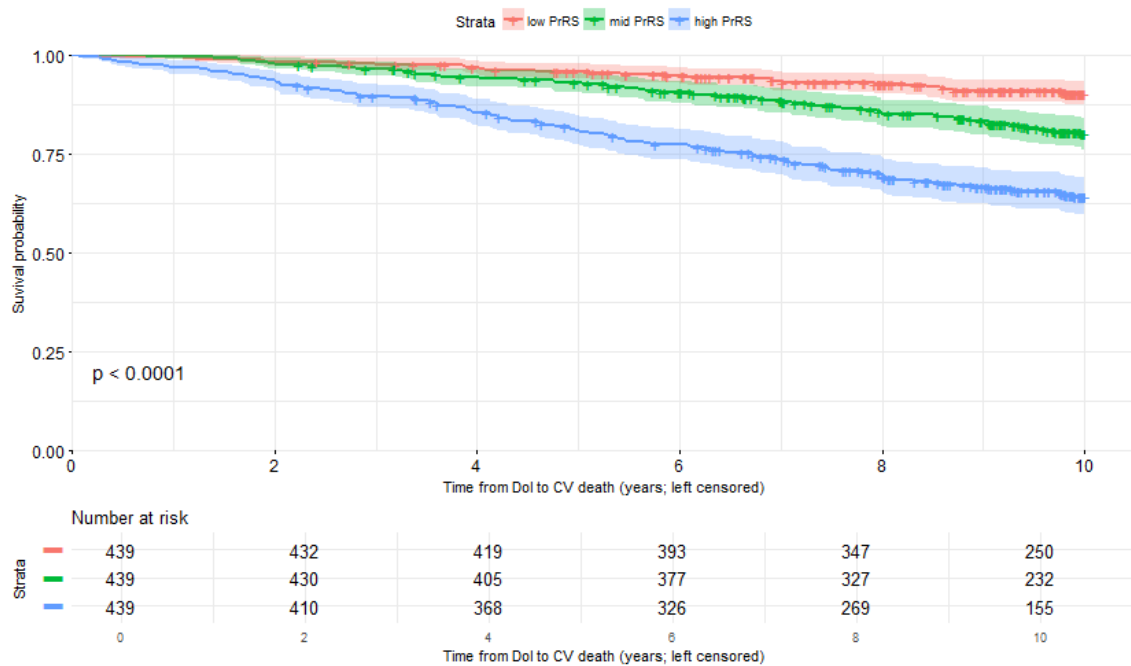
(e) Right eye, actual risk 0.41, predicted risk 0.37

(f) Right eye, actual risk 0.78, predicted risk 0.58

Fig. A.8 More examples of sample grad-CAM heatmaps for PCE risk prediction.

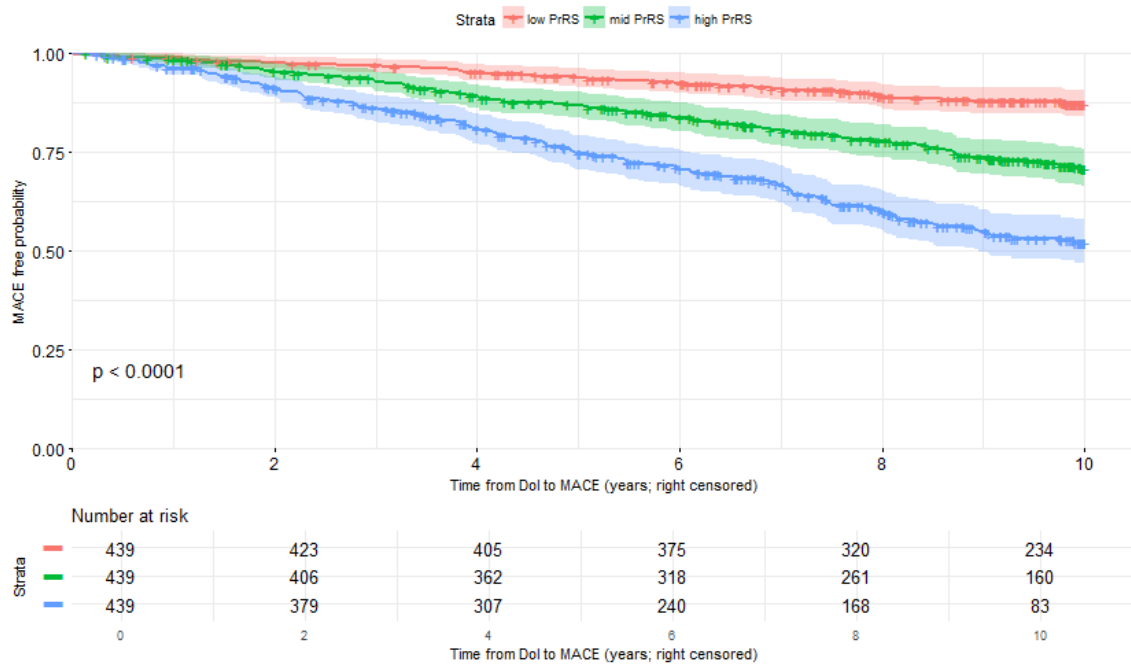


(a) KM curves for MACE

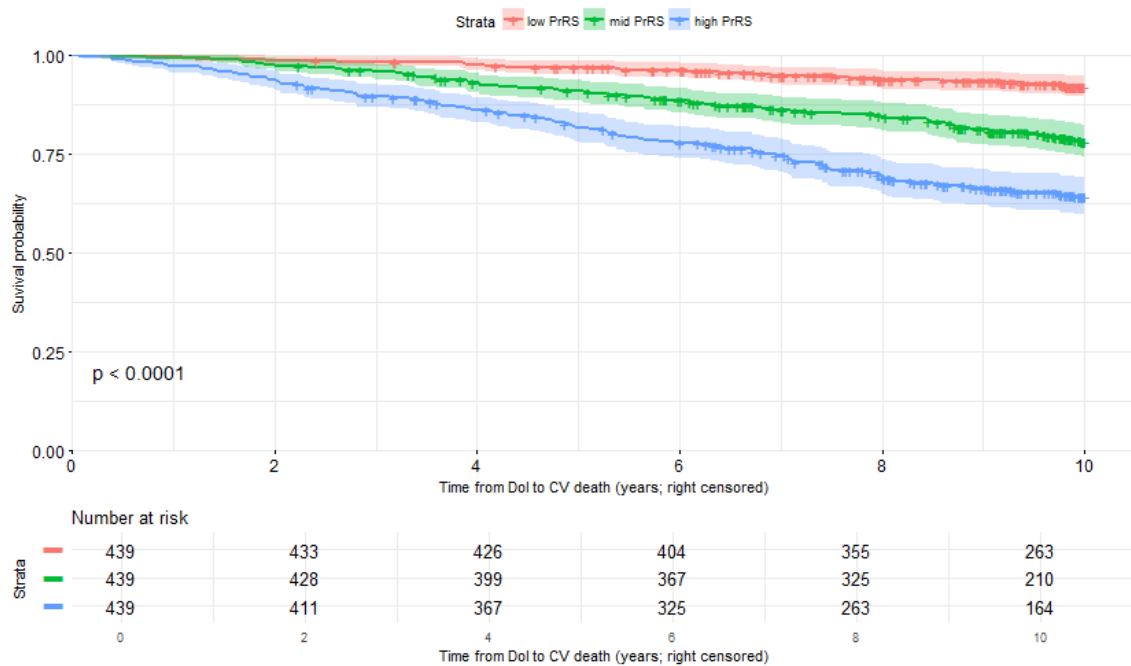


(b) KM curves for CV Death

Fig. A.9 KM curves of right censored survival data for upper (high PRS), middle and lower (low PRS) tertiles of retinal PRS groups derived from only left retina.

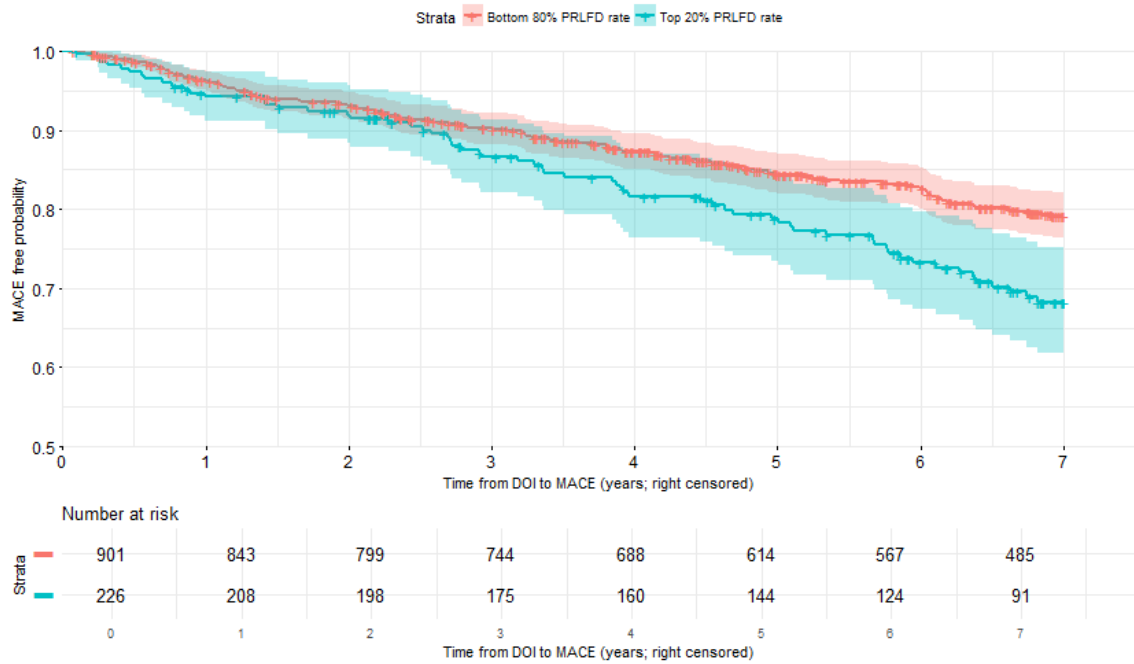


(a) KM curves for MACE

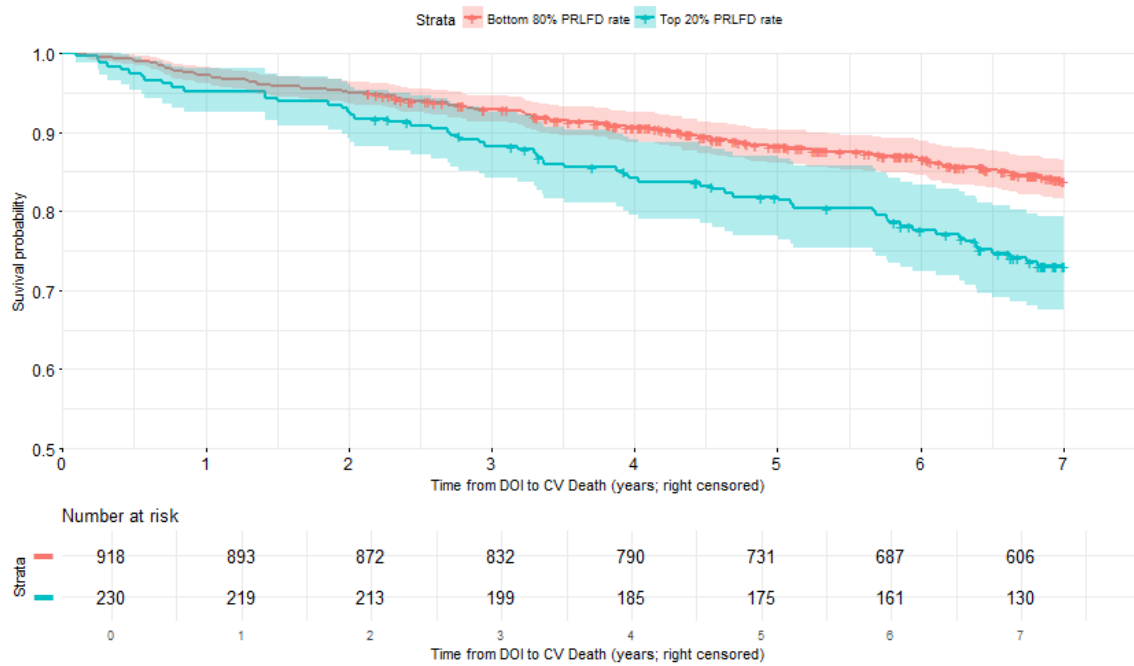


(b) KM curves for CV Death

Fig. A.10 KM curves of right-censored survival data for upper (high PRS), middle and lower (low PRS) tertiles of retinal PRS groups derived from only right retina.

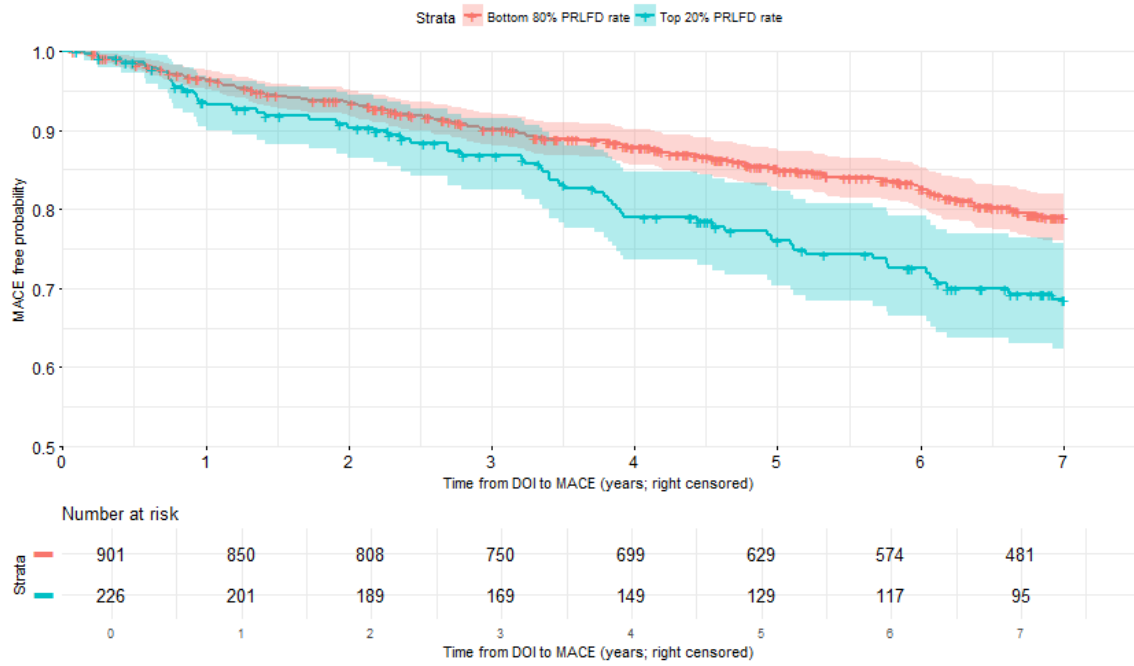


(a) KM curves for MACE

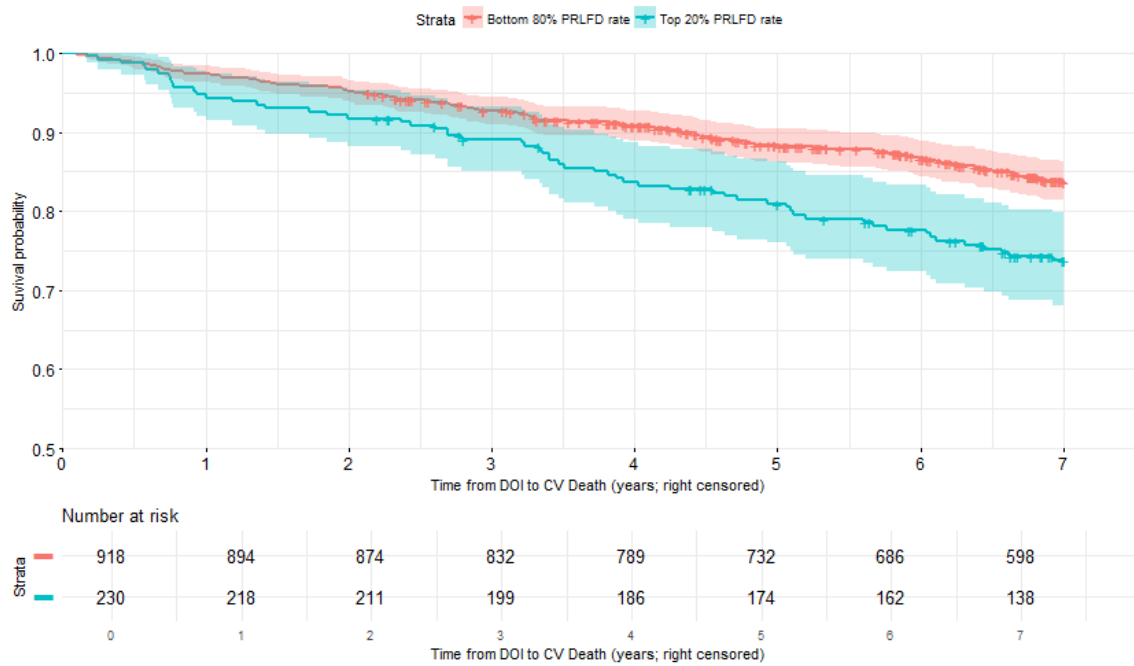


(b) KM curves for CV Death

Fig. A.11 KM curves of right censored survival data for top 20% and bottom 80% of the retinal Ω_{rate} groups computed from only left eye image predictions.



(a) KM curves for CV Death



(b) KM curves for CV Death

Fig. A.12 KM curves of right censored survival data for top 20% and bottom 80% of the retinal Ω_{rate} groups computed from only right eye image predictions.