

UCLA

UCLA Electronic Theses and Dissertations

Title

Analytical Methods for Diagnosis and Prediction of Health Conditions

Permalink

<https://escholarship.org/uc/item/7hp389gm>

Author

Davis, Tyler Austin

Publication Date

2022

Peer reviewed|Thesis/dissertation

UNIVERSITY OF CALIFORNIA

Los Angeles

Analytical Methods
for Diagnosis and Prediction
of Health Conditions

A dissertation submitted in partial satisfaction
of the requirements for the degree
Doctor of Philosophy in Computer Science

by

Tyler Austin Davis

2023

© Copyright by
Tyler Austin Davis
2023

ABSTRACT OF THE DISSERTATION

Analytical Methods
for Diagnosis and Prediction
of Health Conditions

by

Tyler Austin Davis
Doctor of Philosophy in Computer Science
University of California, Los Angeles, 2023
Professor Majid Sarrafzadeh, Chair

Recent years have seen a tremendous amount of growth in the performance and adoption of artificial intelligence (AI) and machine learning (ML) systems. These systems now permeate our lives, underpinning everything from web search to credit card fraud detection and photography. In principle, these advancements could also be applied to the domain of healthcare, where they could improve patient outcomes.

However, despite the almost fifty years that have elapsed since the first National Institutes of Health AI in Medicine (AIM) workshop in 1973 and the ubiquity of AI systems in our daily lives, AIM has not yet lived up to its lofty promises. AIM systems have seen limited deployment due to challenges including data missingness, data heterogeneity, explainability, and generalizability across variances in patient populations. The recent increase in the availability of electronic health record information, the variety and cost-effectiveness of mobile sensors, and the capabilities of machine learning algorithms promise to help improve healthcare delivery if challenges can be overcome. Through techniques such as interpretable

analysis of heterogeneous information networks and missingness-aware modeling, we demonstrate that the challenges of AI in Medicine can be overcome in order to improve healthcare access, aid physicians, and generate new insights into disease.

The dissertation of Tyler Austin Davis is approved.

Anthony John Nowatzki

Alex Anh-Tuan Bui

Yizhou Sun

Majid Sarrafzadeh, Committee Chair

University of California, Los Angeles

2023

TABLE OF CONTENTS

1	Introduction	1
1.1	Motivation	1
1.2	Research Objectives	3
1.3	Contributions	4
2	Hierarchical Target-Attentive Diagnosis Prediction (HTAD)	6
2.1	Introduction	7
2.2	Related Work	9
2.3	Preliminaries	11
2.3.1	EHR Network Formation Process	12
2.4	Methodology	13
2.4.1	Model Overview	13
2.4.2	Network Node Embedding	15
2.4.3	Target-attentive Node-Level Aggregation	16
2.4.4	Node-Level Time Series Aggregation	18
2.4.5	Type-Level Aggregation	18
2.4.6	Model Inference and Optimization	19
2.5	Experiments	21
2.5.1	Dataset	21
2.5.2	Baselines	22
2.5.3	Evaluation of Disease Phenotype Classification	24
2.5.4	Evaluation of Exact Diagnosis Code Prediction	26

2.5.5	Analysis of Attention Mechanism	28
2.6	Conclusion	29
3	Psychological Stress Detection in Older Adults With Cognitive Impairment Using Photoplethysmography	30
3.1	Introduction	31
3.2	Related Work	32
3.3	Methods	34
3.3.1	Physiological Recordings	34
3.3.2	Experimental Protocol	34
3.3.3	Feature Extraction	35
3.3.4	Classification Scenarios	36
3.4	Experimental Results	37
3.4.1	PPG Features Across Cohorts	38
3.4.2	Limitations	41
3.5	Conclusion	41
4	Predicting Rapid Kidney Function Decline Using EHR Data Despite High Missingness	44
4.1	Introduction	45
4.2	Related Work	46
4.2.1	Predicting CKD and Progression to ESKD	46
4.2.2	Predicting Rapid Decline	47
4.3	Methods	48
4.3.1	Population	48

4.3.2	Model Variables	50
4.3.3	Data Preparation	52
4.3.4	Predictive Modeling	53
4.3.5	Identification of Risk Factors	55
4.4	Experimental Results	57
4.4.1	Model Performance Results	57
4.4.2	Subgroup Analysis	59
4.4.3	Risk Factor Distributions	60
4.5	Discussion	60
4.5.1	Importance of UACR to the Model	62
4.6	Conclusion	65
5	RimNet: A Deep Neural Network Pipeline for Automated Identification of the Optic Disc Rim	66
5.1	Introduction	67
5.2	Methods	68
5.2.1	Dataset	68
5.2.2	RimNet Model and Hyperparameter Architecture	70
5.2.3	End-to-End mRDR Calculation Procedure	72
5.2.4	External Validation	75
5.2.5	Evaluation Criteria	75
5.3	Results	76
5.3.1	Hyperparameter Architecture	77
5.3.2	Segmentation Network Results	77

5.4	Discussion	81
5.5	Conclusion	85
6	DDLSNet: A Novel Deep Learning-Based System for Grading Funduscopy Images for Glaucomatous Damage	86
6.1	Introduction	87
6.2	Methods	88
6.2.1	Database	90
6.2.2	RimNet	91
6.2.3	DiscNet	92
6.2.4	DDLSNet Pipeline	95
6.2.5	Evaluation Criteria	95
6.3	Results	95
6.3.1	Model Architecture and Hyperparameter Search	97
6.3.2	RimNet	98
6.3.3	DiscNet	98
6.3.4	DDLSNet	98
6.4	Discussion	99
6.5	Conclusion	102
7	A Twin Convolutional Neural Network for the Identification of Glaucoma Progression Using Images of the Optic Nerve Head	103
7.1	Introduction	104
7.2	Methods	105
7.2.1	Dataset	106

7.2.2	Labeling Glaucoma Progressors vs Nonprogressors	107
7.2.3	Dataset Statistics	108
7.3	Image processing	108
7.4	Development of Twin Convolutional Neural Network	109
7.4.1	Segmentation Model	113
7.5	Model Performance and Statistical Analysis	113
7.6	Results	114
7.7	Discussion	118
7.7.1	Limitations	120
7.8	Conclusion	122
8	Conclusion	123
8.1	Future Work	124
	References	127

LIST OF FIGURES

2.1	(a) A visualization of how we map EHRs to an HIN, (b) EHR heterogeneous information network schema.	10
2.2	The architecture of the proposed hierarchical target-attentive HIN, illustrating the aggregation of patient p 's context nodes with respect to diagnosis d	15
2.3	Distribution of attention scores for prediction of kidney disease and diabetes in a patient presenting with both conditions	27
2.4	The distribution of attention weights among various record types	29
3.1	ROC curves and AUC	37
3.2	Distribution of the PPG features	43
4.1	STROBE Diagram: Overview of participant groups by CKD and at-risk CKD categories in the study	49
4.2	Illustration of the subgroup analysis process	57
4.3	Effect of varying threshold on precision and recall for GBTe model on test set	58
4.4	Distributions of features compared in patients with and without predicted decline (threshold = 0.5)	61
5.1	Distribution of mRDRs for Train, Validation, and Test Datasets. For each dataset, a frequency histogram is shown above with a box plot corresponding to the dataset below.	69
5.2	RimNet Pipeline, showing preprocessing, mask generation, and calculation of RDAR along with either mRDR or ARW depending on whether the rim is intact.	73

5.3	Segmentation Results. This figure demonstrates several examples of RimNet segmentation compared to physician segmentation. The left-most column shows the raw image. The middle column overlays the physician segmentation (white) over the raw image. The right-most column overlays the RimNet segmentation (white) over the raw image. In intact rims, green line shows the diameter and the dark blue shows the thinnest rim. In incomplete rims, the dark blue shows the edges of the segmentation.	78
5.4	Bland-Altman plots showing the agreements in mRDR and RDAR between clinician and RimNet in test images. Red dashed lines indicate 95% confidence limits.	79
6.1	The Disc Damage Likelihood Scale as originally proposed, figure by Spaeth et al. [1]	89
6.2	DDLSNet pipeline, illustrating both the RimNet and DiscNet arms. The calculated disc size and mRDR or ARW are used to calculate the DDLS score.	94
7.1	STROBE Diagram: Overview of eyes included in the study	106
7.2	Architecture of the twin neural network	111
7.3	ROC curve on the test set	114
7.4	Precision recall curve on the test set	115
7.5	Precision and recall as a function of threshold on the test set	115
7.6	Maximum variation in scores for a single eye with twin network	116
7.7	XRAI saliency maps for two eyes that were correctly classified as progressors by the deep learning model. Arrows indicate areas of rim loss.	117

LIST OF TABLES

2.1	Notation and Explanations	14
2.2	Phenotype Classification Results	25
2.3	Exact Diagnosis Code Ranking	26
3.1	Demographic of Study Population	35
3.2	F1-score of Stress Detection across Cognitively Impaired Aging and Control Groups	37
3.3	Statistically Significant Correlation with Cognitive Status	39
3.4	Statistically Significant Correlation with Gender	40
4.1	Continuous variables used in analysis	50
4.2	Categorical variables used in analysis	51
4.3	Hyperparameter search space for deep neural network (DNN)	53
4.4	Hyperparameter search space for logistic regression (LR)	53
4.5	Hyperparameter search space for gradient boosted trees (GBT)	54
4.6	Splits used for subgroup analysis	56
4.7	Model performance on test set	58
4.8	Most frequently occurring variables among top 100 highest risk subgroups . . .	59
4.9	Race/ethnicity for patients with predicted risk above versus study population .	60
4.10	Prevalence of $\geq 40\%$ eGFR decline in CURE-CKD registry by eGFR level and albuminuria	63
4.11	Incidence of $\geq 40\%$ eGFR decline in CURE-CKD registry by eGFR level and albuminuria	63
5.1	Demographic distributions for internal dataset	71

5.2	Glaucoma diagnosis for all 1 208 patients included in the RimNet dataset	76
5.3	Hyperparameter search space for RimNet	77
5.4	RimNet Results on internal test set and Drishti-GS dataset. The ARW cannot be calculated on the Drishti-GS dataset because all rims are intact.	79
5.5	DRISHTI-GS segmentation performance of RimNet compared to published segmentation models [2–7].	80
6.1	Glaucoma diagnoses for all 1 208 patients included in the RimNet dataset	90
6.2	Hyperparameter search space for RimNet	92
6.3	Hyperparameter search space for DiscNet	93
6.4	Demographic characteristics for the datasets used for RimNet, DiscNet, DDL- SNet, and DDLSNet reliability	96
6.5	The DDLS distribution for our test set of 120 images, graded by glaucoma specialists	96
6.6	Kappa agreement between DDLSNet and glaucoma specialist grading	98
6.7	Difference in DDLSNet grading between paired images of non-progressing optic disc photographs. All photographs were taken within four years of each other.	99
7.1	Demographic and clinical characteristics of the study population	108
7.2	Hyperparameter search space for twin neural network	112
7.3	Performance metrics for the investigated models on the test set	115

ACKNOWLEDGMENTS

I would like to thank my family for their unending support throughout my program, as they encouraged me to pursue my interests and push boundaries.

I would also like to acknowledge my collaborators and coauthors for their contributions to the works described in this dissertation. A version of Chapter 2 was published as “Hierarchical Target-Attentive Diagnosis Prediction in Heterogeneous Information Networks” [8]. Anahita Hosseini built the system that inspired this work [9], and she was responsible for implementing the attention mechanism and long-short term memory model in addition to her contributions to the paper. Majid Sarrafzadeh served as PI.

Chapter 3 was published as “Psychological Stress Detection in Older Adults with Cognitive Impairment Using Photoplethysmography” [10], and was a collaboration with Migyeong Gwak and Ellen Woo. Migyeong worked with Ellen Woo to design the study, and was responsible for developing the data platform and performing the initial data analysis. Ellen woo served as PI. I contributed to the construction of the model and final data analysis. All authors contributed to writing the paper.

Chapter 4 is a version of a manuscript in preparation for publication, based on a study done in collaboration with the CURE-CKD team. Versions of this work were presented at ASN Kidney Week 2020 [11] and ASN Kidney Week 2022 [12]. My collaborators include Panayiotis Petousis, Davina Zamanzadeh, Susanne Nicholas, Alex Bui, Keith Norris, Obidugwu Duru, and many others. Davina built the core of the data preprocessing pipeline. All coauthors suggested edits to the presented abstracts [11, 12]. Alex, Panayiotis, and Davina were a part of many great conversations seeking to extract meaning from my results. In these studies I was responsible for modeling and data analysis, while the research group helped immensely with data collection, as well as establishing clinical context. Alex Bui and Susanne Nicholas served as PIs.

A version of chapter 5 was published as “RimNet: A Deep Neural Network Pipeline

for Automated Identification of the Optic Disc Rim” [13], and a version of chapter 6 was published as “DDLSNet: A Novel Deep Learning-Based System for Grading Fundusoscopic Images for Glaucomatous Damage” [14]. These studies were done in collaboration with a group based out of the glaucoma division at Stein Eye Institute. This group includes Haroon Rasheed, Esteban Morales, Zhe Fei, Lourdes Grassi, Agustina De Gainza, Kouros Nouri-Mahdavi, and Joseph Caprioli. In these studies I was responsible for model building and evaluation, building upon initial exploratory work by Haroon and Esteban. Esteban Morales built the computer vision pipeline and generated the datasets used for the project. Haroon conducted extensive literature reviews and distilled the more than two years of work in these projects down into the initial drafts of the manuscripts. Zhe Fei provided statistical expertise and contributed to editing. Lourdes Grassi, Joseph Caprioli, Kouros Nouri-Mahdavi, and Agustina De Gainza provided invaluable clinical expertise, ensuring our work was clinically relevant and impactful, and helped to create the high quality datasets without which our systems could not have been built. Joseph Caprioli and Kouros Nouri-Mahdavi were the PIs.

Chapter 7 is a version of a manuscript in preparation for publication. This study was done in collaboration with a group based out of the glaucoma division at Stein Eye Institute. Vahid Mohammadzadeh and I designed the study. I developed the model and performed all model evaluation. Evan Maltz, Alex Broering, and Alon Oyler-Yaniv helped develop the model. Diana Salazar Vega, Golnoush Mahmoudi Nezhad, Jack Martinyan, Sepideh Heydarzadeh, and Esteban Morales were responsible for data management. Fabien Scalzo supervised the model’s development. Kouros Nouri-Mahdavi and Joseph Caprioli were the PIs.

Research reported in this dissertation was supported in part by the Department of Health and Human Services of the National Institutes of Health under award number T32EB166406. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health. Additionally, this work was funded by

UCLA CTSI grant (UL1 TR001881), NIH R01 MD014712, and NIH T32 EB016640. The work described in Chapter 7 was supported by an NIH R01 grant (R01-EY029792) (KNM), and an unrestricted Departmental Grant from Research to Prevent Blindness (KNM).

VITA

- 2013–2017 B.S. in Electrical Engineering and Computer Science, Bioengineering, University of California, Berkeley, Berkeley, California.
- 2019–2022 Teaching Assistant/ Associate/ Fellow, Computer Science Department, UCLA.
- 2021 Teaching Assistant, Computer Science Department, Harvard University.
- 2021 M.S. (Computer Science), UCLA, Los Angeles, California.
- 2017–2022 PhD student in Computer Science, eHealth Research Lab, UCLA.

PUBLICATIONS

Rasheed, H., Davis, T. A., Morales, E., Fei, Z., Grassi, L., De Gainza, A., & Caprioli, J. (2022). DDLSNet: A Novel Deep Learning-Based System for Grading Funduscopy Images for Glaucomatous Damage. doi:10.1016/j.xops.2022.100255

Rasheed, H. A., Davis, T., Morales, E., Fei, Z., Grassi, L., De Gainza, A., ... Caprioli, J. (2022). RimNet: A Deep Neural Network Pipeline for Automated Identification of the Optic Disc Rim. *Ophthalmology Science*. doi:10.1016/j.xops.2022.100244

Davis, T. A., Petousis, P., Zamanzadeh, D. J., Norris, K. C., Duru, O., Tuttle, K., ... & CURE-CKD Registry Study Team. (2022). PO937: Predicting Rapid eGFR Decline in the CURE-CKD Registry. Poster at American Society of Nephrology Kidney Week 2022.

Zamanzadeh, D. J., Petousis, P., Davis, T. A., Nicholas, S. B., Norris, K. C., Tuttle, K. R., ... & Sarrafzadeh, M. (2021, November). Autopopulus: A Novel Framework for Autoencoder Imputation on Large Clinical Datasets. In 2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC) (pp. 2303-2309). IEEE.

Gwak, M., Davis, T., Sarrafzadeh, M., & Woo, E. (2021, August). Psychological Stress Detection in Older Adults with Cognitive Impairment Using Photoplethysmography. In 2021 IEEE 9th International Conference on Healthcare Informatics (ICHI) (pp. 209-213). IEEE.

Davis, T. A., Petousis, P., Zamanzadeh, D. J., Wang, X., Norris, K. C., Duru, O., ... & CURE-CKD Registry Study Team. (2020). PO0528: Predicting Rapid eGFR Decline Using Electronic Health Record (EHR) Data Despite High Missingness in the CURE-CKD Registry. Poster at American Society of Nephrology Kidney Week 2020.

Nicholas, S. B., Follett, R. W., Tacorda, T. T., Wang, X., Ruenger, D., Petousis, P., ... & Bui, A. (2021). Disparities in CKD risks: Data from the cure-CKD COVID-19 sub-registry. *Journal of the American Society of Nephrology*, 84-84.

Hosseini, A., Davis, T., & Sarrafzadeh, M. (2019, November). Hierarchical target-attentive diagnosis prediction in heterogeneous information networks. In 2019 International Conference on Data Mining Workshops (ICDMW) (pp. 949-957). IEEE.

CHAPTER 1

Introduction

1.1 Motivation

Not long after the term “artificial intelligence” (AI) was first coined, researchers began to dream of how computer systems may one day be used to automate and improve patient care [15, 16]. However, the path towards this goal was hampered by challenges such as the difficulty of encoding knowledge into a system, which resulted in narrowly scoped systems [17], limited performance [18, 19], and “AI winters” where research dramatically slowed for decades at a time [15, 16]. The early 2010s saw immense breakthroughs in AI and an associated boom in research [20] that led to the widespread adoption of AI systems for everything from self-driving cars [21], natural language understanding [22], and even photography [23]. However, despite this widespread adoption in other domains, AI has seen comparatively limited adoption in the field of medicine. This has been due to factors such as outsized expectations in conjunction with limited trust between clinicians and tools [24, 25] and the difficulty in extracting meaning from noisy, biased, and multidimensional data [16].

Meanwhile, mounting challenges to healthcare delivery in the United States are incentivizing researchers and healthcare systems to look for new ways to lower costs while ensuring quality care. Patient care costs are currently increasing at a remarkable rate, up 130% since 2000 to \$11,582 in 2019, outpacing inflation [26] and resulting in a total of 3.8 trillion dollars in expenditures [27, 28]. This expenditure is equivalent to 17.7 percent of the 2019 US GDP. Unfortunately, these costs are only expected to rise as the US population continues to

age [29], with an expected average growth rate of 5.4% per year. In addition, there is predicted to be a shortage of 40,800 to 109,000 physicians by 2030 [30], meaning that resources may be stretched thin, potentially threatening the quality of care that patients receive. Some recent reports in medical journals argue that the impending shortage can be addressed in ways other than simply training more physicians, such as by using new technologies to enable existing physicians to deliver care to more patients and to allow for better delivery of care to underserved populations [31, 32].

Healthcare delivery, and the technologies supporting it, have undergone a remarkable amount of transformation within the last fifteen years, the effects of which are still only beginning to be felt. In 2008, only 9% of non-federal acute care hospitals in the United States were maintaining patient records in a basic electronic health record (EHR) system [33]. Today things are a bit different, with over 90% of non-federal acute care hospitals electronically storing patient information in a way that it can be exchanged with other care providers [34]. It is estimated that a single hospital stay now generates approximately 150,000 discrete pieces of data [35]. Additionally, the average healthcare system manages more than 8.4 petabytes of data as of 2018, an almost ninefold increase from 2016 [35]. This rapid digitization of healthcare records means that not only is it easier for patients to have their information follow them as they move between physicians, but that records are now richer and easier to search through, lowering the barrier to running large scale retrospective studies.

At the same time as this rapid digitization of health records, advancements in wireless systems, low cost sensors, and personal computing began enabling continuous gathering of information, such as blood glucose or respiratory rate, that would have previously been prohibitively expensive. Add on top of this recent explosion in information-rich and easily accessible healthcare data a renewed interest in AI research, and the stage is set for the application of new techniques to datasets the likes of which have never been seen before. The recency of these developments in the artificial intelligence and healthcare communities means that there is an enormous amount of potential to explore the possibilities for easing

modern challenges in healthcare delivery and epidemiological research.

1.2 Research Objectives

The goal of this work is to develop new analytical techniques that will incorporate new sensors, new information rich heterogeneous data sources, and extensive but imperfect health records in order to overcome long-standing hurdles to the adoption of artificial intelligence in medicine including:

1. **Heterogeneity:** A single patient’s records may contain time series sensor readings, tabular data such as biographical information, text based notes, and medical scans saved as images.
2. **Missingness:** Each patient’s health record is composed of a unique combination of clinic visits and tests, so there is no guarantee that two patients will have the same types of data their record, or that they will have similarly dense records, often leading to high levels of missingness.
3. **Generalizability:** There is a great amount of variation between patients due to biological differences from factors such as age, sex, and disease. These differences make it difficult to generalize algorithms and insights to the diverse population at large.
4. **Interpretability:** Many modern analytical techniques are “black box” systems where it is not possible to peer inside the box and determine why the system arrived at its final conclusion. This is particularly an issue in medicine as clinicians want to verify a system is arriving at a decision for the right reasons, understand why a system erred, or even update their own beliefs using the model’s insights.

Each of the works described in this dissertation deals with a different set of challenges, and addressing these challenges is a core component of the research. By overcoming these challenges, future AIM systems may be able to:

1. **Aid physicians:** Advances in machine intelligence may allow for time-consuming and laborious tasks to be partially automated, or for helpful suggestions to be provided as guides for diagnosis and analysis.
2. **Gain new insights into disease:** The increased prevalence of EHR systems means that not only is it easier than ever before to find retrospective cohorts of patients, but also that these records are richer than ever before. Careful analysis of this information may lead to new insights into less common disease patterns.
3. **Increase healthcare access:** New sensors allow for patient monitoring to be done in ways not previously possible, allowing for high quality monitoring without the monetary or time cost of a clinic visit.

1.3 Contributions

This dissertation describes works which all work towards the research objectives listed above while addressing unique domain-specific challenges. In these works:

- We introduce HTAD, a novel model for diagnosis prediction using electronic health records represented as heterogeneous information networks. Our model introduces a target-aware hierarchical attention mechanism that allows it to learn to attend to the most important clinical records when aggregating their representations for prediction of a diagnosis.
- We describe a physiological signal data collection system based on a portable device and a smartphone. Through a classifier trained on this data and our own analysis, we show the different impacts of psychological stress in healthy and cognitively impaired older adults as well as in males and females. Our proposed system can be used as a continuous stress monitoring system in real-world settings that is non-invasive, portable, and easy to use.

- We describe a deep neural network for predicting the risk of rapid kidney function decline and identify populations at higher risk of rapid decline using the CURE-CKD Registry. Our model achieves strong performance despite high levels of data missingness in the registry.
- We describe RimNet, a fully automated system for accurate segmentation of the optic disc rim, and the first study to report performance for this task on incomplete rims.
- We describe DDLSNet, the first system for automated estimation of DDLS, enabling faster evaluation with less variability. Additionally, this is the first study to report on the problem of determining optic disc size solely using optic disc photos without any external aids.
- We describe a system that evaluates pairs of images for the same markers of progression that glaucoma specialists look for. Combined with image saliency techniques, such a system is a promising add-on for clinical decision-making.

CHAPTER 2

Hierarchical Target-Attentive Diagnosis Prediction (HTAD)

In this chapter we introduce HTAD, a novel model for diagnosis prediction using Electronic Health Records (EHR) represented as heterogeneous information networks. Recent studies on modeling EHR have shown success in automatically learning representations of the clinical records in order to avoid the need for manual feature selection. However, these representations are often learned and aggregated without specificity for the different possible targets being predicted. Our model introduces a target-aware hierarchical attention mechanism that allows it to learn to attend to the most important clinical records when aggregating their representations for prediction of a diagnosis. Additionally, our model is built to handle the heterogeneity of data types in EHR, as it can accept many forms of input data simultaneously, including both time series and tabular data.

We evaluate our model using a publicly available benchmark dataset and demonstrate that the use of target-aware attention significantly improves performance compared to the current state of the art. Additionally, we propose a method for incorporating non-categorical data into our predictions and demonstrate that this technique leads to further performance improvements. Lastly, we demonstrate that the predictions made by our proposed model are easily interpretable.

2.1 Introduction

Electronic Health Records (EHR) provide a comprehensive picture of patients’ medical histories, consisting of information such as written clinician notes, medical imagery, prescriptions, and diagnoses. With the recent availability of EHR datasets to researchers, there has been a significant amount of interest in using this information to improve patient outcomes. In this study, we focus on the problem of predicting patients’ diagnoses based on their health records.

Some of the challenges in mining health data are its high heterogeneity and its sparse record distribution, which have led many studies to rely on expert knowledge and manual selection of a set of dense features [36, 37]. One way in which these challenges have been approached is through an unsupervised record embedding technique, first proposed by Med2Vec [38]. Med2Vec, as well as successive studies such as [39], use a skip-gram [40] based technique to learn latent representations for health records based on their co-occurrence relations. In this approach, predictions are commonly made by training supervised models on patient representations, which are obtained by aggregating the embeddings of the items in a patient’s health records. Another work using a similar approach is HeteroMed [9], which demonstrates the advantages of modeling EHR data using Heterogeneous Information Networks (HIN). HeteroMed shows that HINs can capture the structure and semantically important relations of EHR and model its heterogeneity. In this study we continue to explore the promise of HINs for modeling EHR, addressing the shortcomings of prior record embedding approaches along the way.

One shortcoming in these past works stems from the relatively simple aggregation process they use, in which they treat records with equal importance regardless of what diagnosis is being predicted. Taking diabetes and kidney failure as an example, we can see how this is an issue: prior models generate a single patient representation by combining records with fixed weights, which is then used for the prediction of both diagnoses; however, the importance

of tests should vary based on the diagnosis being predicted, with blood glucose levels being more important than blood albumin levels when predicting diabetes than when predicting kidney failure and vice versa. Another shortcoming of these past approaches is that the predictions generated by these models are not easily interpretable, with no way for an end user to understand how the model arrived at its conclusion. Lastly, past approaches only make use of records whose values can be mapped to distinct categories, leaving out other important information such as time series vital signs and medical imagery.

Inspired by the very recent success of attention mechanisms in network embedding [41,42], we propose HTAD, a novel approach for modeling EHR data that leverages hierarchical attention, to overcome these shortcomings. HTAD produces diagnosis-aware patient representations, as well as explainable predictions. We also suggest how non-categorical data, in particular, time series data, can be integrated into HTAD.

Considering EHR in the context of HIN with patients and records mapped to network nodes, our model’s goal is to aggregate a patient’s neighborhood such that the obtained representation is tailored to the prediction of a specific target diagnosis. Recognizing heterogeneity of nodes, we perform the neighborhood aggregation at two levels: first, at node-level and among nodes having similar type to obtain a set of type representations, and then at the type-level to achieve a comprehensive patient representation. In node-level aggregation, we propose employing a target-aware attention mechanism to learn the importance of various nodes with respect to the given diagnosis. We also show ways for the incorporation of time-series data at this level. We apply similar attention technique at the type-level to allow the model to learn preference towards various record types for the prediction of the specified disease. We then pass the resulting patient representation into our objective function for prediction. Importantly, attention weights generated in our model improve the interpretability by providing insight as to which nodes and types the model finds most important for the prediction.

We evaluate our proposed model’s performance on two diagnosis prediction tasks: ex-

act diagnosis code prediction and high-level diagnosis group prediction, using the publicly available MIMIC-III EHR dataset [43]. We compare HTAD to several existing models that represent the state of the art for diagnosis prediction using EHRs. Our experiments show that HTAD outperforms these benchmarked models on both tasks, in multiple cases beating them by a margin of over 10%.

Additionally, we evaluate our model’s interpretability, something that has not been explored in past models for diagnosis prediction that represented patients based on their aggregated EHR embeddings. In summary, we make the following contributions in this paper:

1. We propose Hierarchical Target Attentive Diagnosis (HTAD) in an HIN setting and demonstrate that it significantly improves diagnosis prediction performance.
2. We demonstrate that HTAD’s use of target-aware hierarchical attention can improve interpretability.
3. We demonstrate that non-categorical data can be incorporated when mining EHR data represented as an HIN.

2.2 Related Work

In this section, we highlight prior representative works in three areas that come together in this study: EHR data mining, Heterogeneous Information Network embedding, and attention-based modeling.

2.2.0.1 EHR Modeling

When modeling EHR, there are two main challenges that prior studies have approached. First, clinical records are heterogeneous and are sparsely distributed among patients. To tackle this, manual feature selection has been a method of choice in many studies, leading to two recent works on benchmarking a public EHR dataset [36, 37] and introducing

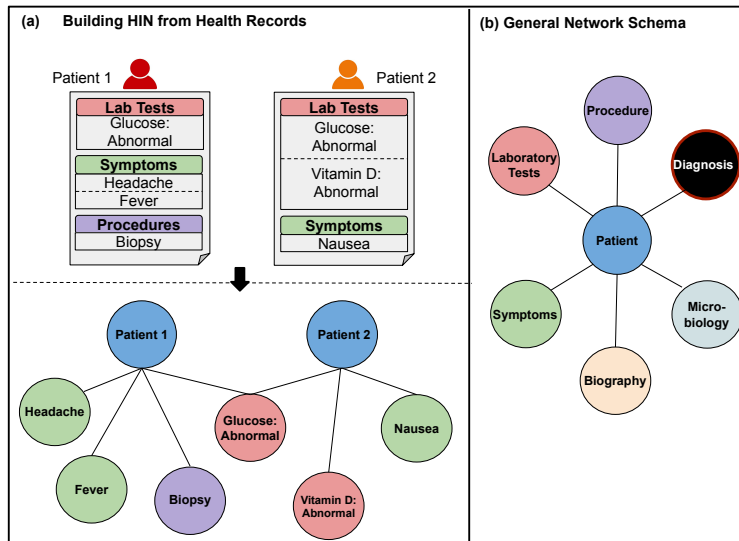


Figure 2.1: (a) A visualization of how we map EHRs to an HIN, (b) EHR heterogeneous information network schema.

a set of features to be extracted for various tasks [36]. In another direction, studies such as Med2Vec [38] introduced the unsupervised embedding of clinical records using a skip-gram which was adopted by a number of later studies [39, 44, 45] and was extended by HeteroMed [9].

Second, it can be difficult to model the complex structure and relations in EHRs. Recurrent Neural Networks (RNNs) have been one of the most widely adopted techniques. However, RNNs lose efficiency and performance when working on long sequences, and clinical records may contain thousands of items. Moreover, they fail to capture the structure and semantics of relations in EHR. HeteroMed [9] proposes the use of HINs for the analysis of EHRs, allowing to capture both node and relation semantics. Our work is inspired by the success of HeteroMed in representing EHRs as an HIN and works to overcome prior studies shortcomings in disregarding the importance of records and providing integrative modeling.

2.2.0.2 Heterogeneous Information Network Embedding

Heterogeneous Information Networks (HIN) have recently gained considerable attention, especially in the domain of recommendation systems. These networks are able to capture various types of entities and relation semantics, which is essential in modeling real-world settings. Embedding an information network refers to learning compact representation vectors for its nodes. Many homogeneous network embedding approaches, such as DeepWalk [46] and node2vec [47], employ random walks or neighbor prediction mechanisms, paired with skip-gram based models. For HINs, relation-based walks have been introduced to incorporate the heterogeneity of data [48].

2.2.0.3 Attention Mechanisms

Attention mechanisms for learning algorithms have gained huge success in the domains of natural language processing [49], with the goal of allowing a model to attend to the most important parts of text while ignoring less relevant portions. Attention for network analysis is a growing topic of interest, with recent studies [41, 42] employing it in the selection of important neighbor nodes, random walks, and meta paths, respectively. In this study, we explore attention in HINs for target-aware node importance scoring when modeling EHR.

2.3 Preliminaries

Definition 1. *Heterogeneous Information Networks* [50] A *Heterogeneous Information Network (HIN)* is defined as a graph $G = (V, E)$ with two type functions $h : V \mapsto A$ and $g : E \mapsto R$ that map nodes and edges to their predefined types A and R , respectively.

Definition 2. *Meta Path* [50] Given A and R , representing sets of all node and edge types in graph G , a meta path is defined by a schema in the form of $A_1 \xrightarrow{R_1} A_2 \xrightarrow{R_2} \dots \xrightarrow{R_m} A_{m+1}$. Any two nodes with a connecting path matching this schema will be linked through this meta

path.

2.3.1 EHR Network Formation Process

In general, an EHR can be viewed as a set of patients $P = \{p_1, p_2, \dots, p_{|P|}\}$ and clinical records $C = \{c_1, c_2, \dots, c_{|C|}\}$. We first put forward a formal view of clinical records.

Definition 3. *Clinical Record* *A clinical record is defined as a triple: $c = (i, t, v)$, where i , t , and v respectively denote the ID of the recorded item (e.g., blood glucose level), its type (e.g., laboratory test), and its value which can be null for some record types, such as symptoms.*

To model EHR as an HIN we rely on a function mapping clinical records to nodes, defined as: $f_c: C \mapsto V$, which projects $c = \{i, t, v\} \in C$ to a node $v \in V$ identified by the tuple (i, v) and having type t . Similarly, $f_p: P \mapsto V$ maps each patient to a node with the same type and identified by the patient ID. Furthermore, the basic links of the network are formed between patient nodes and the nodes representing their clinical records. Fig. 2.1 illustrates this process. To interpret the clinical record values in an EHR, we follow the strategies introduced in [9], which attempt to categorize all node values. However, unlike their approach, we do not discard information that remains in a non-categorical format, and we later present a way for incorporating this data into our model.

Definition 4. *Target/Context Nodes* *Target nodes are defined as the nodes for which the presence of the link to a patient should be predicted (diagnosis nodes in this study). All nodes other than patient and target are considered as context nodes.*

Given these preliminaries, the diagnosis prediction task in an HIN representing EHR data can be defined as:

Definition 5. *Clinical Prediction in an HIN Setting* *Given a patient p with context nodes $N(p) = \{N_1(p), N_2(p), \dots, N_T(p)\}$ where $N_t(p)$ denotes the type t neighborhood of p , predict p 's target neighborhood: $N_d(p) = \{d_1, d_2, \dots, d_{|N_d(p)|}\}$, where d_i is the i th target node.*

When working with diagnosis prediction task, it is important to note that many medical ontologies, such as the ICD-9 system [51], provide a hierarchical and multi-resolution view of diagnoses, with the highest level of the hierarchy identifying the general disease group (e.g., cardiovascular disorders) and lower levels providing more specificity as to the exact diagnosis. Importantly, clinicians may assign codes to a patient at any level. Therefore, the diagnosis prediction task can be defined at two levels:

- Low-level (exact) code prediction: Due to the large number of diagnosis codes, this task is approached as a ranking problem, with the aim of scoring positively labeled codes higher than others.
- High-level (grouped) code prediction: In this task, we aim to predict all diagnosis groups associated with a patient, formulated as a multi-label classification task.

2.4 Methodology

In this section, we present our proposed HIN-based EHR model, leveraging a hierarchical target-attentive architecture.

2.4.1 Model Overview

To model health records and patients, we rely on learning embedding vectors for all these entities. In this approach, a patient representation is often obtained by an aggregation of the embeddings of his/her clinical records and is used for the target prediction task. Different from prior studies where a single patient representation was generated, our model learns to obtain a distinct patient representation for each target node, achieved by favoring the most predictive records for that specific target. The overall architecture for our target-attentive patient aggregation is depicted in Fig. 2.2.

Describing the process in HIN setting, we first aggregate context nodes based on their type

Table 2.1: Notation and Explanations

Symbol	Explanation
h_n	Embedding of node n
h'_n	Transformed embedding of node n
$N_t(p)$	Type t neighborhood of patient p
$z_{p,d}^t$	Aggregated embedding of nodes in $N_t(p)$ with respect to diagnosis d
q^d	Node-level attention vector for diagnosis d
s^d	Type-level attention vector for diagnosis d
$\alpha_{n,d}^t$	Node-level attention score assigned to node $n \in N_t(p)$ when predicting for diagnosis d
$\beta_{p,d}^t$	Type-level attention score assigned to type t representation of patient p , when predicting for diagnosis d
$f_{p,d}$	Aggregated patient p embedding with respect to diagnosis d
M	Node embedding lookup matrix
Q	Node-level attention lookup matrix
S	Type-level attention lookup matrix
W_c^t, b_c^t	Transformation parameters for context nodes with type t
W_d, b_d	Transformation parameters for target (diagnosis) nodes
W_q, b_q	Transformation parameters to obtain node-level attention
W_s, b_s	Transformation parameters to obtain type-level attention
W_t, b_t	Transformation parameters for time series type embedding

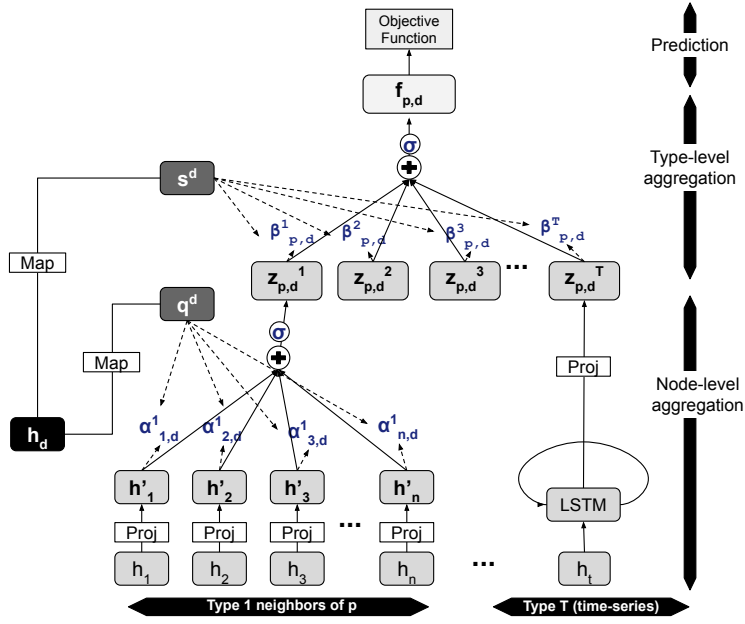


Figure 2.2: The architecture of the proposed hierarchical target-attentive HIN, illustrating the aggregation of patient p 's context nodes with respect to diagnosis d

using a node-level attention mechanism, generating type-specific embedding vectors. The attention weights are assigned based on the importance of the node in the prediction of the diagnosis. We also present a type-level attention layer to learn the importance of each type in predicting the target, further helping to obtain a diagnosis-aware patient representation. Finally, to generate the aggregated type embedding for time-series nodes as well, we replace the node-level attention mechanism with a deep sequential model.

In addition to learning node embeddings using the supervised model described above, we use an unsupervised approach for learning embeddings in order to capture the structure and semantically important relations in EHRs.

2.4.2 Network Node Embedding

Having N as the set of all network nodes, the embedding of $n \in N$ is denoted as h_n and is obtained by looking up the corresponding vector from a trainable embedding matrix

$M \in \mathbb{R}^{|N| \times F}$, where F is the length of the embedding vector.

2.4.3 Target-attentive Node-Level Aggregation

As EHRs are composed of data of heterogeneous types, each node type can carry specific semantic and diagnostic information. Therefore, we start the aggregation process of a patient’s neighborhood by combining the context nodes based on their types, thus obtaining type representation vectors. With this in mind, given a patient p , its type t neighborhood, $N_t(p)$, and a diagnosis node d with corresponding embedding vector h_d , the node level target-attention works as follows:

We first utilize a linear transformation layer, parameterized by a type-specific weight matrix $W_c^t \in \mathbb{R}^{F' \times F}$ and bias vector $b_c^t \in \mathbb{F}'$, to project p ’s context nodes into a new feature space that is more expressive for attention-based node scoring:

$$h'_n = W_c^t h_n + b_c^t \tag{2.1}$$

where h_n and h'_n , having length F and F' , denote the original and transformed embeddings of context node $n \in N_t(P)$.

The importance of each node is then measured based on the similarity of its transformed embedding to a diagnosis-specific attention vector $q^d \in F'$. In the most general design, this vector is obtained by applying a linear transformation, parameterized by weight $W_q \in \mathbb{R}^{F' \times F}$ and bias vector $b_q \in \mathbb{F}'$, to the diagnosis node embedding h_d , formulated as:

$$q^d = W_q h_d + b_q \tag{2.2}$$

where h_d is the original diagnosis node embedding.

However, when working with low-level diagnosis codes, there is a significant imbalance in their frequency in a real-world setting. Therefore, the prior approach may face trouble in learning attention vectors for sparser codes. As such, grouping together those with similar diagnostic processes and allowing them to share attention vectors can improve the expressive

power of attention for sparser codes.

Following this idea and taking D' as the set of such a grouping with size $|D'|$, q^d can be looked up from an attention matrix $Q \in \mathbb{R}^{|D'| \times F'}$, after mapping d to one of the $|D'|$ diagnosis groups. Q is randomly initialized and jointly trained by the model. It is important to note that for high-level diagnosis classification task these groups can be defined the same as diagnosis groups we are predicting for. We refer to this approach for the rest of this paper as **group-based** attention.

Having the transformed node embedding h'_n and diagnosis attention vector q^d obtained, the importance score between them denoted as $e_{n,d}^t$, is calculated as:

$$e_{n,d}^t = \frac{q^d \cdot h'_n}{\sqrt{F'}} \quad (2.3)$$

where t shows the type of node n and division by $\sqrt{F'}$ is used to scale the score for improved performance, following [52].

We then normalize the node importance scores using a softmax function to obtain the attention coefficient $\alpha_{n,d}^t$.

$$\alpha_{n,d}^t = \frac{\exp(e_{n,d}^t)}{\sum_{n' \in N_t(p)} \exp(e_{n',d}^t)} \quad (2.4)$$

Lastly, the normalized attention coefficients are used as weights for linear aggregation of transformed node embeddings, which is then followed by a non-linearity function to form the type embedding:

$$z_{p,d}^t = \sigma \left(\sum_{n \in N_t(p)} \alpha_{n,d}^t \cdot h'_n \right) \quad (2.5)$$

where $z_{p,d}^t$ denotes the representation of type t neighbors of p when predicting for diagnosis d .

2.4.4 Node-Level Time Series Aggregation

As discussed in section 2.4.2, the node embeddings used in the node-level aggregation process are obtained using a shallow embedding lookup process. However, such a technique is not usable for records kept in a time series format, as these records cannot be easily mapped to a small fixed set of categorical values and as there would be too little sharing of nodes between patients if each unique time series value were mapped to a node. Therefore, to incorporate such records into our proposed information network, we employ a Long-Short Term Memory (LSTM) [53] sequential model similar to [36]. In particular, patient p 's time series records $S_t(p) = \{s_1, s_2, s_3, \dots, s_T\}$ is first fed to the LSTM model and then the hidden state of the last LSTM cell, denoted as v_t , is transformed to a vector with embedding size F' , forming the type t representation:

$$z_{p,d}^t = W_t v_t + b_t \quad (2.6)$$

It is worth noting that the embedding obtained is not diagnosis specific, but we have included d to keep the type representation notation consistent throughout the paper.

2.4.5 Type-Level Aggregation

After deriving type representations, $Z_{p,d} = \{z_{p,d}^1, z_{p,d}^2, \dots, z_{p,d}^T\}$, our next step is to combine them to generate the patient representation. Similar to nodes, the predictive power of the different types may vary across diagnoses. For example, the diagnosis of some diseases relies more upon the laboratory tests while others on symptoms.

Therefore, we propose to use another layer of diagnosis-aware aggregation. Similar to node-level aggregation, a type-level attention vector is employed that can either be obtained by a linear transformation of the original diagnosis embedding, parameterized by weight W^s and bias b^s , or be looked up from the attention-matrix $S \in \mathbb{R}^{|D'| \times F'}$.

The normalized attention coefficient between the type t representation ($z_{p,d}^t$) and attention vector s^d is defined as:

$$\beta_{p,d}^t = \frac{\exp \frac{s^d \cdot z_{p,d}^t}{\sqrt{F'}}}{\sum_{z'_{p,d} \in Z_{p,d}} \exp \frac{s^d \cdot z'_{p,d}}{\sqrt{F'}}} \quad (2.7)$$

In the final step, the comprehensive patient representation, specific to prediction of diagnosis d , is denoted as $f_{p,d}$ and is obtained by combining the type representations as follows:

$$f_{p,d} = \sigma \left(\sum_{t \in T} \beta_{p,d}^t \cdot z_{p,d}^t \right) \quad (2.8)$$

2.4.6 Model Inference and Optimization

In section 2.4.5, we explained how we obtain a set of patient representations $F_p = \{f_{p,d_1}, f_{p,d_2}, \dots, f_{p,d_k}\}$, in order to predict each of the k diagnoses in $D = \{d_1, d_2, \dots, d_k\}$. In this section, we describe the optimization and inference of the two prediction tasks built on top of these representations.

2.4.6.1 High-level Diagnosis Code Classification

As this task is formulated as a multi-label classification problem, we first feed the representations into a Multi Layer Perceptron (MLP) that maps $F_p \mapsto D$ and is implemented in two layers: the first one shared among all patient representations and the second one specific to each diagnosis group. We then optimize the model by the following loss function:

$$\begin{aligned} L &= \text{mean}(l_1, l_2, \dots, l_k) \\ l_i &= -y_i \log \sigma(x_i) - (1 - y_i) \log(1 - \sigma(x_i)) \end{aligned} \quad (2.9)$$

where y_i denotes the ground-truth label for diagnosis d_i in patient p 's records and x_i is the prediction made by the model.

2.4.6.2 Low-level Diagnosis Code Ranking

As this task is framed as a ranking problem, we rely on score calculation between a patient and diagnoses. In particular, given a patient representation $f_{p,d} \in \mathbb{R}^{F'}$ learned with respect to diagnosis d , the score of diagnosis d for patient p is defined as the dot product between their representations:

$$\text{score}(p, d) = f_{p,d} \cdot h'_d \quad (2.10)$$

where h'_d denotes the transformed diagnosis node embedding parameterized by $W_d \in \mathbb{R}^{F' \times F}$ and bias vector $b_d \in F'$, which is in the same space as $f_{p,d}$.

Using this score definition, we optimize the model using a hinge loss formulated as:

$$\max(0, -\text{score}(d, p) + \text{score}(\sim d, p) + \epsilon) \quad (2.11)$$

where $\sim d$ is a negative diagnosis sampled for this patient and ϵ is the hinge margin.

2.4.6.3 Unsupervised Node Embedding

Besides the guidance of the supervised task, the network structure and relation of nodes can provide additional information that can be embedded in node representations. To capture this information, we employ an unsupervised network embedding objective similar to [9]. Formally, given a node i and its random neighbor j , we calculate the probability of observing j as a neighbor of i , conditioned on the type of the simple or meta path r connecting them, as follows:

$$P(j|i; r) = \frac{\exp(h_i \cdot h_j)}{\sum_{j' \in \text{Dest}(r)} \exp(h_i \cdot h_{j'})} \quad (2.12)$$

where $\text{Dest}(r)$ is the set of all nodes that are possible destinations on a path of type r and h_i and h_j are the embedding vectors of nodes i and j , respectively. As the above probability becomes expensive to compute in large networks, we instead use negative sampling [40] to

approximate the probability:

$$\begin{aligned} \log P(j|i; r) &\approx \log \sigma(h_i \cdot h_j + b_r) + \\ &\sum_{l=1}^k \mathbb{E}_{j' \sim P_n^r(j')} [\log \sigma(-h_i \cdot h_{j'} - b_r)] \end{aligned} \quad (2.13)$$

The supervised objectives we introduced, try to learn the node embeddings suitable for the diagnosis prediction task, while the unsupervised model embeds more general knowledge about the relation and proximity of nodes. To combine these two types of models, we follow the joint optimization approach suggested in [54] and define the following objective:

2.4.6.4 Combining the Supervised and Unsupervised Models

The supervised objectives we introduced learn the node embeddings suitable for the diagnosis prediction task, while the unsupervised model embeds more general knowledge about the relations and proximity of nodes. To combine these two types of models, we follow the joint optimization approach suggested in [54] and define the following objective:

$$\begin{aligned} \mathbb{L}_{joint} &= \omega \mathbb{L}_{unsupervised} + \\ &(1 - \omega) \mathbb{L}_{supervised} + \lambda \sum_i \|h_i\|_2^2 \end{aligned} \quad (2.14)$$

where $\omega \in [0, 1]$ sets the weight used when sampling a model to train at each training step.

2.5 Experiments

In this section, we provide qualitative and quantitative evaluations of HTAD, demonstrating its superior performance to existing models and its interpretability advantages.

2.5.1 Dataset

All evaluation experiments in this study are conducted using MIMIC-III database [43]. For data preparation and preprocessing, we follow the steps introduced a recent study on stan-

standardizing and benchmarking this dataset [36]. Accordingly, a total of 42,019 unique hospital admissions are included for modeling, 35,725 of which are used for training and 6,294 of which are used for testing. A mean of 11 diagnosis codes are recorded for each admission with 6016 diagnosis codes overall. [36] also introduces a set of manually selected features for model training, which we rely upon in our time series node aggregation process. Furthermore, for the task of high-level diagnosis prediction, we rely on the 25 disease phenotype groups introduced in this study.

2.5.1.1 Evaluation Metrics

Prediction of high-level disease groups is considered a multi-label classification problem. Accordingly, we follow existing works and employ Micro, Macro, and Weighted AUC-ROC scores to evaluate this task.

On the other hand, the exact diagnosis code prediction task is considered a ranking problem. Following the common approaches in the evaluation of large-scale ranking tasks [55], the ranking is conducted on a list of 100 codes, consisting of the original positive codes and a number of negatively sampled diagnosis codes. We evaluate our performance on this task using the Mean Average Precision at K (MAP@K), where K is set to 4, 6, 8, and 10.

2.5.2 Baselines

We compare our proposed model, HTAD, to recent studies that have achieved state-of-the-art results in diagnosis prediction, including those using manual feature selection as well as those relying on unsupervised EHR embedding. We also evaluate variants of HTAD to demonstrate the effectiveness of each of its components. A comprehensive list of models evaluated is as follows:

- Std-LSTM [36]: An LSTM-based model for predicting high-level diagnosis groups, introduced as the standard baseline for diagnosis prediction task.

- MMDL [37]: A multi-modal deep model for diagnosis group prediction that relies on a comprehensive set of hand selected features extracted from categorical and time series records in EHR.
- SAnD [56]: A recent study that employs a self-attention mechanism when modeling the EHR data. This study relies on manual feature extraction as well.
- Med2Vec [38]: An influential skip-gram based model for embedding health records. As this model is used to learn node embeddings and not for prediction, we employ mean aggregation of the embeddings it learns to represent patients based on their records and rely on supervised prediction methods similar to those used in HTAD.
- HeteroMed [9]: An HIN embedding method for modeling EHR data. Comparing to HeteroMed can directly reveal the benefits of learning record importance scores, as its basic architecture is similar to HTAD’s.
- HeteroMed_{MLP}: A variant of HeteroMed that we use for the group-based diagnosis classification task, obtained by replacing the hinge loss objective with HTAD’s multi-label classification one, to achieve a fair comparison.
- HTAD_{noAttnGrp/noTS}: A variant of HTAD that does not employ the group-based attention introduced in section 2.4.3. This model also excludes time series data so that the performance comparison to HeteroMed is solely focused on the attention mechanism used.
- HTAD_{AttnGrp/noTS}: A variant of HTAD that employs the group-based attention. For fair comparison with HeteroMed, this model excludes the time series data as well.
- HTAD: Our proposed model, employing group-based attention along with time series node aggregation.

2.5.2.1 Implementation Details

We implemented HTAD in Python using TensorFlow [57]. HTAD is trained using the Adam optimizer [58] and the learning rate of the optimizer, the batch size, the node embedding size, and the attention vector size are set to 0.001 and 32,256, and 128 respectively. When using grouped attention vectors, diagnosis groups are formed based on the CCS hierarchical coding system [59]. Furthermore, the LSTM model used in node-level time series aggregation is pre-trained using the model configuration proposed by the Std-LSTM model [36].

Our implementation of HeteroMed shares its code base with HTAD, particularly in network formation and unsupervised node embedding training. For a fair comparison, both models use the same set of hyperparameters and meta paths when training the unsupervised node embedding task. The metapaths used are: $labt \leftarrow pati \rightarrow diag$, $diag \leftarrow pati \rightarrow symp$, $labt \leftarrow pati \rightarrow symp$. Furthermore, we observed that running the unsupervised part as a pre-training step provided the best results for low-level prediction in HTAD, and as such for both models we do not employ joint training for this task. However, joint training is employed in all other tasks. Med2Vec is trained with an embedding size of 256, and the MMDL and SAnD models are run using the same parameters and setups suggested in their studies. Experiments were run on one NVIDIA GeForce RTX 2080 Ti GPU and two cores on an Intel Core i9-7920X CPU.

2.5.3 Evaluation of Disease Phenotype Classification

Table 2.2 lists the results obtained from evaluating our models on the diagnosis group classification task. Overall, we observe that HTAD outperforms all the baselines we investigated. Inspection of results further demonstrates that:

- $HTAD_{AttnGrp/noTS}$ shows notably higher performance than $HeteroMed_{MLP}$. This comparison is important as it demonstrates the effectiveness of our target-attentive aggregation mechanism versus models that otherwise share the same structure.

Table 2.2: Phenotype Classification Results

Model	AUC-ROC		
	Micro	Macro	Weighted
Std-LSTM	0.821	0.77	0.757
MMDL	0.819	0.754	0.738
SAnD	0.816	0.766	0.754
Med2Vec	0.815	0.748	0.741
HeteroMed	0.831	0.745	0.739
HeteroMed _{MLP}	0.864	0.788	0.786
HTAD _{noAttnGrp/noTS}	0.871	0.829	0.815
HTAD _{AttnGrp/noTS}	0.874	0.832	0.818
HTAD	0.880	0.843	0.828

- Compared to HTAD_{noAttnGrp/noTS}, HTAD_{AttnGrp/noTS} shows slightly better performance. This indicates that defining independent attention vectors as in group-based attention can be easier to train even when we are working with limited set of diagnoses.
- HTAD shows better performance than HTAD_{AttnGrp/noTS}, which is expected as the latter does not utilize the time series information in our dataset.
- HeteroMed_{MLP} outperforms HeteroMed by a considerable margin. This is in line with our expectations, as the original ranking objective used in HeteroMed may not be op-

Table 2.3: Exact Diagnosis Code Ranking

Model	MAP@4	MAP@6	MAP@8	MAP@10
Med2Vec	0.752	0.743	0.738	0.714
HeteroMed	0.866	0.843	0.814	0.805
HTAD _{noAttnGrp/noTS}	0.867	0.842	0.813	0.806
HTAD _{AttnGrp/noTS}	0.888	0.848	0.821	0.810
HTAD	0.890	0.881	0.865	0.923

timal for multi-label classification, and we expected that adjusting that could improve the performance.

- HeteroMed_{MLP} shows performance distinctly superior to that of the methods that rely on deep neural networks (SAnD, Std-LSTM, MMDL). This can be attributed to the fact that information networks eliminate the need for manual feature selection and allow for the incorporation of all clinical records. HeteroMed_{MLP} also outperforms Med2Vec, which is expected as it employs a more semantic-aware node representation learning approach.

2.5.4 Evaluation of Exact Diagnosis Code Prediction

The feature extraction based studies introduced for evaluation of the previous task have not approached the task of exact disease code prediction, mainly due to the huge size of the prediction space. In this study, we evaluate variants of our model against HeteroMed and Med2Vec, results of which are presented in Table 2.3 that shows:

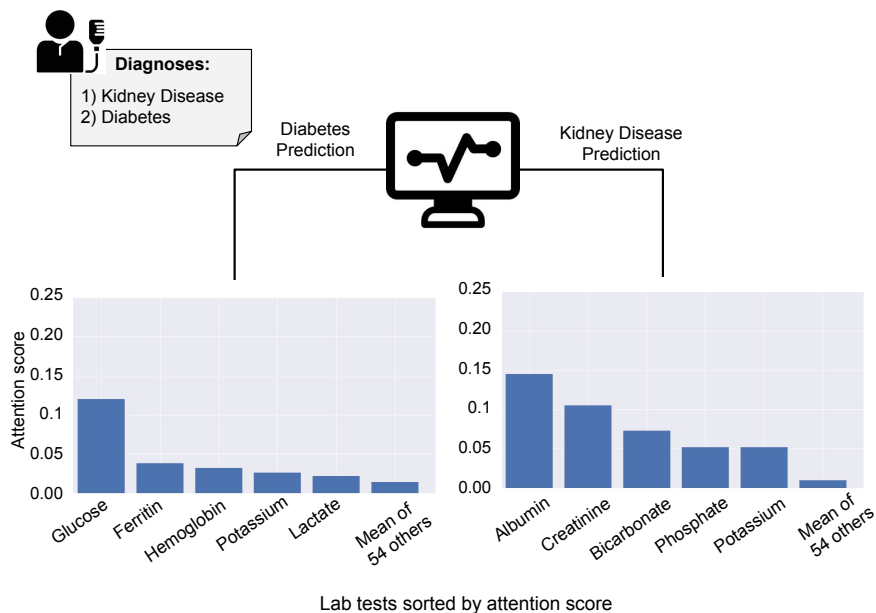


Figure 2.3: Distribution of attention scores for prediction of kidney disease and diabetes in a patient presenting with both conditions

- HTAD, which incorporates time series data as well as group-based attention, outperforms all other models.
- Similar to the high-level classification task, a comparison between $\text{HTAD}_{\text{AttnGrp/noTS}}$ and HeteroMed reveals the significance of employing hierarchical attention mechanism in node-aggregation.
- The performance gain of $\text{HTAD}_{\text{AttnGrp/noTS}}$ compared to $\text{HTAD}_{\text{noAttnGrp/noTS}}$ is significantly greater in this task. This gain can better demonstrate the advantage of using the group-based attention mechanism. As discussed before, sharing attention vectors among similar diagnoses can result in better performance for less common ones that otherwise remain under-trained.

2.5.5 Analysis of Attention Mechanism

Besides the performance improvement that our proposed hierarchical attentive architecture offers, one major benefit it provides is the interpretability of its results. We illustrate this in the node-level aggregation process in Fig. 2.3. We consider a patient diagnosed with both diabetes and kidney failure and study the importance score assigned to each of his 59 laboratory tests when predicting these two conditions.

The first important observation from this figure is that the set of laboratory tests the model attends to varies between the two diagnoses. As the figure shows, the highest attention score for the detection of diabetes is given to blood glucose level, which is a key predictor for diabetes. Similarly, the laboratory tests listed for kidney failure are highly indicative of this condition.

Additionally, we observe a larger skewness in attention scores when predicting for diabetes, with glucose having a notably higher score than other labs, than we do when predicting for kidney disease, where attention scores are more evenly distributed. This can be attributed to the fact that kidney failure is indicated by multiple factors while blood glucose is a single key indicator of diabetes. Insights such as these can be highly beneficial in supporting the diagnosis decision process.

We next analyze the attention scores in the type-level aggregation. Fig. 2.4 is a box plot demonstrating the range of attention weights assigned to different type-level embeddings across all the diagnoses in our test set. As we can see, the procedures and laboratory tests are overall our main predictors of diagnoses. However, there is more variance in procedure scores than in laboratory test scores, indicating that the predictive power of this category varies across diagnoses.

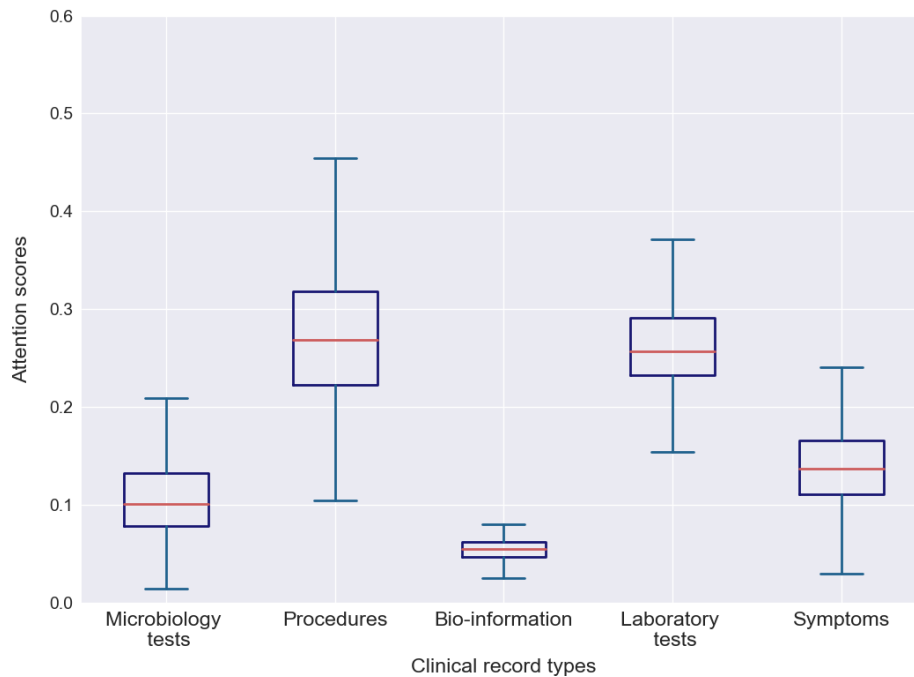


Figure 2.4: The distribution of attention weights among various record types

2.6 Conclusion

In this chapter, we introduced HTAD, an HIN based model incorporating a hierarchical attention mechanism for diagnosis prediction using EHRs. In HTAD, a patient representation is learned through a target-attentive aggregation of its clinical records’ embeddings, a process that allows distinguishing important record items for the prediction of a specific diagnosis. The novelty of this approach lies also in the interpretability it offers. Additionally, HTAD is capable of incorporating non-categorical records unused by past approaches. Experimental results demonstrate HTAD’s superior performance compared to the previous state-of-the-art methods and the interpretability of its predictions.

CHAPTER 3

Psychological Stress Detection in Older Adults With Cognitive Impairment Using Photoplethysmography

In this chapter, we discuss a system that leverages low cost mobile sensors to continuously monitor elderly patients for cognitive impairment risk factors. Our system generalizes well across patients of different genders and cognitive status, something that is important due to the underlying biological differences that can exist between these populations.

Psychological stress can have significant impacts on both physical and mental health. Chronic stress brings multiple adverse health outcomes, including cognitive difficulties. Unfortunately, there is a dearth of literature on the use of physiological sensor data to detect stress and analysis on the effects of gender and cognitive impairment on stress response in older adults, an especially important cohort as the population of the United States continues to age. We developed a physiological signal data collection system based on a portable device and a smartphone and used it to acquire signal data from 62 older adults (72 ± 10 years old; 30 cognitively healthy, 31 with mild cognitive impairment, and 1 with Alzheimer’s disease) in three conditions: rest, psychological stress, and recovery. Through a classifier trained on this data and our own analysis, we show the different impacts of psychological stress in healthy and cognitively impaired older adults as well as in males and females. Our classifier achieved a 0.84 F1-score when discriminating between the rest and stress conditions. Our proposed system can be used as a continuous stress monitoring system in real-world settings that is non-invasive, portable, and easy to use.

3.1 Introduction

Cognitive impairment, especially that associated with aging, has a significant cost both for our society and our healthcare system. The impact of Alzheimer’s disease (AD) and other forms of dementia is consistently increasing, with over 8.3 million years of potential life lost worldwide in 2012, more than the double the number lost in 2000 [60]. The estimated cost of lifetime care for an individual with dementia in the United States was \$350,000 in 2018, and in 2018 alone the unpaid assistance given by caregivers totaled over 18.5 billion hours [61]. The increasing prevalence of cognitive impairment, coupled with an aging US population, means that maintaining health in aging for our elderly will be an important challenge to address in the next decade [62].

Early identification of risk factors for cognitive impairment can allow for effective treatment and intervention, helping to reduce their impact on AD. Physiological dysregulation due to stress is associated with age-related cognitive deficits [63]. Studies have shown a positive correlation between cortisol levels and cognitive decline. Lupien et al. [64] followed a group of healthy adults as they aged, and they found that the adults that developed memory deficits had increasing cortisol secretion every year, resulting in high cortisol levels at the end of the study [65]. This study also found that persons with mild cognitive impairment (MCI), the risk state for developing dementia, tend to release more cortisol than healthy adults, while AD patients release more cortisol than those with MCI [65].

Unfortunately, cortisol measurement is not always readily attainable, and traditional methods of psychiatric assessment, such as clinical interviews and self-reports, have limitations. The latter assessment methods depend on retrospective summaries and subjective observations, which can result in reporting biases, inaccurate recall, and delayed treatment [66]. These assessments are also too coarse-grained to capture the dynamic nature of daily stress [67]. Timely intervention with a reliable mental stress detection algorithm can prevent stress from becoming chronic [68].

One alternative method for identifying stress is the use of a Photoplethysmography (PPG) sensor. PPG signals can provide valuable information about the cardiovascular system, such as irregularities in heart rhythms [69], and they can be captured non-invasively and at low cost using commercially available devices such as pulse oximeters, smartwatches, and even some smartphones. PPG sensors use a light source and a photodetector to measure the volume of blood flow based on the received light, giving insight into heart activity [69]. Heart rate variability (HRV), the variation in the time between heart beats, is the most critical marker for recognizing the reaction of the autonomic nervous system in response to stress [70]. Stress reactivity and recovery data obtained from HRV signals may deliver critical information on stress-related cardiovascular disease [71]. Unfortunately, there are few studies that use sensor data to examine how gender and cognitive differences in older adults correlate with stress response.

This study aims to address these unknowns by evaluating the feasibility of using a novel PPG-based stress monitoring system to continuously and reliably identify psychological stress. Building a useful stress monitoring and management solution requires the gathering of physiological sensor data for stress detection and validation through data-driven analysis. In our previous publication [72], we developed a mobile application to collect PPG signals using a wireless pulse oximeter and validated that PPG data can contribute to the identification of MCI in conjunction with conventional cognitive tests. In this paper, we apply machine learning algorithms to investigate psychological stress detection in older adults using only PPG-derived features. We also identify statistically significant correlations between cognitive status, gender, and stress response that may be useful for future analysis.

3.2 Related Work

Several techniques for detection of mental stress using objective readings from physiological sensors have been proposed. Some of these proposed techniques synthesize readings from

multiple sensors to perform their detection. In [73], researchers collected ECG, respiration, skin conductance, and electromyography (EMG) signals from 30 young adults. Mental stress was induced by having participants perform mathematical calculations, a logic puzzle task, and a memory task. The classification accuracy of the resulting system when identifying stress and non-stress conditions was 80%. In [70], 10 young adults wore a commercial smartwatch to obtain skin conductance, heart rate, and body temperature. Researchers induced stress using logic tasks and identified stress conditions with 84.5% accuracy using a K-Nearest Neighbors classifier. These studies above reported that mean heart rate, mean heart peak-to-peak intervals, and the standard deviation of the peak-to-peak intervals are the most useful time-domain cardiovascular features for stress detection.

Following along these lines, several studies have performed stress detection using only one wearable sensor. Stress Hacker [67] used a commercially available Empatica E4 wristband to obtain PPG signals in real-life settings. Twelve study participants provided a list of their daily stress dynamics, split into four tiers of stress, and classification accuracy was 88.6%. A study in [71] analyzed how age and gender affected heart rate changes in response to a social stress test. The researchers focused not only on the peak stress response, but also the recovery process after stress, allowing for an understanding of the physiological resilience of individuals. They collected heart rates using chest ECG device from 28 children, 34 younger adults, and 26 older adults. They found that the older group had lower heart rates, both while resting and after the induction of stress, than younger subjects. Additionally, they found that during the stress recovery phase the heart rates of older men returned to baseline, while the heart rates of older women remained elevated. These studies validate the feasibility of stress detection using only one sensor. However, while these studies did investigate how physiological responses to stress differ in older adults, they did not investigate the effects of cognitively impaired aging on stress response.

3.3 Methods

3.3.1 Physiological Recordings

For our study, participants are equipped with the Nonin fingertip pulse oximeter (Nonin Onyx II 9550; Nonin Medical, Plymouth, MN) [74] on the index finger of the non-dominant hand. The Nonin pulse oximeter is a clip type PPG device that measures heart rate (HR) and peripheral capillary oxygen saturation (SpO_2) with a 3 Hz sampling rate and PPG with a 75 Hz sampling rate.

We use a study-provided smartphone to wirelessly collect PPG signals from the pulse oximeter. The details of the system architecture and the data collection protocol are described in [72].

3.3.2 Experimental Protocol

Our data collection system measured the physiological state of sixty-two older adults (Table 3.1) in three study phases:

1. **Rest phase:** The baseline PPG recordings were three minutes long and made when participants were in a relaxed state and seated.
2. **Psychological stress phase:** Mental stress was induced by three cognitively challenging tests: the California Verbal Learning Test-II (CVLT-II) [75], the Auditory Consonant Trigrams (ACT) test [76], and the Stroop test [77]. The PPG recordings for this phase are from 20 to 30 minutes long, depending on how long participants took to complete the tests.
3. **Recovery phase:** After the completion of a multi-hour neuropsychological evaluation, the participants rested while another three minute PPG Signal was recorded.

Table 3.1: Demographic of Study Population

	Control (N=30)	Cognitively Impaired (N=32)	Total (N=62)
Age			
Mean(SD)	70.7 (11.6)	74.9 (9.3)	72.9 (10.6)
Median	73.0	76.5	74.0
Range	53, 92	56, 91	53, 92
Gender, n(%)			
Male	14 (46.7%)	16 (50.0%)	30 (48.4%)
Female	16 (53.3%)	16 (50.0%)	32 (51.6%)

3.3.3 Feature Extraction

We divide the time-series physiological recordings into one minute windows to generate features. Each window contains 180 samples for HR, 180 samples for SpO₂, and 4500 samples for PPG, and is labeled as rest, stress, or recovery. Statistical features, including minimum, maximum, range, median, mean, standard deviation, variance, and kurtosis are extracted from each signal in each window. A peak detection algorithm [78] is used to generate features from PPG signals in the time domain, such as the number of peaks per window (countPeaks), the mean of peak-to-peak (PP) intervals (meanPP), the standard deviation of PP intervals (SDNN), and root mean square of successive differences between successive peaks (RMSSD). A total of twenty-eight features are extracted from the physiological recordings. These features are rescaled using min-max normalization before they are used for model training.

3.3.4 Classification Scenarios

The dataset has 1954 unique 1-minute recordings from 62 participants. To compare stress reactivity based on cognitive status, the dataset is split into healthy aging (control) and cognitively impaired aging (MCI and AD). As the psychological stress portion of the study took 20-30 minutes per patient versus 3 minutes for rest or recovery, we have almost ten times more psychological stress samples (1582) than rest or recovery samples. To help models learn effectively despite this class imbalance we employed undersampling, keeping only 300 random samples of the psychological stress class. In the end we have a total 672 samples, with 350 from the cognitively impaired aging group and 322 from the control dataset.

We investigated five classic machine learning models for the task of detecting the psychological stress phase. These models included Logistic regression (LR), Random Forest (RF), Extra Trees (ET), Support Vector Machine (SVM), and Multilayer Perceptron (MLP). Each of these models was trained using scikit-learn [79]. We randomly split the subjects into 75% train set and 25% validation set, ensuring that samples from any subject are not used for both training and validation. We refrained from using deep learning models as these can require tens of thousands or even millions of training samples, and our dataset is relatively small.

As physiological state at baseline, during stress, and after stress may be different, we measured classification accuracy on four different classification tasks:

- *Rest vs. Stress*: a binary classification between the rest phase and the mental stress phase.
- *Stress vs. Recovery*: a binary classification between the mental stress phase and the recovery (post-stress) phase.
- *Stress vs. Non-stress*: a binary classification between mental stress and non-stress conditions (both rest and recovery phases).

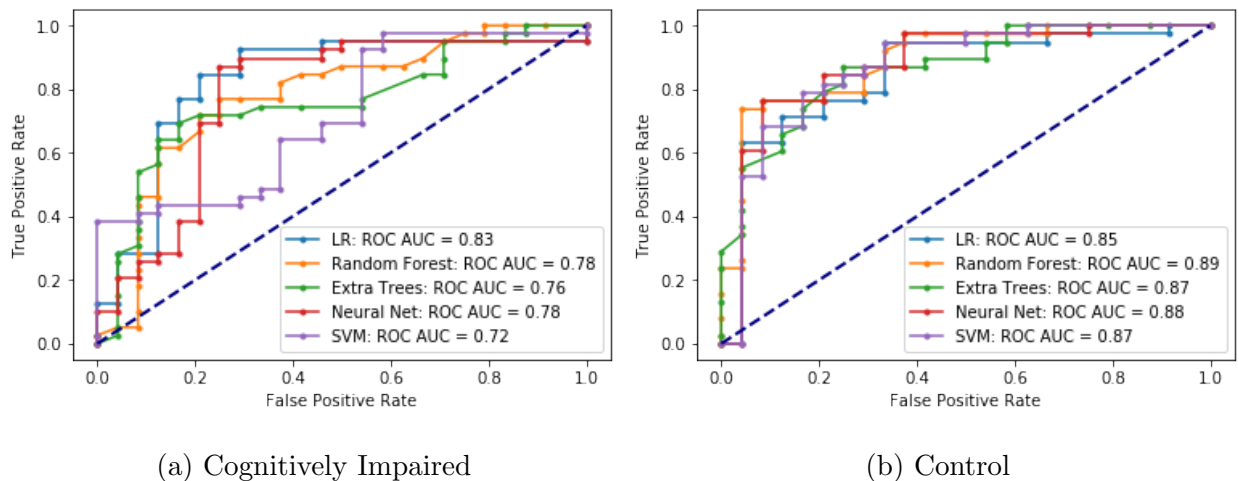


Figure 3.1: ROC curves and AUC

- *Rest vs. Stress vs. Recovery*: a 3-class classification among rest, mental stress, and recovery phases.

3.4 Experimental Results

Table 3.2: F1-score of Stress Detection across Cognitively Impaired Aging and Control Groups

Classification	Rest vs. Stress		Stress vs. Recovery		Stress vs. Non-stress		Rest vs. Stress vs. Recovery	
	Yes	No	Yes	No	Yes	No	Yes	No
Cognitive Impaired								
LR	0.76	0.83	0.63	0.76	0.71	0.78	0.54	0.61
Random Forest	0.69	0.78	0.56	0.62	0.56	0.80	0.45	0.55
Extra Trees	0.70	0.84	0.53	0.72	0.58	0.80	0.39	0.55
SVM	0.60	0.83	0.58	0.75	0.57	0.78	0.46	0.59
MLP	0.75	0.79	0.59	0.82	0.62	0.78	0.48	0.61

Table 3.2 shows the F1-score of each classification scenario for each of the learning models. The *Rest vs. Stress* binary classification tasks have the best classification accuracy (0.76 and 0.84) among the classification scenarios. We interpret this to mean that PPG-derived features have clear differences between rest and mental stress states. The next highest stress detection

accuracy is observed in the *Stress vs. Non-stress* task, where rest and recovery classes are combined for *Non-stress*. We observe the lowest classification performance in the three way *Rest vs. Stress vs. Recovery* task. One possible explanation for the lower performance on this task is that our PPG-derived features may not always significantly differ between the rest and recovery periods.

Table 3.2 also breaks down performance based on cohort. We observe that the learning models generally perform worse at identifying stress states in the cognitively impaired adults, perhaps indicating a less well defined physiological response in this group. The receiver operating characteristic (ROC) curves and the area under the curve (AUC) for the *Rest vs. Stress* classification (Figure 3.1) lend support to this possibility because the AUC of the cognitively impaired group (Figure 3.1a) is less than the control group (Figure 3.1b). These results indicate that stressors typically prompt cardiovascular changes in older adults, but the changes can be less pronounced in individuals with cognitive deficits. It is possible that since cognitively impaired older adults experience cognitive difficulties in everyday life, the cognitive tests we administered may not have been as stress-inducing to them as the tests were to the controls. We conclude that psychological stress detection using one minute windows of easily obtained PPG data is feasible, particularly using the logistic regression model, which attained high F1-scores while also being lightweight and efficient to run.

3.4.1 PPG Features Across Cohorts

In addition to investigating whether machine learning models could discriminate between stress states, we also investigated whether there might be any differences in the PPG-derived features that would allow a human to differentiate between the cognitively impaired and the control groups.

We plot the distribution of mean heart rate (meanHR), range of PPG signals (rangePleth), and RMSSD, shown in Figure 3.2. In [71], the heart rates of older women were found to increase after mental stress. However, our analysis suggests that women with cognitive

Table 3.3: Statistically Significant Correlation with Cognitive Status

Feature
RMSSD during stress
SDNN during stress
meanPP during stress
meanHR during stress (female only)

deficits do not have increased heart rates in response to these stressors. The basal heart rate of older women is higher than older men (Figure 3.2a). The mean rangePleth of older men is highest in the rest state, followed by the stress phase and lastly the recovery phase. On the other hand, older women typically have the highest rangePleth in the stress phase, and cognitively impaired women generally have higher rangePleth than the control (Figure 3.2b). RMSSD is an indicator of the psychological stress period for both genders, but RMSSD for women is higher and the difference between phases is bigger than for men (Figure 3.2c).

We also used a Python implementation of an independent t-test [80] to determine which measures of stress response were statistically different across cohorts, comparing the features of the MCI group’s PPG signals in each phase to those of the control group, split by gender.

3.4.1.1 Cognitive Decline

We show the list of all features we found to differ statistically significantly ($ps < .05$) when comparing participants of a given gender with and without cognitive impairment in Table 3.3. For meanHR, there was no statistically significant difference ($p < .05$) between normally aging and cognitively impaired males, and for females there was only a significant difference in the “during stress” phase.

Table 3.4: Statistically Significant Correlation with Gender

Feature
RMSSD at rest
RMSSD during stress
RMSSD during recovery (healthy aging only)
MeanHR at rest
MeanHR during stress (healthy aging only)
MeanHR during recovery
SDNN at rest (cognitively impaired aging only)
SDNN during stress
SDNN during recovery (healthy aging only)
meanPP (all stages)

Our analysis suggests that cognitive decline was associated with a statistically significant difference in meanPP during the mental stress phase in both males and females, and that no significant difference existed in the rest or recovery phases. We observe a similar pattern for RMSSD and SDNN, with cognitive decline being associated with a significant increase in HRV during the stress phase, but not the rest or recovery phases.

3.4.1.2 Gender

We show the list of all features we found to differ statistically significantly ($ps < .05$) when comparing participants of a given cognitive status across gender in Table 3.4.

3.4.2 Limitations

One limitation of our proposed approach is that the PPG sensor used may still generate poor readings if the subject is moving. In order to reduce possible motion artifacts in this study, we disallowed hand-movement while recording PPG signals. While the goal is to create a system that could continuously monitor mental stress throughout the day, it may be difficult to generate readings in situations where a user is not completely stationary. Reducing motion artifacts in PPG is an active area of research [81,82], and such techniques may help to make our system more flexible. While the Nonin fingertip pulse oximeter provides readings of high enough quality to be used in healthcare settings, it can still provide low quality readings when certain factors are in play, such as extreme ambient light, moisture, or fingernail polish [74].

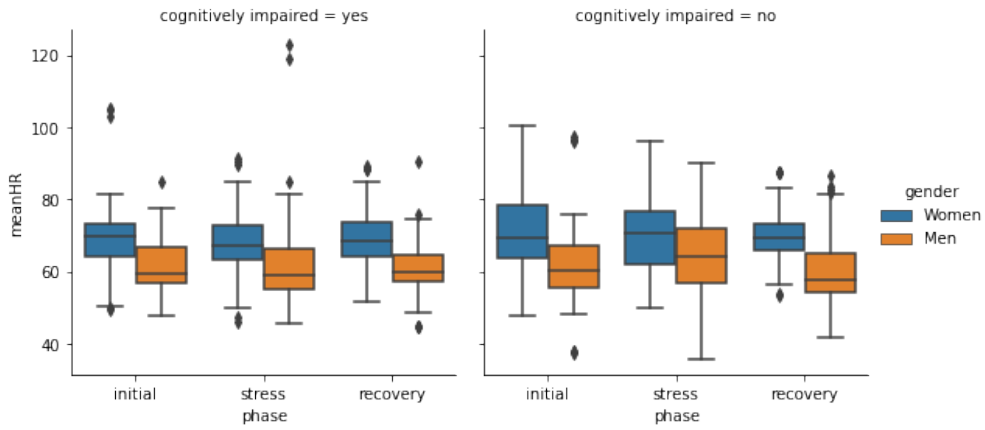
Another limitation of this study is that while many forms of mental stress exist, we only induced mental stress using three cognitively challenging tasks. While our analysis shows that our proposed system is useful for identifying the stress induced by these tasks, we have not yet proven whether it is also able to capture other forms of stress, such as financial or social factors, as well.

Lastly, stress is not simply an all or nothing state: stress exists on a spectrum, and individuals may experience widely varying stress levels throughout the day, sometimes rapidly transitioning between them. This is in contrast to our study, in which we sought only to determine whether an individual was in a relaxed or stress-induced state. More research must be done in order to determine whether our method can be extended to detect various stress levels in an elderly population.

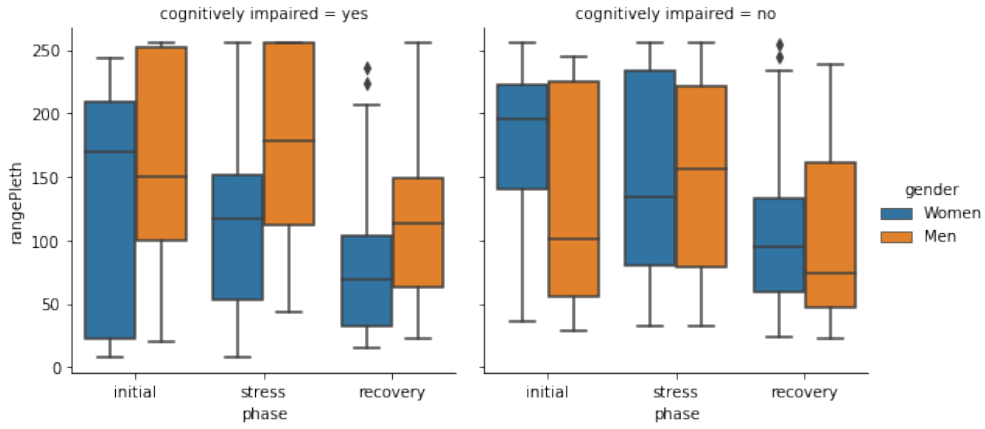
3.5 Conclusion

We successfully identified mental stress conditions based on features extracted from PPG sensor data from 62 older adults. Our high-accuracy classification results indicate that monitoring heart activity with a PPG sensor can assess psychological stress in older adults.

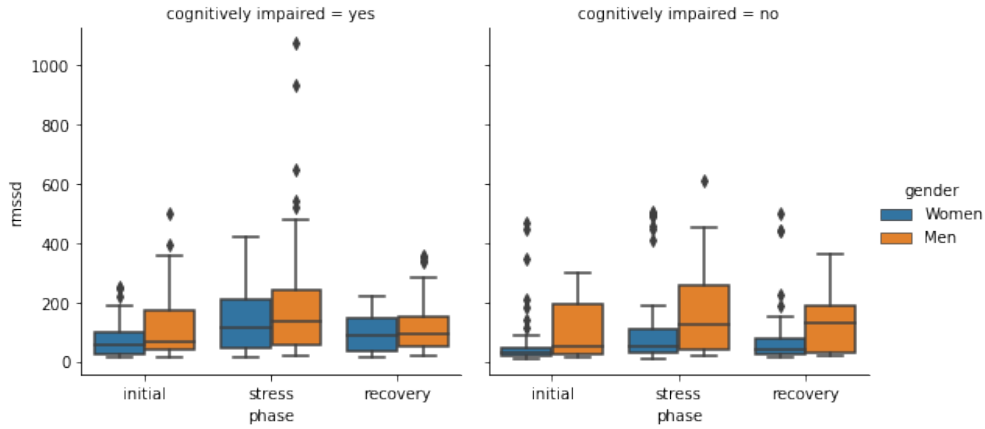
Based on our classification results, our features may not sharply distinguish between rest (pre-stress) and recovery (post-stress) states. Our further analysis with learning models and statistical analysis provides insight into the associations between gender, cognitive impairment, and psychological stress reactivity and recovery. With insight into these associations, we were able to achieve high accuracy in a diverse population with physiological differences. The system's portability and convenience, especially if coupled with a wrist-worn PPG device, give it the potential to monitor daily stress dynamics in real-life environments. The classification models we used can be run on smartphones, allowing for real-time stress detection, a stark contrast to the survey-based conventional methods of stress and MCI detection. Personalized stress monitoring and management using a PPG sensor is an exciting option for promoting positive health outcomes among older adults.



(a) Mean Heart Rate



(b) Range of PPG signal



(c) RMSSD

Figure 3.2: Distribution of the PPG features

CHAPTER 4

Predicting Rapid Kidney Function Decline Using EHR Data Despite High Missingness

In this chapter we describe a system for the prediction of rapid kidney function decline, as well as how the system can be used to identify risk factors for this rapid decline. While the progression of chronic kidney disease (CKD) to end stage kidney disease is typically a gradual process occurring over 1-2 decades depending on the cause, in some cases progression occurs rapidly and unexpectedly, with patients losing over 40% of their kidney function in just two years. Rare disease phenotypes such as this are easier to study when researchers have access to detailed accounts of the outcome in question. Automated tools can identify individuals at risk of renal function decline and facilitate disease mitigation, but electronic health record (EHR) datasets can be challenging to work with due to high levels of missing data. This paper describes a model for the prediction of rapid kidney function decline despite high missingness, capable of generating quality predictions despite many of the most useful features being missing more than 50% of the time. Additionally, the model's scores are used with combinations of features to identify population cohorts at risk for rapid kidney function decline. This study finds that that rapid eGFR decline can be detected most reliably among middle-aged patients with reduced eGFR, elevated urine albumin-to-creatinine ratio and elevated systolic blood pressure.

4.1 Introduction

Chronic kidney disease (CKD) is one of the most prevalent chronic health conditions in the world, with an estimated 10-13% of all individuals affected [83]. Not only can CKD lead to severe complications such as end-stage kidney disease (ESKD), but it is also a significant risk factor for cardiovascular disease, death from COVID-19, and overall mortality [84–88]. As a chronic condition, kidney health management typically includes lifestyle modifications, patient monitoring, and the use of medications to slow disease progression [89], as well as kidney transplantation, when indicated and available [90, 91]. However, some patients progress rapidly or non-linearly in the span of just a few years [90, 92, 93]. In fact, rapid decline that results in the loss of 40% or more of kidney function within two years is such a strong predictor of kidney failure that it is used as an alternative outcome to ESKD in clinical trials [94]. Little is currently known about why some patients experience more rapid kidney function decline than others. To better understand the complex interactions between various potential risk factors, a prediction system was built to examine 40% renal function decline, as based on estimated glomerular filtration rate (eGFR), using the CURE-CKD registry [95, 96] with two goals in mind:

1. To identify individual patients at risk of 40% renal function decline in order to facilitate disease mitigation, as well as to identify associated risk factors; and
2. To identify populations at higher risk of 40% decline, creating insights into key differences between (sub)groups

This paper describes the development of a set of predictive models for assessing the risk of rapid eGFR decline, defined as a $> 40\%$ decrease in eGFR over two (2) years [94], and identifies populations at higher risk of this rapid decline. The modeling and analysis conducted show that rapid eGFR decline can be detected most reliably among middle-aged patients with reduced eGFR, elevated urine albumin-to-creatinine ratio, and elevated systolic

blood pressure. The patients and subpopulations identified are strong candidates for closer study.

4.2 Related Work

4.2.1 Predicting CKD and Progression to ESKD

While the study of rapid kidney function decline is a relatively recent phenomenon [97], a great deal of work has been done to understand CKD progression [90,98,99], particularly to assess potential renal failure. One of the most widely used and validated tools for predicting the 2- and 5-year risk of renal failure is a Cox proportional hazards-based model developed by Tangri et al. [100]. This model has been validated in a multinational study of more than 700,000 patients, demonstrating good discrimination across cohorts diverse in their age, race, and diabetes status [101]. A review of other models for prediction of kidney failure shows that they often rely on myriad labs, including some such as albuminuria and plasma biomarkers that are inconsistently collected, even in at-risk patients [98,99,102–106]. As such, many of these tools have seen limited adoption given missing data and pragmatic implementation [98]. Some models, such as Tangri’s [100], address this challenge by providing multiple separate risk equations: a higher accuracy model incorporating many labs, and a simpler model using just more common lab values [107–110]. However, these models cannot conclude on the significance of the presence or absence of a specific variable, and the simpler models are typically less accurate [103]. Moreover, falling back to a simpler model relies on the assumption that data is missing completely at random [111] and can result in a loss of performance in situations where the absence of information, (e.g., fewer blood pressure readings in healthier patients) [112].

Still, a noted strength of proportional hazards models and other regression-based models is that each risk factor is assigned a particular hazard ratio or risk score. These interpretable scores then enable insights into factors associated with eGFR decline (e.g., hyperfiltration

in patients with diabetes [113,114], cardiovascular disease [115,116]). However, the insights provided by these models are limited as they do not consider the interactions between risk factors, instead assuming an additive risk model for each factor.

A comprehensive external validation study of eleven kidney failure risk models [103] found that when accounting for competing risks, the models investigated had average c-statistics of 0.74 on the European Quality Study dataset and 0.80 on the Swedish Renal Registry dataset [117]. This stands in contrast to an average c-statistic of 0.89 in past validations, which did not account for competing risk factors [103]. Most models with longer prediction horizons were found to overpredict risk considerably, with the 5-year kidney failure risk equation [101] overpredicting risk by 10-18% [103].

4.2.2 Predicting Rapid Decline

Although the above approaches are useful for understanding kidney function decline in CKD patients and patients at-risk of CKD on a population level, some studies have found that the trajectory of kidney function decline in the sickest patients varies greatly from others at risk of kidney failure. For example, O’Hare et al. analyzed the eGFR trajectories of Veterans Affairs patients in the two years before initiation of long-term dialysis [118], finding that the trajectories could be split into four distinct categories. For most patients, the decline into kidney failure was gradual and prolonged, with a decline of as little as 5 mL/min/1.73 m² per year. Yet some patients were observed to decline more precipitously, with an initial period of stable kidney function followed by a decline of more than 20 mL/min/1.73 m² per year - a greater than 40% decrease in eGFR within two years, a phenotype that is both rare [119,120] and clinically significant [97,119–121].

Indeed, while general kidney failure risk models incorporate risk factors to understand unique patient characteristics, they have not been built in a way that allows them to accurately predict those who would have rapid decline. Some effort has been undertaken to address this challenge, building prediction models that focus explicitly on rapid kidney func-

tion decline, variably defined as a loss of at least 3-10 mL/min/1.73 m² per year. Such models often incorporate additional variables beyond the ones used in standard CKD progression and kidney failure models, including blood biomarkers [122] and longitudinal electronic health record (EHR) data [123]. Though these works show good performance, they markedly require even more information in an already high data missingness environment. Additionally, their definitions of rapid decline are less strict than the definition used in this study (a > 40% decrease in eGFR within two years), resulting in decreased specificity for finding the most severe cases of rapid decline.

To address these issues, the chosen approach for this study is based on machine learning (ML) models capable of learning the complex nonlinear interactions between risk factors in the patients with a > 40% decrease in eGFR within two years. Additionally, the technique is designed with data missingness in mind, assuming it to be the norm rather than the exception, so that the models may use as many variables as possible while also understanding what the presence or absence of data may indicate.

4.3 Methods

4.3.1 Population

Analysis was performed using the CURE-CKD registry [95, 96], a large EHR dataset that contains records for more than three million patients (Figure 4.1). Patients are split into two groups based on kidney function: those with diagnosed CKD and those at-risk for CKD. Here, the CKD group consists of patients with eGFR < 60 ml/min/1.73 m², a CKD diagnosis by ICD-9/ICD-10 codes, or albuminuria; and the at-risk for CKD group consists of patients with CKD risk factors including diabetes and hypertension. These groups are based on EHR coding from patients with CKD (N=599,121) and at-risk for CKD (N=2,566,172). All patients have at least two years of follow-up. To reduce uncertainty, all patients without a year two eGFR value were excluded. In total, 21,641 (0.68%) patients met the definition of

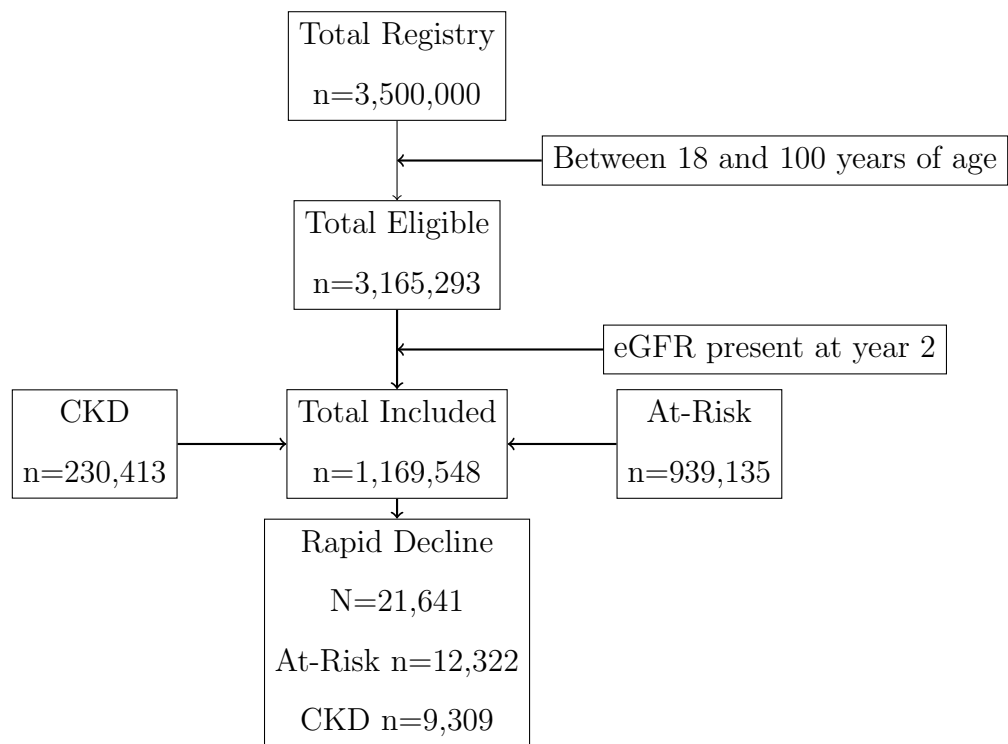


Figure 4.1: STROBE Diagram: Overview of participant groups by CKD and at-risk CKD categories in the study

rapid kidney function decline used in this study at the 2-year mark. The prevalence of rapid decline differs between the CKD and at-risk groups, with 9,400 patients (1.6%) meeting the criteria in the CKD group and 12,241 (0.48%) meeting the threshold in the at-risk group.

Table 4.1: Continuous variables used in analysis

Variable Name	Mean (SD)	Median (IQR)	Percent Missing
Age, years	58.34 (16.94)	59 (23)	0.00%
eGFR, mL/min/1.73 m ²	82.08 (23.76)	83.7 (31.6)	0.07%
HbA1c, percent	6.60 (1.63)	6 (1.4)	76.57%
UACR, mg/g	101.19 (443.45)	11.6 (27.6)	93.82%
UPCR, g/g	1.45 (2.56)	0.42 (1.34)	99.53%
Diastolic blood pressure, mm Hg	74.85 (10.20)	75 (13)	50.14%
Systolic blood pressure, mm Hg	128.32 (16.17)	127 (21)	50.07%
Ambulatory visit count	5.32 (6.60)	3 (6)	0.00%
Inpatient visit count	0.11 (0.41)	0 (0)	0.00%
ACE inhibitor, ARB use, days	7.75 (24.60)	0 (0)	0.11%
SGLT2 inhibitor use, days	0.07 (2.38)	0 (0)	0.00%
GLP-1 agonist use, days	0.13 (3.18)	0 (0)	0.00%
NSAID use, days	5.01 (19.65)	0 (0)	0.10%
Proton pump inhibitor use, days	4.81 (19.38)	0 (0)	0.12%

4.3.2 Model Variables

Variables in the CURE-CKD registry (Table 4.1 and Table 4.2) include: age; sex; race/ethnicity; eGFR; systolic blood pressure (SBP); hemoglobin A1C (HbA1c); the number of days of use at study entry for common medications that may affect kidney function, such as

Table 4.2: Categorical variables used in analysis

Characteristic	No. (%)	Characteristic	No. (%)
Sex		Rural-Urban Commuting Area	
Male	648 238 (55.4%)	code	
Female	521 310 (44.6%)	Urban	1 049 717 (90.6%)
Site Source		Large rural	57 284 (4.9%)
UCLA	978 602 (83.7%)	Small rural	26 179 (2.2%)
Providence	190 946 (16.3%)	Isolated	24 816 (2.1%)
Health		Medical Conditions	
Race/ethnicity		CKD from eGFR	205 896 (17.6%)
White	762 903 (66.9%)	CKD from eGFR, no race	210 093 (18.0%)
Non-Latino		CKD from ICD-9/10	35 649 (3.0%)
White Latino	41 143 (3.6%)	CKD from	20 816 (1.8%)
Black	54 455 (4.8%)	albuminuria	
Asian	69 476 (6.1%)	Diabetes mellitus	182 911 (15.6%)
American	11 203 (1.0%)	Prediabetes	115 929 (9.9%)
Indian		Hypertension	406 035 (34.7%)
Hawaiian	6 862 (0.60%)		
Other	113 892 (10.0%)		
Not categorized	80 148 (7.0%)		

angiotensin-converting enzyme inhibitors (ACEi) and angiotensin receptor blockers (ARB); and the diagnosis of hypertension, type 2 diabetes, and/or CKD. Urine albumin-to-creatinine ratio (UACR) and urine protein-to-creatinine ratio (UPCR) are also used, two lab tests that are used to establish kidney disease clinically [89]. Every variable in the CURE-CKD registry is used in this analysis. This approach was used as the importance of the various variables for rapid decline and the potentially complex interactions between variables are not known, and it would be undesirable to prematurely filter out information that may be useful.

4.3.3 Data Preparation

One challenge presented by data obtained from many registries, compared to clinical trial data, for instance, is that the patients' records are typically not complete. As such, there are varying levels of missingness in the dataset. For example, 50% of patients are missing SBP readings at study entry, as shown in Table 4.1.

The data preprocessing for the predictive models takes steps to minimize the effects of this missingness. The data preparation consists of three steps. First, as correlations between a missing value and a patient's health status are expected, missing values are not simply imputed. Instead, to preserve the information that a patient was missing a particular variable, indicator variables are added for each feature that contains missing information [124]. This indicator flag is then set whenever the corresponding variable is missing. By way of illustration, if a patient has no systolic blood pressure reading at study entry, the value of the "SBP is missing" variable is set to "1" for that patient. Second, tabular data is converted into a normalized form for ingestion by the model. This step consists of one hot encoding all categorical variables and normalizing continuous variables by subtracting the mean and scaling to unit variance. Lastly, after specifying the indicator variables, mean imputation is used to fill in missing values. Mean imputation was chosen so that the imputed values will have a minimal effect on the overall distribution of continuous variables.

The data processing pipeline is built using Pandas 1.1 [125] and NumPy 1.19 [126] for

Table 4.3: Hyperparameter search space for deep neural network (DNN)

Hyperparameter Name	Possible Values
L2 Regularization Weight	Range(.0001, .25), log sampled
No. Hidden Layers	Range(1, 16), linearly sampled
Learning Rate	Range(1^{-6} , 1^{-2}), log sampled
Units in first hidden layer	Range(16, 512), linearly sampled

Table 4.4: Hyperparameter search space for logistic regression (LR)

Hyperparameter Name	Possible Values
Inverse Regularization strength	20 values logarithmically spaced from 1^{-4} to 1^3
Regularization	L1, L2

data manipulation, and scikit-learn 0.24.2 [79] for imputation and scaling.

4.3.4 Predictive Modeling

With a goal of training a high performing model for the task of predicting rapid kidney function decline, three different families of classification models were investigated:

- **Deep neural network (DNN):** A multilayer perceptron with dropout and L2 regularization was developed. The number of layers in the model, the width of each layer, the weight for L2 regularization, and the learning rate were selected through a hyperparameter search with the Hyperband algorithm [127] implemented in KerasTuner [128]. The model that performs best on the validation dataset consists of one hidden layer, an L2 regularization weight of 2.9×10^{-4} , 272 units in the first hidden layer, and a learning rate of 8.6×10^{-4} . The DNN was built using TensorFlow 2.2 [129]. The full

Table 4.5: Hyperparameter search space for gradient boosted trees (GBT)

Hyperparameter Name	Possible Values
Learning Rate	0.05, 0.10, 0.15
Min Child Weight	1, 3, 5, 7, 10
Gamma	0.1, 0.2, 0.3, 0.4, 0.5, 1, 1.5, 2, 5
Subsample	0.6, 0.8, 1.0
colsample_by_tree	0.5, 0.6, 0.7, 0.8, 1.0
Max Depth	3, 6, 8, 10, 12, 15
No. Estimators	32, 100, 300, 500, 1 000, 2 000, 10 000, 15 000

hyperparameter search space is shown in Table 4.3.

- Logistic regression:** Logistic regression models can achieve high performance for many classification tasks, though they perform best when there is a linear relationship between model inputs and the target variable. Scikit-learn 0.24.2 [79] was used to train the logistic regression model. The SAGA solver [130] was used due to its speed on large datasets and support for L1 regularization. An exhaustive hyperparameter search was performed over the parameter grid shown in Table 4.4.
- Gradient boosted trees (GBT):** Gradient boosted trees are an adaptation of the random forest family of models that excel at learning nonlinear relationships between model inputs and variables of interest. A random search algorithm was used to identify the best hyperparameters with undersampling to balance the dataset. 600 hyperparameter combinations were sampled from the search space. The gradient boosted tree model was implemented using XGBoost 1.1.0 [131]. The full hyperparameter search space is shown in Table 4.5
- Gradient boosted trees ensemble (GBTe):** The use of an undersampling ap-

proach for rebalancing class frequencies in this dataset reduces the size of the training dataset by 96%. As such, differently undersampled training datasets may result in models specialized on subsets of the dataset. Combining multiple models with unique strengths to generate a final prediction is an approach that has seen great success in recent years, including in the medical domain [132, 133]. To explore this idea, the hyperparameter search employed for the gradient boosted trees model was repeated 100 times, each time with a unique seed for the undersampling process. Two ensemble methods were investigated. The first ensemble investigated was one built using the top 10% of undersampled GBT models based on validation set performance. Another ensemble using the center 90% of undersampled GBT models was also investigated, with the goal of minimizing overfitting to the validation set. Both ensembles generate predictions by taking the mean of the predictions from each submodel.

A variety of techniques were employed to mitigate the effects of dataset imbalance on model performance. Each model except for the GBTe model was trained with both an undersampling and an oversampling approach for equalizing class frequencies. Imbalanced-learn 0.9.0 was used to equalize class frequencies for the logistic regression and gradient boosted trees models [134]. All models were trained using a stratified 60/20/20 train/validation/test split, and the test split was a complete holdout. Bootstrapping with 100 bootstraps was done on the test set in order to determine confidence intervals.

4.3.5 Identification of Risk Factors

After training the predictive models, the risk distribution of 8,503,055 subgroups of the test set, obtained from all possible expert-defined combinations of the variable splits shown in Table 4.6, were computed. To ensure that all subgroups analyzed are clinically relevant, subgroups containing fewer than 15 patients were excluded. In the event that multiple subgroups contained identical sets of patients, the subgroups were collapsed into one. The

Table 4.6: Splits used for subgroup analysis

Split Criteria
Age 18-45 years, 56-65 years, or ≥ 66 years
Sex
Race/Ethnicity
Hypertension
ACEi/ARB use
SGLT2 inhibitor use
GLP-1 agonist use
NSAID use
Proton Pump Inhibitor use
HbA1c $> 8\%$
SBP > 140 mm Hg
CKD diagnosis ¹
Diabetes status

¹ Defined as union of all CKD diagnosis categories in Table 4.2

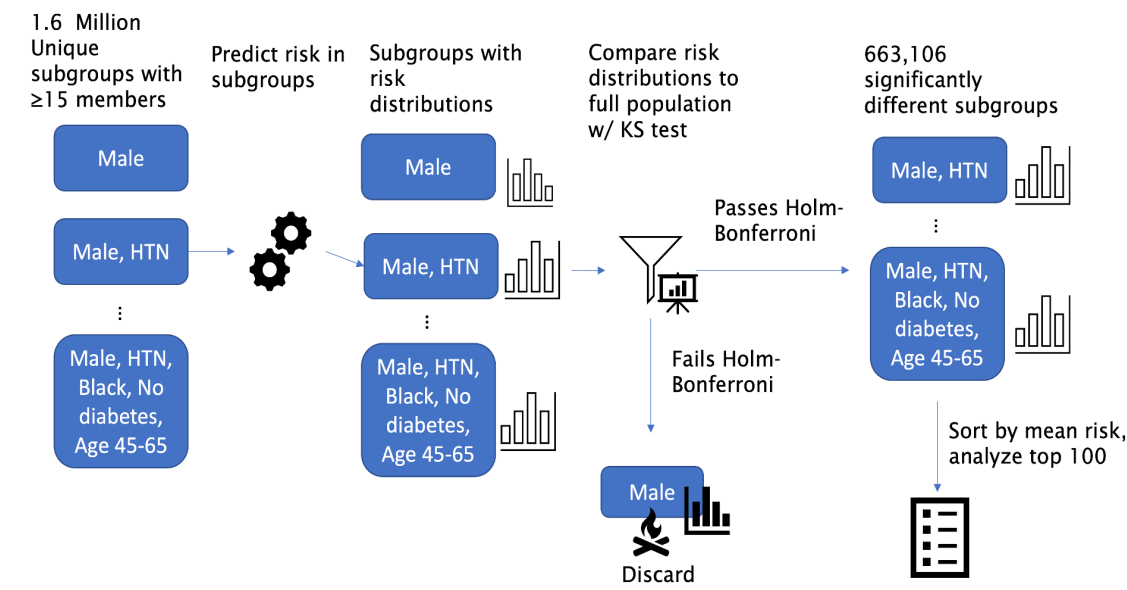


Figure 4.2: Illustration of the subgroup analysis process

risk distribution for each subgroup was then compared against the whole population’s risk distribution using the Kolmogorov-Smirnov (KS) test [135] implemented in SciPy 1.4.1 [136]. Subgroups with the highest risk of decline are identified using the KS test (Holm-Bonferroni method [137] with $\alpha = 0.05$) on the highest performing model. This process is illustrated in Figure 4.2.

4.4 Experimental Results

4.4.1 Model Performance Results

Results for each model are shown in Table 4.7. The primary metric for determining performance is the average precision (AP) score, as this metric provides good insight into the performance of each model despite the significant class imbalance in the dataset. The area under the receiver operating characteristic curve is also reported.

The center 90% GBTe model achieved an average precision of 0.099 on the test set, very similar to the 0.098 PR-AUC achieved by the top 10% GBTe model. The best performing

Table 4.7: Model performance on test set

Model Name	AP Score (95% CI)	AUC (95% CI)
Logistic Regression, undersampled	0.065 (± 0.00047)	0.72 (± 0.00085)
Logistic Regression, oversampled	0.064 (± 0.00049)	0.73 (± 0.00082)
Gradient boosted trees, undersampled	0.093 (± 0.00041)	0.74 (± 0.00077)
Gradient boosted trees, oversampled	0.092 (± 0.00041)	0.74 (± 0.00079)
Gradient boosted trees, center 90% ensemble, undersampled	0.099 (± 0.00043)	0.75 (± 0.00076)
Gradient boosted trees, top 10% ensemble, undersampled	0.098 (± 0.00043)	0.752 (± 0.00077)
DNN, oversampled	0.096 (± 0.00040)	0.75 (± 0.00085)
DNN, undersampled	0.094 (± 0.00042)	0.76 (± 0.00082)

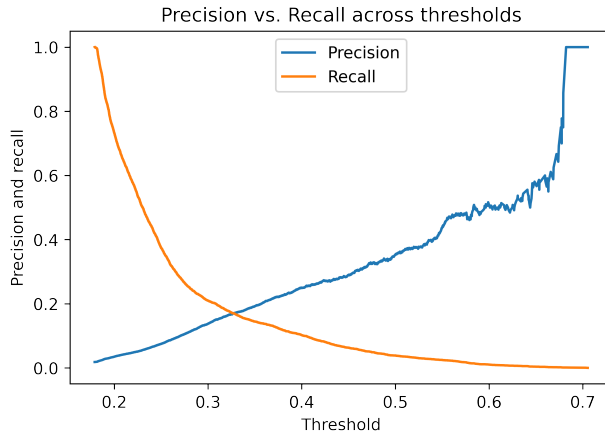


Figure 4.3: Effect of varying threshold on precision and recall for GBTe model on test set

DNN achieved an average precision of 0.096, and the best performing GBT and logistic regression models achieved an average precision of 0.093 and 0.065 respectively. The strong and similar performance from the DNN, GBT, and GBTe models may indicate that the interactions between risk factors are nonlinear and/or too complex to be captured by simpler models with less learning capacity, such as logistic regression without feature crosses. As the center 90% GBTe model achieved the highest average precision, its predictions are used for the remainder of the analysis. Precision and recall as a function of threshold for the center 90% GBTe model are shown in Figure 4.3.

4.4.2 Subgroup Analysis

Table 4.8: Most frequently occurring variables among top 100 highest risk subgroups

Variable name	Prevalence
CKD ¹	100%
Proton Pump Inhibitor use	100%
SBP >140 mm Hg	98%
HbA1c > 8%	87%
Age 45-66 years	79%

¹ Defined as union of all CKD diagnosis categories in Table 4.2

Of the 8,503,055 subgroups investigated, 1,640,355 were eligible for further analysis after collapsing identical subgroups and filtering out undersized groups. Of these 1,640,355 subgroups, 503,578 had a statistically significant increase in their predicted risk score above the population mean. The most frequent predictors of rapid eGFR decline across the 100 highest risk populations were identified, shown in Table 4.8. Patients in these 100 subgroups

Table 4.9: Race/ethnicity for patients with predicted risk above versus study population

Race/Ethnicity	Prevalence above 0.5 (%)	Prevalence overall (%)	Ratio
White Non-Latino	66.4%	66.9%	0.99
White Latino	3.2%	3.6%	0.89
Black	6.7%	4.8%	1.40
Asian	4.9%	6.1%	0.80
American Indian	1.5%	1.0%	1.50
Hawaiian	1.1%	0.6%	1.83
Other	10.5%	10.0%	1.05
Not categorized	5.6%	7.0%	0.80

were 13.7 times more likely to experience rapid decline than the overall study population (23.4% prevalence vs 1.8% prevalence).

4.4.3 Risk Factor Distributions

Figure 4.4 shows the distributions for several continuous variables in the patients with predicted risk above 0.5 compared with the distribution in the patients with predicted risk below 0.5. All differences in distribution are statistically significant (KS test, $p < .00001$). Table 4.9 shows the breakdown of race and ethnicity in the patients with predicted risk above 0.5 compared with the overall study population.

4.5 Discussion

Identifying 40% decline at two years in a dataset as sparse and imbalanced as the CURE-CKD registry is a challenging task. Despite this, the explored models were able to achieve adequate performance, enabling the detection of certain high risk patients with good speci-

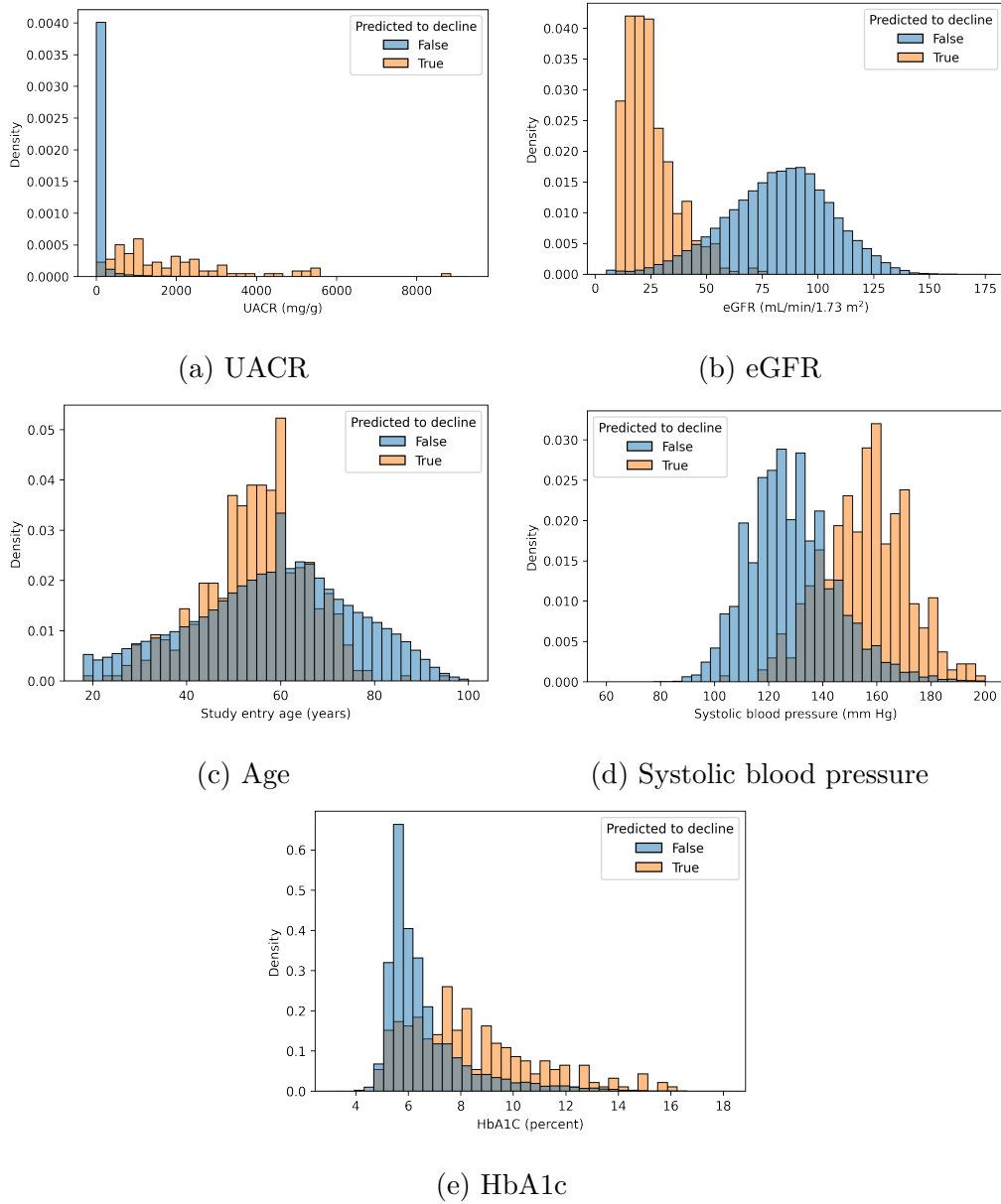


Figure 4.4: Distributions of features compared in patients with and without predicted decline (threshold = 0.5)

ficity. Additionally, the subgroup analysis provides interesting insights into the types of patients that the model is identifying as high risk. Notably, 79% of the top 100 high-risk groups include only middle-aged patients. This finding suggests that middle-aged patients should still be considered to have risk of rapid kidney function decline, even though they are younger than many patients on dialysis [138]. This observation is also supported by Figure 4.4c. Additionally, the highest risk patients (threshold=0.5) are consistently in CKD Stage 3a or lower, often with coexisting hypertension. Taken together, this suggests that the patients at highest risk of rapid decline tend to be those with CKD and hypertension, two significant health issues, despite being younger age than many CKD patients in the CURE-CKD registry (mean 54.6 years vs 71.2 years). UACR levels are also elevated for many of the patients identified as high risk, but 80% of patients identified as high risk have no UACR data in the study entry period.

Table 4.9 shows that Black, American Indian, and Hawaiian patients were overrepresented in the group with predicted risk > 0.5 , a finding in line with previous analysis of patterns of ESKD [138].

4.5.1 Importance of UACR to the Model

As shown in Figure 4.4a, the highest risk patients also had UACR values far from the overall study population’s median, suggesting that these high UACR values are connected to an outcome of rapid decline. As such, to assess the importance of UACR on the model’s performance on the test set, an experiment was conducted in which all UACR values from the 6% of patients with UACR lab results were removed, and the values were marked as missing. Risk scores were then generated for all patients without retraining the model. The hypothesis was that removing UACR records would result in a marked drop in performance because UACR is known to be one of the most important tests for assessing the risk of kidney failure [89, 139]. However, removing UACR values from the test set only resulted in a minor decrease in average precision score, with the average precision of the best performing gradient

Table 4.10: Prevalence of $\geq 40\%$ eGFR decline in CURE-CKD registry by eGFR level and albuminuria

eGFR Category	Albuminuria Category			
	Unknown	A1	A2	A3
G1	1.13%	0.72%	1.35%	6.74%
G2	1.59%	0.73%	2.39%	9.97%
G3a	2.61%	0.74%	2.55%	12.99%
G3b	4.54%	1.51%	2.33%	17.08%
G4	12.29%	3.31%	4.55%	29.52%
G5	7.63%	0%	2.33%	15.29%

Table 4.11: Incidence of $\geq 40\%$ eGFR decline in CURE-CKD registry by eGFR level and albuminuria

eGFR Category	Albuminuria Category			
	Unknown	A1	A2	A3
G1	4 838	153	71	65
G2	7 575	170	135	120
G3a	3 021	41	55	93
G3b	2 329	33	33	124
G4	2 020	14	24	152
G5	550	0	1	24
Percent	94%	1.9%	1.5%	2.6%

boosted tree ensemble decreasing from 0.099 to 0.096. Additionally, the true positive rate with a threshold of 0.5 did not change significantly, decreasing from 3.9% to 3.7%. While this result may seem at first to contradict the well established importance of UACR for assessing the risk of kidney failure [89, 139], this result may largely be a result of two characteristics of the CURE-CKD registry.

First, while most equations for kidney failure and renal decline risk have been developed on populations with dense features, lab results in the CURE-CKD dataset, especially for UACR, are very sparse, with only 6.2% of all patients having a UACR reading, with 23% of diabetic patients and 3.1% of nondiabetic patients having readings. This screening rate stands in contrast to CKD screening guidelines for diabetic patients, which recommend yearly UACR tests [140]. The low screening rates observed in the CURE-CKD registry are in line with those observed at other centers, which have reported UACR screening rates among type 2 diabetes patients of 40 to 60% [104, 105]. One possible reason for the lower screening rate reported in this study is that the analysis performed uses patients' records at entry into the CURE-CKD registry, dating back as far as 2005, when screening rates have been reported to be lower [106].

Second, in addition to UACR readings being rare in the CURE-CKD dataset, the presence or absence of a UACR reading is not strongly correlated with a $> 40\%$ decline in eGFR. In the full study population, the Pearson's correlation between rapid decline and the absence of a UACR reading is 7.7×10^{-4} , with a p-value of 0.4. This high p-value indicates that the null hypothesis is true and there is no association between the presence of a UACR reading and the outcome of rapid decline. In the subset of the population diagnosed with diabetes, the correlation coefficient is 0.045 with a p-value of 2.01×10^{-82} . This indicates that within the diabetic subset of the study population, patients with UACR readings are slightly less likely to experience rapid decline.

Taken together, the high missingness rate for UACR readings and the weak connection between the presence of a UACR reading and rapid decline means that patients without

UACR readings are as equal to experience rapid decline as patients with UACR readings (shown in Table 4.10), and that the majority of cases of rapid decline occur in patients without UACR lab results (shown in Table 4.11). As such, the model seems to have learned that while a high UACR result is a risk factor for rapid decline, the absence of a UACR lab result is uninformative, and other features must be relied upon for prediction. This is supported by the small change in TPR when UACR information is removed, indicating that it is likely that the model is relying on a blend of features to determine risk.

4.6 Conclusion

A risk model was developed for rapid eGFR decline using big data and its predictions, along with the KS test, were used to identify subpopulations with significantly high risk for rapid eGFR decline. This was achieved despite the high missingness in the dataset, showing that retrospective cohort studies can be feasible even when the data is collected from an EHR system and not from a purpose-built study. Additionally, subgroup and risk analysis were performed to identify common patterns among the patients at highest risk of rapid eGFR decline. These patients tended to be of poor health at a relatively young age, with multiple comorbidities. The patients and subpopulations identified are strong candidates for closer study.

CHAPTER 5

RimNet: A Deep Neural Network Pipeline for Automated Identification of the Optic Disc Rim

In this chapter we describe RimNet, a system for accurate optic disc rim segmentation across a spectrum of cameras and disease severities. Accurate neural rim measurement based on optic disc imaging is an important part of glaucoma severity grading and is typically performed by trained glaucoma specialists. There is room for error in this process as clinicians may not agree on the size of the optic disc or where exactly to define the edge of the optic cup. Additionally, glaucoma specialists comprise only a fraction of ophthalmologists, and less trained physicians may not be able to grade eyes as consistently, especially in edge cases. Neural rim measurement seems a prime target for the development of an assistive tool that could not only help to turn a manual process into a partially automated process, but could also provide an external check to help verify the decisions of clinicians. We aim to improve upon existing automated tools by building a fully automated system (RimNet) for direct rim identification in glaucomatous eyes and measurement of the minimum rim-to-disc ratio (mRDR) in intact rims, the angle of absent rim width (ARW) in incomplete rims, and the rim-to-disc-area ratio (RDAR) with the goal of grading optic disc damage.

We evaluate RimNet using both an internal dataset and the Drishti-GS dataset, which is used for external validation. Performance is evaluated by using clinician segmentations as ground truth and then measuring both intersection over union and error in mRDR, ARW, and RDAR. RimNet demonstrates acceptably accurate rim segmentation and mRDR and ARW measurements on the internal dataset, as well as competitive performance on Drishti-

GS.

5.1 Introduction

Glaucoma is the leading cause of irreversible blindness and the second leading cause of blindness worldwide [141]. Roughly half of all glaucoma cases are undiagnosed according to population-based studies [142, 143]. Early treatment preserves patient quality of life and reduces disease burden [144]. Therefore, identification of early glaucoma is key to preventative care.

Glaucoma diagnosis and grading are performed, in part, by evaluation of the optic nerve head’s neuroretinal rim of the optic disc. Metrics often include cup-to-disc ratio (CDR), rim-to-disc ratio (mRDR), and the *inferior > superior > nasal > temporal* (ISNT) rule, which compares the regional width of the neuroretinal rim [1]. Recent studies have shown the advantages of mRDR compared to ISNT and CDR for glaucoma classification accuracy [145].

The mRDR cannot adequately account for the degree of damage in optic discs with localized rim loss where the neuroretinal rim is noncontinuous or “incomplete”. A solution can be found in the Disc Damage Likelihood Scale (DDLS) proposed by Spaeth et al. [1]. DDLS accounts for incomplete rims by measuring the angle for which a rim is absent. This is called the absent rim width (ARW). Additionally, the scale accounts for disc size which affects the significance of the mRDR or ARW [1]. It is commonly accepted and has been incorporated into eye health guidelines for optometrists and ophthalmologists [146, 147]. DDLS is limited as a diagnostic tool by the need for expert time to accurately grade images. Automated high-efficacy DDLS grading could offer a powerful screening method.

In recent years, a confluence of several factors has led to efforts in automated glaucoma diagnosis and grading. First, studies have shown that automated algorithms can offer more consistent and reliable grading than human graders [147]. Second, there has been a rapid advancement in image segmentation, image processing, and deep learning neural networks.

In other fields, several neural networks outperformed human graders in image classification tasks [148]. This could allow for unprecedented accuracy in optic rim segmentation and glaucoma grading [2]. Finally, the optic disc exhibits characteristic alterations in glaucomatous patients, a prime candidate for automated segmentation and analysis. Together, these factors make automated glaucoma diagnosis and grading a possibility.

While DDLS also requires disc size analysis, automated rim segmentation with mRDR calculation for intact neuroretinal rims and ARW calculation for incomplete neuroretinal rims offers a step towards creating an efficacious, high-throughput diagnostic system for glaucomatous disc damage. Such a segmentation algorithm would need to be broadly applicable. Additionally, it would require an expansive learning capacity that could be applied to a variety of fundus images taken with different imaging modalities and with concurrent pathologies and normal variations. Convolutional neural networks offer such an approach [149].

The goal of this paper is to present a novel convolutional neural network algorithm for neuroretinal rim segmentation, automated mRDR calculation for intact rims, and ARW calculation for incomplete neuroretinal rims. This neural network algorithm offers an important step towards building an automated DDLS screening tool.

5.2 Methods

The study adhered to the tenets of the Declaration of Helsinki, was approved by UCLA's Human Research Protection Program, and conformed to the Health Insurance Portability and Accountability Act (HIPAA) policies.

5.2.1 Dataset

Optic disc photographs were taken from the UCLA Stein Eye Glaucoma database. The images were of varied magnifications and taken from slides and three different digital fundus cameras. All cameras were visible light cameras. No infrared, laser scanning, red-free, aut-

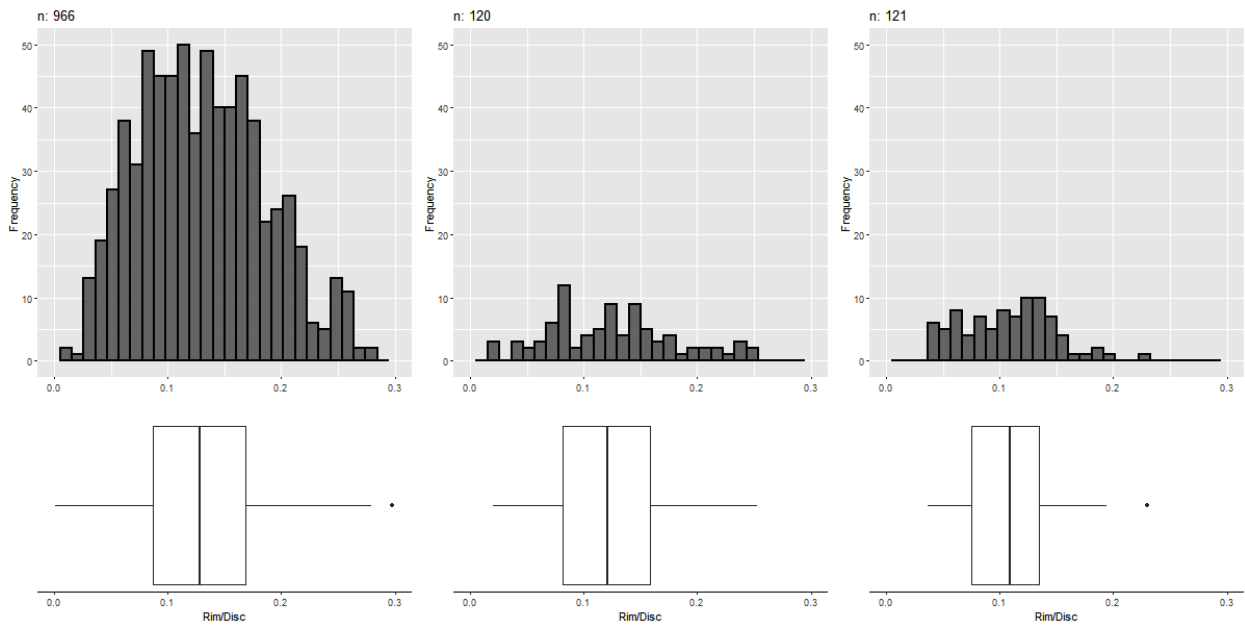


Figure 5.1: Distribution of mRDRs for Train, Validation, and Test Datasets. For each dataset, a frequency histogram is shown above with a box plot corresponding to the dataset below.

offluorescence, or hand-held smartphone-based cameras were used. Slide films were scanned and digitized at a third-party location.

The enrolled images met the following inclusion and exclusion criteria as deemed by two board-certified glaucoma specialists. Inclusion criteria include: (i) evidence of glaucomatous damage in the posterior pole; (ii) images had to be in focus, with discernible posterior pole and vasculature details. Exclusion criteria were concurrent non-glaucoma disease including optic neuritis, optic disc neovascularization, and vitreous hemorrhage that would impair visualization of the posterior pole. Globally, it was ensured that the full spectrum of glaucomatous damage, from early-stage intact neuroretinal rims to late-stage incomplete rims, were included while abiding by the inclusion and exclusion criteria. Figure 5.1 shows the mRDR distributions of our train, validate, and test set. The neuroretinal rim and optic cup were then manually segmented by one of three glaucoma specialists with a smart tablet and the image editing program GIMP. These masks were used as ground truth. The diagnostic categories for patients are shown in Table 5.1.

5.2.2 RimNet Model and Hyperparameter Architecture

A deep learning model for rim segmentation was developed as the centerpiece of the RimNet pipeline. The model was developed with Python 3.9.7 [150]. Libraries used include TensorFlow 2.6.0 [129], Segmentation Models 1.0.1, Keras Tuner 1.04 [128], OpenCV Python 4.5.3 [151], NumPy 1.19.5 [126], SciPy 1.7.1 [80], and scikit-learn 0.24.2 [79].

Optimizing the deep learning model requires a careful choice of model architecture and hyperparameters. The choice of hyperparameters can greatly influence the prediction speed, processing requirements, and accuracy of a neural network model [152]. These hyperparameters include the decoder, learning rate, optimizer, and loss function as shown in Table 5.3. The optimal combination of these parameters is task dependent. Whereas trial and error has been used in the past, newer architecture search techniques allow for the rapid evaluation of combinations of hyperparameters with the goal of optimizing a selected metric [152].

Table 5.1: Demographic distributions for internal dataset

		Scanned	Digital	Digital	Digital
		Slides	Camera 1	Camera 2	Camera 3
Gender	F	407	119	55	12
	M	302	85	44	11
Age	Mean	60.72	67.13	72.80	66.92
	SD	13.48	17.43	12.75	17.71
	Median	61.87	71.06	73.91	72.37
	IQR	15.79	16.33	12.86	22.54
	Min	9.36	6.92	16.19	17.48
	Max	90.05	96.10	94.41	86.17
Race/Ethnicity	Asian	90	34	24	2
	Black	63	22	8	1
	Hispanic	66	20	16	6
	White	366	100	45	12
	Other	53	5	3	0
	Unknown	71	22	3	2

To narrow the search space, an encoder of InceptionV3 was chosen based on literature review and computational efficiency. InceptionV3 was first published in 2015, outperforming popular encoders at the time with a fraction of the computation costs [153]. It has previously been used for medical segmentation [154, 155]. Our workstation uses NVIDIA 2080 RTX Ti graphics cards. Therefore, with limited computational efficiency, the selection of InceptionV3 was appropriate.

Transfer learning with ImageNet weights was used to initialize InceptionV3. No transfer learning was done for the decoder. Augmentations were used including a 20-degree rotation, a 10% vertical shift, a 10% horizontal shift, a horizontal flip, a vertical flip, up to a 30% random crop, a brightness change by ± 50 units, and a contrast limited adaptive histogram equalization (CLAHE) filter. Image down sampling was completed via a nearest neighbor algorithm. Color information was encoded using RGB channels with 8 bits per channel. The encoder and decoder were coupled using the Segmentation Models 1.0.1 library. The total number of trainable parameters was 29 896 979. No dropout layers were manually added.

Finally, a random search was performed using the Keras Tuner library [128]. The search parameters included the decoder, loss function, learning rate, and the optimizer. The rim Intersection over Union (IoU) was used as the segmentation metric. The full search space is documented in Table 5.3.

5.2.3 End-to-End mRDR Calculation Procedure

mRDR, ARW, and rim-to-disc-area ratio (RDAR) measurements are the final output of RimNet, which can be accomplished by accurate rim segmentation followed by image analysis. These two steps, along with preprocessing, led to the final framework for RimNet as shown in Figure 5.2.

The optic disc photographs were first resized to 224x224 with nearest neighbor interpolation in order to meet model specifications. A contrast limited adaptive histogram equal-

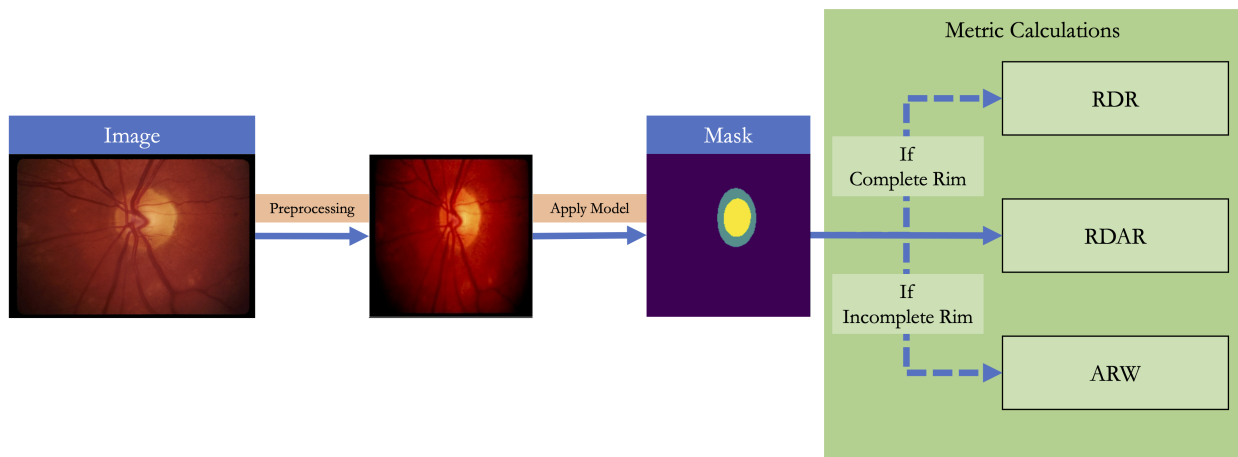


Figure 5.2: RimNet Pipeline, showing preprocessing, mask generation, and calculation of RDAR along with either mRDR or ARW depending on whether the rim is intact.

ization (CLAHE) filter was then applied to highlight distinctive features. The preprocessed image was submitted to the neural network model which generated a segmentation mask of the optic rim and cup. While a segmentation of the optic cup is not directly needed for mRDR or RDAR calculations, it was found that training the model to identify and segment the optic cup improved identification of incomplete rims and ARW calculations. Finally, the rim segmentation mask was resized to the dimensions of the original image to allow for accurate mRDR calculation and submitted to image analysis algorithms.

For mRDR, the algorithm first identified the center of the segmented optic cup using OpenCV. Vectors were created from the center of the cup to the boundary points. Boundary points were found using OpenCV. The number of vectors depended on the number of boundary points detected in the segmented rim. The intersection between the vectors and the segmented rim was taken as the rim width. The shortest rim width was identified and, through boundary point analysis of the rim, the disc diameter was found. Hence, the mRDR was calculated by dividing the rim width by the diameter. The RDAR was calculated by dividing the number of pixels of the segmented rim by the number of pixels in optic disc.

The Absent Rim Width (ARW) was calculated by first applying contour hierarchies to identify shapes within the rim segmentation. We rely on the fact that intact rims will have a “second shape” within the segmentation, the elliptical or circular form of the optic cup. Incomplete rims will not have this second shape. If the rim is classified as broken, 360 radial segments from the center are drawn to the edge of the rim. The radial segments that do not intersect the rim are those within the “broken” segment of the neuroretinal rim. The number of radial segments within the incomplete segment are added to give the ARW, one radial segment for each degree. If there were two breaks in a neuroretinal rims, the angles were added together and reported as one ARW. Examples of this can be found in the neuroretinal rim segmentation shown in Figure 5.3.

5.2.4 External Validation

The Drishti-GS database is a publicly available dataset of retinal images of glaucomatous eyes with manual cup and disc segmentations [156, 157]. Each image in the dataset is accompanied by four disc segmentations and four cup segmentations. In order to arrive at a single mask for each image, we took the true cup and disc masks to be the region of total agreement between all four segmentations. Before being used for evaluation, the images were first cropped around the optic disk, as they are available at a field of view of 30 degrees. Then, by subtracting the Drishti-GS cup segmentations from the disc segmentations, rim segmentations were acquired. These were used as “ground truth” for validation testing. The database has been used to compare performance between published optic cup and disc segmentation models through metrics such as IoU and Dice score [2–7]. Few investigators have attempted rim segmentations on the Drishti-GS database [6]. Therefore, RimNet rim segmentations were used to recreate cup segmentations to allow for comparison with other segmentation models. The intersection over union for cup segmentations (CupIoU) and disc segmentations (DiscIoU) were reported. Additionally, the Dice scores for the cup (CupDice) and disc (DiscDice) were reported.

5.2.5 Evaluation Criteria

The main outcome measures are the median absolute error (MAE) difference between the glaucoma specialists and RimNet for three metrics: mRDR, RDAR, and ARW. A secondary measure is the RimIoU, the IoU of the RimNet rim segmentation compared to that of the glaucoma specialists.

The mRDR, RDAR, and ARW have been explained above. Two measures of segmentation accuracy are also reported: Intersection over Union and Dice scores. The Intersection over Union (IoU), also known as the Jaccard distance, is a measure of segmentation accuracy. It compares the ground truth with the segmentation by reporting the ratio of the intersection

area over the union area. The Dice score for cup and disc segmentations are reported for the Drishti-GS dataset to compare segmentation performance. The Dice score compares the ground truth with the segmentation by reporting the ratio of two times the intersection area over the summed area of the ground truth and segmentation.

5.3 Results

Table 5.2: Glaucoma diagnosis for all 1 208 patients included in the RimNet dataset

Diagnosis	Count
Primary Open-Angle Glaucoma	530
Glaucoma Suspect	403
Chronic Angle-Closure Glaucoma	71
Low-Tension Glaucoma	47
Secondary Open-Angle Glaucoma	35
Capsular glaucoma with psuedoexfoliation	33
Anatomical Narrow Angle	27
Glaucoma secondary to Eye Infection	24
Pigmentary Glaucoma	15
Secondary Angle Closure	11
Congenital glaucoma	7
Juvenile Glaucoma	3
Acute angle-closure glaucoma	2

A database of 1 208 optic disc photographs of 121 eyes from 903 glaucoma patients were used for training, validation, and testing in an 80/10/10 split. Both scanned slides and original digital images were represented in the dataset. The average (\pm SD) age of the patients

was $63.7 (\pm 14.9)$ with a 43:57 male-to-female ratio. Full demographics including gender, age, and race/ethnicity are listed in Table 1. The average (\pm SD) visual field mean deviation (MD) was -8.03 ± 8.59 dB (range: -31.64, 3.59). Of the 1 208 optic disc photographs, 340 had incomplete neuroretinal rims. The diagnoses for the patients are listed in Table 5.2.

5.3.1 Hyperparameter Architecture

Table 5.3: Hyperparameter search space for RimNet

Hyperparameter Name	Possible Values
Decoders	U-Net, FPN, LinkNet, PSPnet
Loss Function	Categorical_Crossentropy, Categorical_Focal_Loss
Learning Rate	10^{-3} , 10^{-4} , 10^{-5} , 10^{-6}
Optimizer	Adam, SGD

Optimized parameters were found through the random search of 64 model combinations, detailed in Table 5.3 [58, 153, 158–167]. The combination of the InceptionV3 backbone and LinkNet architecture proved to be the most accurate [153, 159]. LinkNet is a lightweight decoder first published in 2017 [159]. Other parameters identified include the loss function of binary cross-entropy, learning rate of 10^{-3} , and the Adam optimizer [58].

5.3.2 Segmentation Network Results

The code used to train, run, and evaluate RimNet can be found on our public repository at <https://github.com/TylerADavis/GlaucomaML>. On the test set, an mRDR MAE (IQR) of 0.03 (0.05) was achieved on the intact rims while an ARW MAE (IQR) of 31 (89) degrees was achieved on the incomplete rims. 22 of 34 eyes with incomplete rims were correctly identified as incomplete on segmentation. A RDAR MAE (IQR) of 0.09 (0.10) was achieved

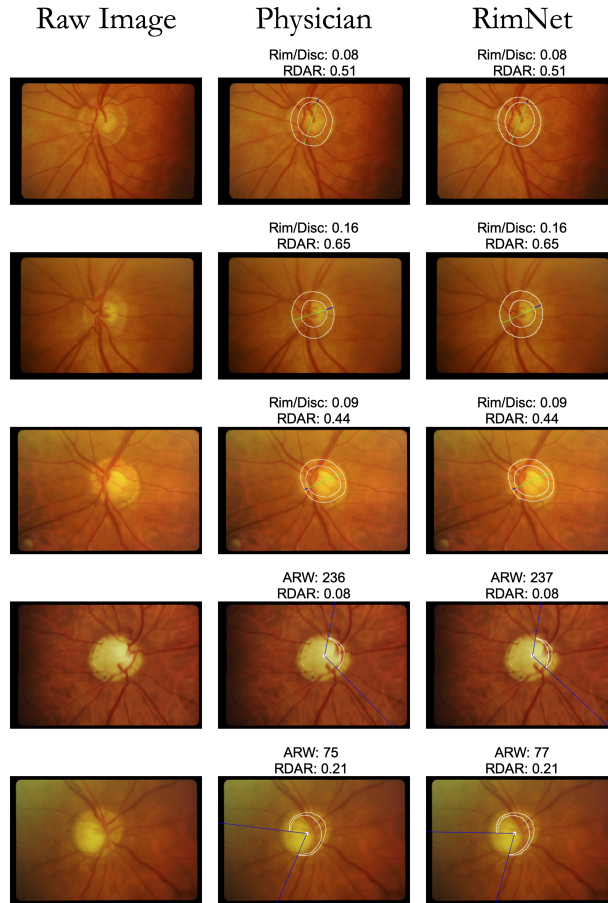


Figure 5.3: Segmentation Results. This figure demonstrates several examples of RimNet segmentation compared to physician segmentation. The left-most column shows the raw image. The middle column overlays the physician segmentation (white) over the raw image. The right-most column overlays the RimNet segmentation (white) over the raw image. In intact rims, green line shows the diameter and the dark blue shows the thinnest rim. In incomplete rims, the dark blue shows the edges of the segmentation.

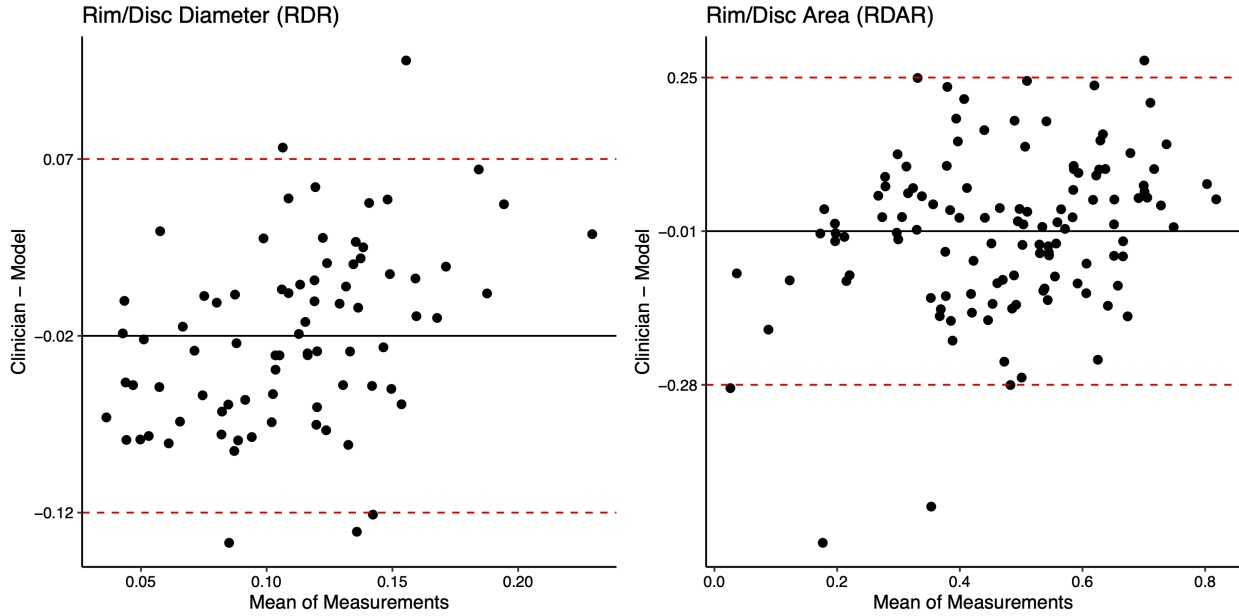


Figure 5.4: Bland-Altman plots showing the agreements in mRDR and RDAR between clinician and RimNet in test images. Red dashed lines indicate 95% confidence limits.

Table 5.4: RimNet Results on internal test set and Drishti-GS dataset. The ARW cannot be calculated on the Drishti-GS dataset because all rims are intact.

Metric Name	Internal Dataset	Drishti-GS
mRDR MAE (IQR)	0.03 (0.05)	0.03 (0.04)
ARW MAE (IQR)	31.00 (89.00)	N/A
RDAR MAE (IQR)	0.09 (0.10)	0.09 (0.10)
RimIoU (Intact Rims)	0.68	0.67
No.	121 (87 Intact, 34 Incomplete)	101 (101 Intact, 0 Incomplete)

Table 5.5: DRISHTI-GS segmentation performance of RimNet compared to published segmentation models [2–7].

Model	CupIoU	DiscIoU	CupDice	DiscDice
RimNet	0.77	0.91	0.86	0.95
Zilly et al. (2017)	0.85	-	0.87	0.87
Sevastopolsky (2017)	0.75	-	-	-
Edupuganti et al. (2018)	0.81	0.69	-	-
Al-Bander and Zheng et al. (2018)	-	-	0.83	0.95
Joshua et al. (2019)	0.79	-	-	-
Yu et al. (2019)	-	-	0.88	0.97

on all images. A RimIoU of 0.68 was achieved on intact rims, while a RimIoU of 0.45 was achieved on incomplete rims. The results of RimNet are presented in Table 5.4. Figure 5.3 demonstrates examples of RimNet segmentation results. To better examine the accuracy of the mRDR and RDAR calculations, the difference between the estimated values and the ground truths were calculated. Bland-Altman plots comparing the estimated and ground truth mRDR and RDAR are shown in Figure 5.4.

A comparison of RimNet segmentation on the Drishti-GS dataset to other published works is presented in Table 5.5. The mRDR MAE (IQR) was 0.03 (0.04) and the RDAR MAE (IQR) was 0.09 (0.10). The IoU of the optic cup (CupIoU) was 0.77 and a IoU of the optic disc (DiscIoU) was 0.91. The Dice score of the cup (CupDice) was 0.86 and Dice score of the optic disc (DiscDice) 0.95 was achieved.

5.4 Discussion

These results demonstrate that RimNet is capable of reasonably accurate segmentation and analysis of optic discs with both intact and incomplete rims. Spaeth et al. distinguished different DDLS grades by mRDR steps of 0.1 [1]. The MAE of the mRDR is well within this value, showing that RimNet segmentations are clinically relevant. For more advanced glaucoma with DDLS grades of 6 and above, the neuroretinal rim is incomplete and Spaeth et al. uses the ARW to distinguish grades. The five categories are less than 45 degrees, 45 degrees to 90 degrees, 90 degrees to 180 degrees, 180 to 270 degrees, and greater than 270 degrees. The minimum step is 45 degrees; the MAE falls slightly below that category at 31 degrees with 22 of 34 total incomplete rims correctly identified as incomplete. However, the IQR demonstrates a broad range of ARW. The error echoes the difficulties faced by the glaucoma specialists. While segmenting these severely glaucomatous rims to create the “ground truth” masks, glaucoma specialists often differed regarding where rims were interrupted and if rims were incomplete or intact. Though a forced consensus was eventually reached, this demonstrates the difficulty of the task and the variability of this “ground truth”. RimNet offers 65% accuracy in identifying incomplete rims and a relatively low ARW MAE. To our knowledge, RimNet is the first to offer such capabilities in published literature.

This work offers three improvements in the current landscape of optic disc segmentation. First, we utilized a dataset of 1 208 images with external validation on Drishti-GS [157]. Second, while we have still reported IoU and Dice scores, we have focused on more clinically relevant metrics such as mRDR, RDAR, and ARW. Third, our study is the first to focus on accurate segmentation of incomplete rims. RimNet is a useful step towards completely automating the DDLS algorithm.

Automated segmentation of the optic disc and cup have been previously explored. The original studies were initially based on image processing functions such as thresholding, level set, active contour, clustering, and component extraction with success on local and publicly

available datasets [168]. As early as 2001, Chrástek et al. offered an automated method of optic disc segmentation with filtering and edge detection, which achieved a segmentation accuracy of 82% [169]. In 2008, Liu and collaborators used level set and thresholding methods to achieve 97% accuracy when comparing algorithm-determined CDR ratio to clinical CDR ratio on a dataset of 73 images from the Singapore Eye Research Centre [170]. In 2015, Lotankar et al. used active contouring to achieve a 99% pixel-to-pixel accuracy on a private database of 150 images [171]. However, each of these approaches were limited in scope. Level-setting and thresholding would fail with images with decreased or increased intensity caused by pathological findings, which can be commonly seen on optic disc photographs such as peripapillary atrophy. This leads to overestimating or underestimating CDRs. Active contouring may similarly be affected by abnormal pathology or bright artifacts fixating on local maxima or minima within the image. Therefore, though these methods have proven efficacy, they can be improved upon.

An automated grading system for glaucoma diagnosis and progression needs a high efficiency, broadly applicable segmentation algorithm with an expansive learning capacity which could be applied to a variety of fundus images acquired with different imaging modalities with concurrent pathologies and variations. Though further work must be done, deep learning and convolutional neural networks may play an important role in the solution. They have an enormous learning capacity relative to their size [149]. Rapid advances in computational memory and processing speed have made neural networks more accessible for optic disc segmentation. Zilly et al. used ensemble learning to achieve 89% IoU on disc segmentation and 84% IoU on cup segmentation on the Drishti-GS dataset [3]. Sevastopolsky and coworkers furthered this work by using a modified U-Net to achieve a comparable accuracy in less than a tenth of the time [4]. More on segmentation efforts, both image processing functions and neural network attempts, can be found on a review article by Thakur and Juneja et al. [168].

Several groups have pursued automated mRDR and RDAR calculations. In 2019, Kumar

et al. proposed using an imaging processing technique called active discs to segment the optic disc and cup and perform general glaucoma classification (normal, moderate, severe) based on mRDR [145]. Though direct mRDR accuracy was not reported, an mRDR-based approach demonstrated high classification accuracy. In 2020, Martins et al. proposed a smartphone-based glaucoma diagnosis pipeline, which focuses on glaucoma classification and calculates RDAR [172]. However, RDAR results were not directly reported. More recently, Pachade et al. proposed an NENet model consisting of EfficientNetB4 and adversarial learning that achieved an area-under-the-curve (AUC) of 0.901 on RDAR calculation for Drishti-GS [173].

To the best of our knowledge, RimNet is the first engineering attempt to pursue segmentation and glaucoma grading efforts with incomplete neuroretinal rims. Thus, direct comparison of RimNet to other segmentation models is difficult. However, through the Drishti-GS dataset, an artificially-derived segmentation comparison is possible by recreating cup and disc masks from the RimNet rim segmentations. Table 5.5 demonstrates that RimNet performed well overall compared to recent segmentation models on the Drishti-GS dataset. While it outperformed several other models in CupDice, DiscIoU, and DiscDice segmentations, it was below average in CupIoU. These results must be understood in the context of three factors. First, the Drishti-GS images were available as 30-degree field of views. However, RimNet requires images centered and cropped near the optic disc margin. Therefore, RimNet has a significant information loss compared to other models that use the 30-degree field of view. Second, RimNet is unique in that it has been trained on both complete and incomplete rims. The models compared to RimNet have been trained only on complete rims. It is reasonable to expect a higher segmentation accuracy in these cases. Finally, the cup and disc segmentations produced by RimNet were artificially derived from the RimNet’s rim segmentation. By not directly predicting on the cup and disc, accuracy was lost. Considering these three factors, RimNet’s performance on Drishti-GS is acceptable. This is corroborated by the Drishti-GS mRDR MAE of 0.03 (0.04) and RDAR MAE of 0.09 (0.10), both of which are low.

The findings of this study need to be interpreted with the shortcomings in mind. First, the hyperparameter architecture search was limited by the computational and memory limits of our workstation, which uses NVIDIA RTX 2080 Ti graphics cards. We could not include larger models such as ResNet152 or EfficientNetB3 into our search due to these memory constraints. Second, the number of ground truth masks and optic disc images, particularly those of more severe glaucoma is limited. Greater numbers of samples diverse in glaucoma severity and race/ethnicity would allow RimNet to generate better segmentations, and thus increase the accuracy of its mRDR and ARW calculations. Lastly, while multiple clinicians created segmentations for this project, each of the images in the internal dataset used to train and evaluate this model was segmented by only one clinician. Leveraging the expertise of multiple clinicians for each segmentation, whether by averaging segmentations, segmenting by consensus, or conducting manual review of the dataset, may distill the knowledge of multiple clinicians into the model, potentially improving performance.

As we continue to develop this system, it may be worth exploring newer convolutional neural network segmentation architectures, such as UNet++ [174] and UNet+++ [175], as well as transformer based techniques such as TransUNet [176] and Swin-Unet [177]. These architectures have shown good performance on some medical segmentation tasks, and it would be interesting if they would be able to show such benefits for the task of neuroretinal rim segmentation as well. Additionally, Stein Eye has a wealth of stereoscopic fundus images available. It would be interesting to see whether the depth information encoded in these stereo pairs could successfully be leveraged by a model to generate more accurate segmentations.

RimNet brings glaucomatous detection and DDLS grading a step closer to full automation [1]. Automated grading of disc size is a necessary step to fully autonomous DDLS grading. A future goal would be to not only pursue full automation of DDLS grading, but to test their capabilities as diagnostic tools. One promising avenue for further investigation would be screening with smartphone fundoscopy. The increasing quality of smartphone

cameras have made smartphone funduscopy viable as a screening method [178, 179]. This, combined with automated DDLS grading, could provide a powerful screening tool to revolutionize glaucoma detection.

5.5 Conclusion

In conclusion, RimNet provides a method for high efficacy rim segmentation, mRDR, and ARW calculation. It also provides an example of how ophthalmic care be augmented by artificial intelligence. Though more work remains to be done, we believe that detection, diagnosis, and care of glaucoma can integrate with approaches such as these and aid ophthalmologists in decision-making to provide higher quality care for a global population of patients.

CHAPTER 6

DDLSNet: A Novel Deep Learning-Based System for Grading Funduscopy Images for Glaucomatous Damage

In this chapter, we build upon the work described in chapter 5 to build an end-to-end fully automated image analysis pipeline, DDLSNet, consisting of a rim segmentation branch (RimNet) and a disc size classification branch (DiscNet), to estimate the disc damage likelihood scale (DDLS). Extending RimNet by incorporating disc size information and estimating DDLS, a well established grading system, provides an output understandable by non-experts, especially when done in an interpretable manner as described here. By separating the calculation of rim width from the calculation of disc size, the system’s estimations can be validated more easily. Additionally, an automated system allows for rapid screening of large databases of ungraded fundus images, completing in a fraction of a second what may take upwards of a minute for a trained glaucoma specialist. Such rapid screening tools allow for easier creation of datasets for downstream studies.

DDLSNet was tested against manual grading of DDLS by clinicians, with the average score across clinicians used as “ground truth”. Reproducibility of DDLSNet grading was evaluated by repeating DDLS estimation on a dataset of non-progressing paired optic disc photos taken at separate times. On our internal dataset, DDLSNet achieved moderate agreement with clinicians for DDLS grading, as measured by weighted kappa score. This novel approach illustrates the feasibility of automated optic disc photo grading for assessing

glaucoma severity.

6.1 Introduction

Glaucoma is the leading cause of irreversible blindness worldwide with an estimated 80 million people affected in 2020 and a projected rise to 111.8 million people by 2040 [141, 180]. Glaucoma is asymptomatic in the early stages; untested individuals often remain undiagnosed until advanced symptoms are present. In developed countries, up to 70% of patients with glaucoma are undiagnosed, a number that rises in areas with less access to screening [181]. While patients with mild glaucoma have a quality of life comparable to that of healthy patients, the quality of life drastically decreases with more advanced glaucoma [182]. Early diagnosis and treatment allow for preservation of patient quality of life and is at the forefront of strategies for reducing disease burden [144].

Glaucoma diagnostic methods can be grouped into two categories: techniques that evaluate structural changes in the eye and techniques that evaluate functional changes in vision. Among those assessing structural changes, optical coherence tomography (OCT) and fundus photography are most often used in clinical practice. While OCT has been shown to have a high sensitivity for detection of structural glaucomatous changes, the high cost of the technique often restricts the device to large eye clinics or centers [183, 184]. This is especially problematic given that developing regions have the highest rates of undiagnosed glaucoma [181]. Moreover, the World Glaucoma Association considers the largest barrier to glaucoma screening to be cost [185]. In contrast to OCT, fundus photography stands as a lower cost option; new advances such as telemedicine screening and smartphone funduscopy have made fundus photography a feasible and financially viable option even in remote locations [178, 179].

While OCT and fundus photography allow for the structural findings to be captured, a mechanism is needed to classify such changes and correlate them with functional glauco-

matous damage. The Disc Damage Likelihood Scale (DDLS) is one such approach. DDLS is a well-established grading scale to correlate glaucomatous damage with progression of fundus photographs [1, 146, 181]. DDLS has been incorporated into the eye health professional guidelines for optometrists and ophthalmologists [146]. The interobserver agreement of DDLS even among glaucoma specialists can vary from 85% based on optic disc photographs to 70% based on clinical exam although intraobserver reliability is high [186]. This is especially troubling as DDLS scores can be used as the basis for referral by a variety of eye health professionals, and improper grading may result in missed opportunities for early intervention [146].

An ideal screening tool would be high-throughput, accurate, and reliable with high specificity. With the advent of neural network models and an increase in image processing capabilities, high specificity with acceptable sensitivity, together with high throughput, may be achieved with a neural network-based pipeline [187]. In this chapter, we present DDLSNet, a neural network pipeline which aims to accurately grade DDLS based on optic disc photographs with a combination of a rim segmentation neural network (RimNet) and a disc size classification network (DiscNet).

6.2 Methods

The DDLS grading criteria is shown in Figure 6.1. The DDLS score is determined by two features of the optic disc: the disc size and the narrowest rim width. Progression of glaucomatous damage is seen as enlargement of the optic disc cup and subsequent thinning of the optic disc rim. This thinning can be measured by the rim-to-disc ratio (mRDR) in intact rims. However, in severe glaucoma, the rim can be completely absent in certain areas. In these cases, the angle for which the rim is completely lost is measured. We call this the “absent rim width” (ARW) and we call these rims “incomplete”. These three features, mRDR, the absent rim width, and disc size, are the metrics needed to calculate DDLS.




























Stage	The Thinnest Width of the Rim (Rim/Disk Ratio)			Examples		
	Small Disk < 1.5 mm	Average Size Disk 1.5 – 2.0 mm	Large Disk > 2.0 mm	Small Disk	Average Size Disk	Large Disk
0a	0.5	0.4 or more	0.3 or more			
0b	0.4 up to 0.5	0.3 – 0.4	0.2 – 0.3			
1	0.3 up to 0.4	0.2 – 0.3	0.1 – 0.2			
2	0.2 up to 0.3	0.1 – 0.2	0.05 – 0.1			
3	0.1 up to 0.2	0.01 – 0.1	0.01 – 0.05			
4	0.01 – 0.1	no rim < 45 degrees	no rim < 45 degrees			
5	no rim < 45 degrees	no rim 45 – 90 degrees	no rim 45 – 90 degrees			
6	no rim 45 – 90 degrees	no rim 91 – 180 degrees	no rim 91 – 180 degrees			
7	no rim > 90 degrees	no rim > 180 degrees	no rim > 180 degrees			

Figure 6.1: The Disc Damage Likelihood Scale as originally proposed, figure by Spaeth et al. [1]

The latter is crucial as the significance of mRDR or ARW varies depending on disc size [1]. Therefore, the DDLSNet pipeline consists of two components: RimNet, which performs rim and cup segmentation and calculates mRDR or absent rim width, and DiscNet, which classifies the size of the optic disc into small, average, and large.

6.2.1 Database

Table 6.1: Glaucoma diagnoses for all 1 208 patients included in the RimNet dataset

Diagnosis	Count
Primary Open-Angle Glaucoma	530
Glaucoma Suspect	403
Chronic Angle-Closure Glaucoma	71
Low-Tension Glaucoma	47
Secondary Open-Angle Glaucoma	35
Capsular glaucoma with psuedoexfoliation	33
Anatomical Narrow Angle	27
Glaucoma secondary to Eye Infection	24
Pigmentary Glaucoma	15
Secondary Angle Closure	11
Congenital glaucoma	7
Juvenile Glaucoma	3
Acute angle-closure glaucoma	2

Our image database was based on a collection of all the optic disc photographs (ODPs) available in the UCLA Stein Eye Glaucoma Division. For the RimNet database, three glaucoma specialists manually created a mask of the optic disc rim and optic disc cup for each

fundus image using the image editing program GIMP. These masks were used as the ground truth. The RimNet dataset had two inclusion criteria. The images had to show signs of glaucomatous damage and the images had to be in focus and with discernible posterior pole and vasculature details, both as deemed by two board-certified glaucoma specialists. The exclusion criteria was concurrent non-glaucoma disease including optic neuritis, optic disc neovascularization, and vitreous hemorrhage that would impair visualization of the posterior pole. The demographic information for the RimNet dataset is presented in Table 5.1. Table 6.1 presents the glaucoma diagnoses for the RimNet dataset. These requirements result in a database that displays the full range of glaucomatous changes to the optic disc rim, ranging from mild optic disc rim narrowing in early-stage glaucoma to absent optic disc rim in severe glaucoma.

The DiscNet database consisted of optic disc photographs with available corresponding Cirrus high-definition OCT Optic Disc Cubes (200x200). The size of the Bruch’s membrane as measured by Cirrus OCT was used as a proxy for disc area and was used to categorize the disc size into small, average, or large optic discs. The optic disc photographs had to be of “good” quality—in focus with unobstructed view of the posterior pole—as determined by a board-certified glaucoma specialist. The OCT images were required to have a good quality (signal strength > 6) and be free of artifacts based on the review of printouts. To examine reliability, a database of non-progressing glaucomatous eyes was created. Each eye had two optic disc photographs available taken less than four years apart, which were deemed stable as confirmed by a glaucoma specialist. The time restriction was imposed to increase the population included but decrease the chance of glaucoma progression between the two photos.

6.2.2 RimNet

RimNet consists of a pre-processing step of contrast enhancement, an optic disc rim and cup segmentation model, and an image analysis step to calculate the mRDR for intact rims and

Table 6.2: Hyperparameter search space for RimNet

Hyperparameter Name	Possible Values
Decoders	U-Net, FPN, LinkNet, PSPnet
Loss Function	Categorical_Crossentropy, Categorical_Focal_Loss
Learning Rate	10^{-3} , 10^{-4} , 10^{-5} , 10^{-6}
Optimizer	Adam, SGD

ARW for incomplete rims. This latter case occurs in eyes with severe glaucomatous damage. The model was optimized by submitting it to a hyperparameter search with rim intersection over union as the metric. The included hyperparameters were the neural network structure, learning rate, loss function, and optimizer [58, 153, 158–160, 162–165, 167]. Table 6.2 lists the hyperparameter search space. Fifty total hyperparameter combinations were trained with the Keras Tuner library [128] with the rim segmentation proficiency, measured as an intersection-over-union, as the optimized metric. The rim segmentation model was trained, validated, and tested on a database of images from the UCLA Stein Eye Glaucoma Division with an 80/10/10 split.

6.2.3 DiscNet

DiscNet is a deep neural network developed to assign disc size as small, average, or large as an essential process in DDLS grading. The disc photographs included scanned digitized slides and digital photographs. Disc size information taken from paired OCT data was used as the ground truth. While the original DDLS grading defined small, average, and large discs as diameters of <1.50 mm, between 1.50 mm and 2.00 mm, and >2.00 mm respectively [1], we modified the cutoffs slightly to ≤ 1.44 mm, 1.44 mm to 2.28 mm, and ≥ 2.28 mm so that the three disc size categories had more evenly distributed sample sizes. This sorted our

available data into a 15/70/15 split for small, average, and large discs.

When training DiscNet, we used a transfer learning approach and instantiated our model using weights from pretraining on ImageNet [188]. We used a two phase transfer learning approach to improve performance. In the first phase, only the final layer of the model was trainable, ensuring that the steep gradients when adapting the model to an entirely new domain did not result in the weights in the pretrained model getting destroyed. Once the model’s performance stabilized, we then unlocked a number of the backbone’s layers, allowing them to be fine-tuned for the new task. The portion of the model’s layers trained is termed the “tuning fraction” of the model. We used unique learning rates in each phase of training.

Table 6.3: Hyperparameter search space for DiscNet

Hyperparameter Name	Possible Values
Backbones	InceptionV3, EfficientNetB4, EfficientNetB0, ResNet101v2, VGG16, VGG19
Phase One Learning Rate	10^{-4} , 10^{-5}
Phase Two Learning Rate	10^{-5} , 10^{-6} , 10^{-7}
Tune Fraction	0.1, 0.2, 0.5
Optimizer	Adam, SGD, RMSProp

A hyperparameter search was completed to select the optimal learning rates in both phases, the tuning fraction, the optimizer, and the network architecture. Table 6.3 lists the hyperparameter search space, which each hyperparameter was selected from. Thirty total hyperparameter combinations were trained with the Keras Tuner library with classification accuracy as the optimized metric [128].

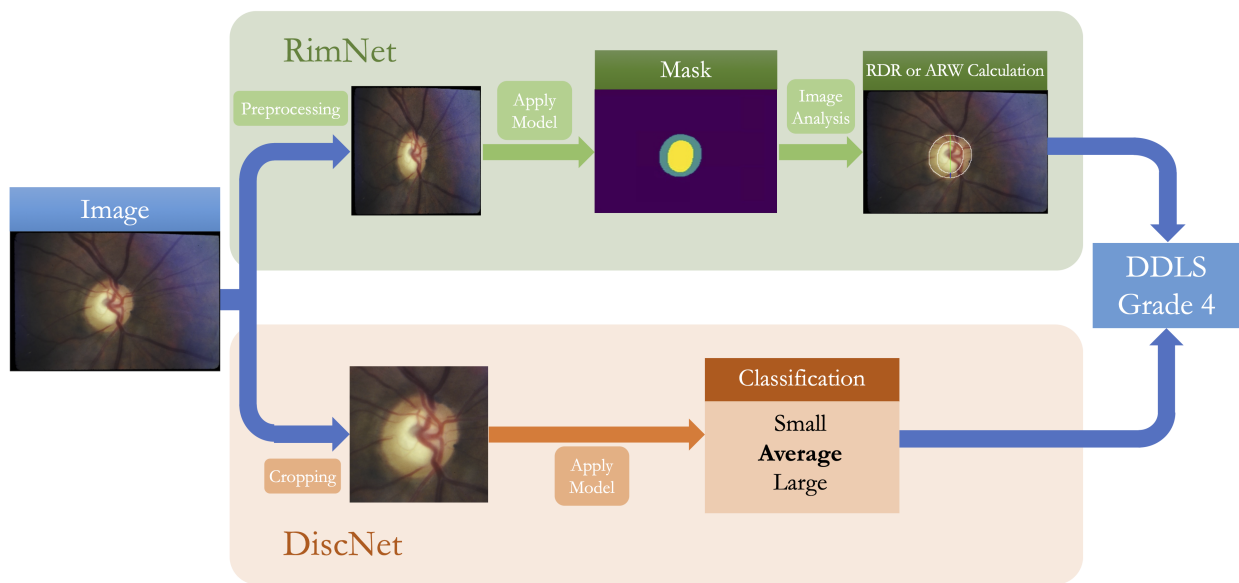


Figure 6.2: DDLSNet pipeline, illustrating both the RimNet and DiscNet arms. The calculated disc size and mRDR or ARW are used to calculate the DDLS score.

6.2.4 DDLSNet Pipeline

The mRDR and ARW from RimNet and the disc size from DiscNet were used to calculate the DDLS score. A full diagram of our pipeline is shown in Figure 6.2. DDLSNet was evaluated against a ground truth database of optic disc photographs, which three glaucoma specialists had graded with DDLS. The weighted kappa agreement ± 1 DDLS grade between the DDLSNet’s output and the average of the grades of three glaucoma specialists was measured. The average of the interobserver agreement for clinicians was also measured. The code used to train, run, and evaluate DDLSNet can be found on our public repository at <https://github.com/TylerADavis/GlaucomaML>.

Evaluating DDLSNet reliability is necessary, as physician intraobserver accuracy for DDLS grading should be matched by our proposed system for it to be clinically useful. A database of pairs of funduscopy photos of 781 non-progressing glaucomatous eyes taken within four years was used to test DDLSNet reliability. Each image was graded via DDLSNet, and the difference between the two images for each eye was recorded. Glaucoma specialists verified that the eyes were non-progressing, based on evaluation of the disc photos and the visual fields.

6.2.5 Evaluation Criteria

The main evaluation criterion was the weighted kappa agreement between DDLSNet and physicians with the ground truth database. Interobserver and intraobserver agreement was also measured as secondary evaluation criteria.

6.3 Results

RimNet was trained, validated, and tested on 1 208 optic disc photographs with an 80/10/10 split respectively. The mean age was 63.7 (± 14.9) years with a male:female ratio of 43:57.

Table 6.4: Demographic characteristics for the datasets used for RimNet, DiscNet, DDLSNet, and DDLSNet reliability

	DDLSNet Test Set	RimNet	DiscNet	DDLSNet Reliability
Total No. of Images	120	1 208	11 536	1 562
Total No. of Eyes	109	1 021	5 213	781
Gender: Male/Female	45:55	43:57	58:42	43:57
Age: Mean (SD)	65.9 (± 14.8)	63.7 (± 14.9)	67.6 (± 14.5)	73.8 (± 11.4)

Table 6.5: The DDLS distribution for our test set of 120 images, graded by glaucoma specialists

DDLS Grading by Clinician	Count
1	0
2	12
3	29
4	30
5	12
6	19
7	12
8	4
9	2
10	0
Total	120

DiscNet was trained, validated, and tested on a database of 11 536 eyes in an 80/10/10 split. The mean age was 67.6 (± 14.5) and had a male:female ratio of 58:42. DDLSNet was tested on 120 optic disc photographs from the RimNet test set manually graded based on DDLS by three glaucoma specialists. The distribution of DDLS grades in the test set is shown in Table 6.5. Reproducibility of DDLSNet was evaluated on 781 eyes, each with two optic disc photographs available (mean age=73.8 (± 11.4) years, male:female ratio=43:57). The eyes were all classified as non-progressing by a glaucoma specialist based on review of the optic disc photographs. The demographic data for the 4 cohorts are presented in Table 6.4.

6.3.1 Model Architecture and Hyperparameter Search

After exploring thirty different combinations of hyperparameters through random search, the following hyperparameters were identified as providing the highest classification accuracy for DiscNet: VGG19 architecture, phase one learning rate of 1^{-4} , phase two learning rate of 1^{-5} , tuning fraction of 0.5, and Adam optimizer [58, 153, 161–163, 189, 190]. VGG19 is a 19-layer convolutional neural network published in 2015 that has previously been used in medical image analysis [163, 191, 192]. For RimNet, fifty different combinations were examined through a random search, which resulted as follows: InceptionV3/LinkNet architecture, binary cross-entropy loss function, learning rate of 10^{-3} , and Adam optimizer [58, 153, 158–160, 162–165, 167]. InceptionV3 was first published in 2015, outperforming popular encoders at the time with a fraction of the computation costs [153]. It has been previously used in medical segmentation [154, 155]. LinkNet is a lightweight decoder first published in 2017 [159]. Given the computational restrictions of our workstation, which uses NVIDIA RTX 2080 Ti graphics cards, these were appropriate choices.

6.3.2 RimNet

The RimNet evaluation criteria were the mean absolute error (MAE) for mRDR for intact rims and the MAE for ARW for incomplete rims between physician grading and RimNet grading with a secondary evaluation criterion of the rim intersection over union (RimIoU). The intersection over union (IoU) is a commonly used measure for segmentation accuracy. RimNet achieved an mRDR MAE of 0.04 (± 0.03), a ARW MAE of 48.9 (± 35.9), and a RimIoU of 0.68.

6.3.3 DiscNet

DiscNet raw classification accuracy was found to be 73% (95% CI: 70, 75) across a test set of 1,137 images, which included both scanned slides and digitally acquired optic disc photographs. Broken down by category, DiscNet had a classification accuracy of 62% (95% CI: 55, 70) for small discs, 77% (95% CI: 74, 80) for average discs, and 60% (95% CI: 52, 68) for large discs. Notably, only three small discs out of 234 (1.2%) were mistakenly classified as large and only two large discs out of 146 (1.3%) were mistakenly classified as small.

6.3.4 DDLSNet

Table 6.6: Kappa agreement between DDLSNet and glaucoma specialist grading

Graders	Kappa (95% CI)
Grader 1 vs. Grader 2	0.52 (0.32, 0.72)
Grader 1 vs. Grader 3	0.56 (0.35, 0.77)
Grader 2 vs. Grader 3	0.49 (0.29, 0.70)
Grader Average vs. DDLSNet	0.54 (0.40, 0.68)

Table 6.7: Difference in DDLSNet grading between paired images of non-progressing optic disc photographs. All photographs were taken within four years of each other.

DDLS Difference	Number of Images
0	481
1	267
2	28
3	1

DDLSNet was evaluated on a testing database of 120 optic disc photographs. Three glaucoma specialists also graded the same 120 funduscopy images with DDLS. The weighted kappa agreement between the average grading of the three glaucoma specialists and DDLSNet was 0.54 (95% CI: 0.4, 0.68). A full breakdown of results can be found in Table 6.6. The model matched the kappa scores between physicians, which included 0.49, 0.52, and 0.56, averaged at 0.52. DDLSNet reproducibility was measured by evaluating pairs of non-progressing optic disc photographs. Of the 781 pairs of eyes, 485 (62%) had DDLS difference of 0, 267 (34%) had a DDLS difference of 1, 28 (4%) had a DDLS difference of 2, and 1 (0.1%) had a DDLS difference of 3 (Table 6.7).

6.4 Discussion

We present an automated pipeline for estimating the DDLS score with optic disc photographs in patients with suspected or established glaucoma to facilitate detection and monitoring of the disease. The DDLSNet weighted kappa agreement of 0.54 (95% CI 0.40-0.68) demonstrated moderate agreement with clinician grading and matching inter-clinician agreement. Moreover, the DDLSNet reproducibility was high with 96% of 781 non-progressing eyes found to have ± 1 DDLS grade difference on stable pairs of optic disc photographs.

Automated glaucoma grading with optic disc photographs has been evolving. Most experimental approaches focus on accurate detection of the cup-to-disc ratio with techniques ranging from thresholding to level setting to artificial intelligence models [168]. As early as 2001, Chrástek et al. offered an automated method of optic disc segmentation with filtering and edge detection, which achieved a segmentation accuracy of 71% with accuracy defined subjectively as “good” or “very good” [169]. More recently, Kumar and Bindu used U-Net [193], a segmentation neural network architecture, to achieve an intersection-over-union (IoU) of 87.9% in optic disc segmentation [194]. Our algorithm for measuring mRDR, RimNet, combines both the image processing techniques used in older segmentation studies and the artificial intelligence of newer studies to achieve a high-efficacy segmentation on a variety of optic disc photographs.

Cup-to-disc ratio has been repeatedly shown to be inferior to DDLS in grading glaucomatous damage [195]. Several papers addressed detection of the minimum optic disc rim width, an important component of calculating the DDLS score [196–198]. However, few have used automated DDLS calculation due to the complexity of the challenge. Two studies examined the results of a 3D stereographic camera (Kowa Nonmyd WX 3D, Kowa, Tokyo, Japan) [199, 200]. The camera automatically displays the DDLS grade in its final report. The study by Han et al. showed moderate agreement (weighted kappa value, 0.59) with one glaucoma specialist [200]. This study has two limitations compared to our study. First, the study only evaluates the camera against one glaucoma specialist rather than the three in our study. Second, such camera-specific software does not offer the generalizability of DDLSNet. While functional on certain cameras, such software would not offer the generalizability of DDLSNet. A third study provided clinical validation for RIA-G, an automated cloud-based optic nerve head analysis software that has been reported to be able to grade optic disc photographs based on DDLS [201]. This study showed a moderate agreement of 0.62 (0.55, 0.69) between three glaucoma specialists and the software. However, the validation set favored photographs of mild glaucoma (average DDLS grade 3, DDLS 1-7 included) and required

fundus photographs with a 30-degree field of view [201]. Our validation set has a wider spectrum of glaucomatous damage (average DDLS grade 4.5, DDLS 2-9) and DDLSNet does not require a 30-degree field of view. Moreover, the RIA-G optic disc cup and disc detection software operates based on contrast detection which would be impaired in photographs with bright artifacts and abnormal pathology⁴⁸. A fourth study implemented a partial-DDLS grading using active discs, where a circular disc shape was assumed and DDLS grades were grouped into normal, moderate, and severe categories [145]. The model achieved a category accuracy of 89% [145]. DDLSNet improves upon this study by directly comparing then ten DDLS grades rather than three categories. Additionally, our network accounts for disc size variations through DiscNet and intact and incomplete rims through RimNet. It is unclear if and to what extent the above studies included optic discs with areas of absent optic disc rim widths, which constitute the most severe DDLS grades.

DDLSNet is the most accurate and generalizable approach developed to date for several reasons. First, it was validated on optic disc photographs with a wide breadth of glaucomatous damage. This included optic disc photographs with areas of absent optic disc rims. Second, it makes no assumptions of the size or shape of the optic disc when grading size. Third, it is built on a neural network model rather than thresholding or contrast-based algorithms which are limited in learning capacity. Finally, it is not restricted to specific fundus cameras, making it more amenable for use in mobile settings where smartphones or portable fundus cameras can be used for fundus photography.

The shortcomings of our study need to be considered. Expanding the dataset could improve performance of both RimNet and DiscNet. The models will also have to be trained on images with significant concurrent pathologies, such as severe diabetic retinopathy and macular degeneration. The hyperparameter search was limited by the processing power and memory constraints of our NVIDIA RTX 2080 Ti graphics cards, which were used to train the model. A more extensive hyperparameter search can be done using larger architectures such as ResNet152 with more powerful computing hardware. Following the hyperparameter

search, the selected DiscNet model and RimNet model had the highest accuracy and rim intersection over union respectively on the validation set. However, their loss functions had evidence of possible overfitting. This would need to be addressed in future study. Finally, the number of physicians grading and segmenting funduscopy images could be increased to allow DDLSNet to learn a wider consensus of gradings.

6.5 Conclusion

In conclusion, DDLSNet offers a unique, high-efficacy, high-throughput, reliable DDLS grading system, which is well-suited to perform as a screening, diagnostic, and prognostic tool for identifying and classifying glaucomatous damage and monitoring disease progression. DDLSNet is also well-suited for mobile applications in a variety of settings, including use by individuals without extensive ophthalmological training such as a neurology resident using a phone camera attachment or optometrists seeking to better evaluate their patients' funduscopy images. Future study directions include increasing the number of physician graders and examining the implementation in remote areas with limited access. With powerful computing technology, glaucoma screening could be enhanced and widely disseminated, improving clinical outcomes for patients.

CHAPTER 7

A Twin Convolutional Neural Network for the Identification of Glaucoma Progression Using Images of the Optic Nerve Head

In this chapter we introduce a system for the detection of glaucoma progression using serial photographs of the optic nerve head. Unlike past systems which have been trained to detect progression by predicting the visual field or optical coherence tomography reading corresponding to a photo, this study uses consensus-derived clinician-generated labels as ground truth. Clinician-derived labels are used with the intention of allowing the recognition of glaucoma progression across a broad variety of phenotypes and disease severities based on the same features that a glaucoma specialist would look for. Additionally, we pair the system with an image saliency technique, allowing for insight into what portions of an image the deep learning model is attending to when generating its prediction.

We evaluate the model against an internal dataset and find that it demonstrates acceptable performance for detecting glaucoma progression, outperforming simpler automated techniques based solely on the width of the optic disc rim. Eyes correctly identified by the model demonstrated clinically relevant functional deterioration. These findings suggest that with further refinement and expansion of datasets and optimizations to the training process, deep learning has promising potential as an ancillary method for clinical decision-making regarding glaucoma progression.

7.1 Introduction

Glaucoma is a progressive optic neuropathy that can cause significant visual disability or blindness if inadequately treated [202]. Timely detection of glaucoma progression is a pressing unmet need. Appropriate remedial action can be taken and further visual loss prevented only if worsening of glaucoma is detected in a timely manner. Progressive damage to retinal ganglion cell axons at the level of lamina cribrosa is currently considered to be the main factor leading to characteristic structural changes within the optic disc [203, 204]. Various imaging modalities, such as optic disc photography, scanning laser ophthalmoscopy, and optical coherence tomography have been utilized for detection of disease deterioration in glaucoma. Among these, optic disc photography is widely available, easy to perform, and does not require sophisticated software for review; hence, it is a viable option in “low-tech” environments [205].

Serial optic disc photography is an established method to detect progressive glaucomatous damage especially in early to moderately severe disease [205–207]. However, review and comparison of serial disc photos (DPs) is time-consuming and requires extensive experience. Additionally, detection of serial change is subjective and there is high interrater variability even among seasoned glaucoma specialists [208–213]. As such, despite the low cost of optic disc photography compared to other imaging methods, and despite the decades of disc photos some patients have in their records, optic disc photography remains underutilized in the care of glaucoma patients. However, an important advantage of optic disc photography is that it has remained relevant despite past technological innovations, and will likely remain relevant into the future. As such, clinicians will always be able to compare prior disc photos to a current exam or recent photographs and make a decision on whether the disease has progressed.

Recent improvements in computing power and the availability of large clinical databases have spurred great interest in artificial intelligence approaches in healthcare [214–216]. Deep

learning in particular has seen great success in imaging tasks, achieving performance par with clinicians in various domains, including ophthalmic applications. Identification of diabetic retinopathy and clinical management of glaucoma both rely heavily on clinicians analyzing images of the eye, and the demonstrated strengths of convolutional neural networks in image classification [217] have translated into promising performance in the ophthalmic domain [218–225]. Studies on glaucoma detection with deep learning models have reported high discriminative capability [218–220, 222, 223]. Li et al. reported an area under receiver operating characteristic curve (AUC) of 0.986 for detection of glaucoma based on color fundus photographs [222]. However, despite the great success of deep learning identifying the presence of glaucoma, there are comparatively few studies that attempted to identify glaucoma progression [226–229]. Identification of glaucoma progression is a unique problem because rather than discriminating solely between healthy and glaucomatous eyes, a grader must be capable of finer discrimination, such as that between moderate and severe glaucoma. Additionally, identifying progression requires a model capable of accepting more than one image a time, and most image-based deep learning models for classification accept only single images.

The purpose of this study is to design a supervised deep learning model for detection of glaucoma progression relying on longitudinal series of disc photos. The model’s performance was investigated according to severity of glaucoma damage at baseline. Use of deep learning techniques could potentially enable non-expert clinicians as well as glaucoma specialists to be able to ascertain disease progression; we speculate that deep learning methods could exceed the performance of glaucoma experts for making this important decision.

7.2 Methods

Patients from the clinical database of the Stein Eye Institute, University of California Los Angeles (UCLA), meeting the inclusion criteria and who were seen between 1998 and 2019 were enrolled. The current study was carried out in accordance with the tenets of the dec-

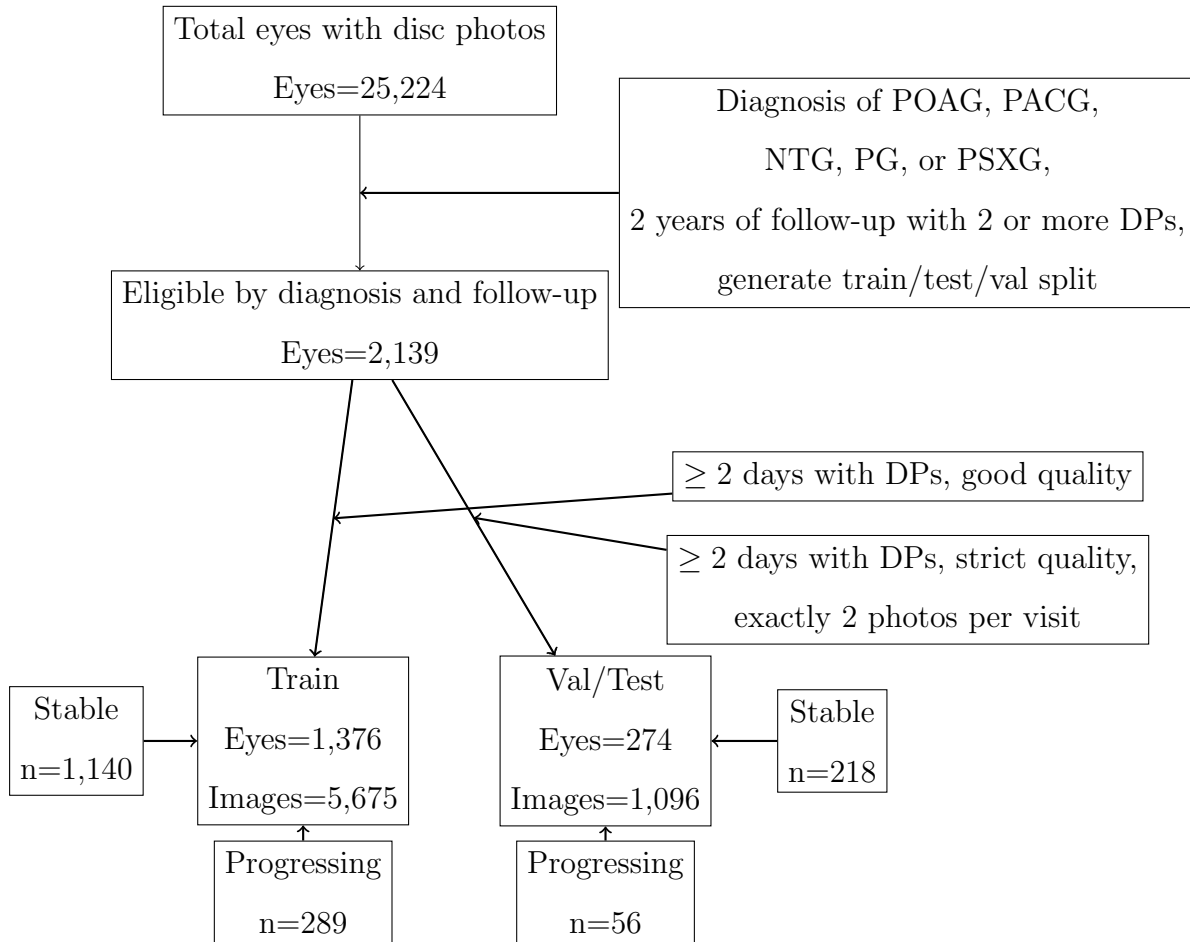


Figure 7.1: STROBE Diagram: Overview of eyes included in the study

laration of Helsinki and the Health Insurance Portability and Accountability Act (HIPAA) and was approved by UCLA’s Human Research Protection Program.

7.2.1 Dataset

The study eyes were required to have at least two years of follow-up with two or more optic disc photographs available during the follow-up period. Only patients with a diagnosis of primary open angle glaucoma (POAG), normal tension glaucoma (NTG), pigmentary glaucoma (PG), pseudoexfoliation glaucoma (PXFG), and primary angle closure glaucoma (PACG) were included. Optic disc photographs were acquired with multiple devices during

the study period including an older version of the Zeiss 450 camera (Carl Zeiss Meditec, Dublin, CA) and the FF450^{plus} Fundus Camera with VISUPACTM Digital Imaging System (Carl Zeiss Meditec, Dublin, CA). The images acquired before 2013 by the older Zeiss fundus camera were digitized prior to the study at a resolution of 4256 x 2832 pixels. The serial DPs were first evaluated by two ophthalmologists (VH and DS) for quality and those deemed to be of poor quality or with any evidence of retinal disease (such as retinal vein occlusion, diabetic retinopathy, etc.) were excluded. Standard achromatic perimetry was performed with a Humphrey Field Analyzer II using 24-2 strategy. Visual field exams with a false positive rate of <15% were included. Lastly, in order to ensure each eye in the test and validation splits was equally represented, we kept only photos in these sets from eyes where there were exactly two photos at both the baseline and final visits. Additionally, an unpublished convolutional neural network was used to reduce the likelihood of human error by identifying images that were likely to be of low quality in the test and validation splits. The restriction on number of images per visit and the second quality check were not applied to the training set in order to maximize the size of the training set and provide additional regularization. The dataset construction process is summarized in Figure 7.1.

7.2.2 Labeling Glaucoma Progressors vs Nonprogressors

Two ophthalmologists (VH and DS) were first retrained on detection of glaucoma progression using disc photos by two glaucoma specialists. VH and DS then independently labelled every (baseline, final visit) image pair as either stable or progressing. In instances where the two clinicians disagreed on presence of progression, the images were reviewed independently by two glaucoma specialists. In the instance that the two glaucoma specialists disagreed, the result was adjudicated by a third glaucoma specialist.

Table 7.1: Demographic and clinical characteristics of the study population

Variable	
Median (IQR) follow-up time (years)	10.26 (5.1-14.5)
Mean (\pm SD) baseline VF MD (dB)	-3.8 (\pm 5.2)
Percent of progressors (%)	20.2%
Baseline glaucoma severity (%)	
Mild severity at baseline	977 (79%)
Moderate severity at baseline	152 (12%)
Severe severity at baseline	113 (9)%

7.2.3 Dataset Statistics

Fourteen thousand two hundred and ninety-seven disc photos from 1,645 eyes of 916 patients were included in this study. Table 7.1 shows the demographics of the study population. Median (IQR) follow-up time was 10.26 (5.1-14.5). The mean (\pm SD) baseline 24-2 visual field Mean Deviation (MD) was -3.8 dB (\pm 5.2). The distribution of glaucoma severity at baseline was as follows: mild glaucoma (24-2 visual field MD ≥ -6 dB): 79%, moderately severe glaucoma (MD between -12 and -6 dB): 12%, and severe glaucoma (MD < -12 dB): 9%. Based on clinical review of DPs as described above, 289 eyes (20.2%) progressed during the follow-up period and 1140 eyes (79.8%) were considered as nonprogressors.

7.3 Image processing

Labeling for laterality (right versus left eye) was performed manually. In order to pass only the most informative region of the image into the model [230], we cropped the raw fundus images to squares centered around the optic disc. Square crops were chosen so that

preprocessing the images before feeding them into the model would not result in changes to the aspect ratio, thus avoiding any apparent warping of the optic disc. To identify the optic disc, we used the segmentation model first described in RimNet [13]. Our cropping algorithm identified the tightest possible square bounding box around the optic disc and cropped the raw image such that the distance from each side of the bounding box to the edge of the image was 40% of the width of bounding box. This results in the bounding square around the disc taking up 31% of the cropped image.

In order to increase the number of (baseline visit, final visit) tuples available to the model for training, we train the model using all possible pairings of images from the baseline visit and final visit. This means that if an eye had two images from the baseline visit and two images from the final visit, we generate $2 \times 2 = 4$ unique pairings of images which we feed into the model. Eyes were randomly divided into an 80/10/10 train/validation/test split before the filtering process described in Figure 7.1. In situations where both of a patient's eyes were present in the dataset, both eyes were assigned to the same split.

7.4 Development of Twin Convolutional Neural Network

In order to identify progression from pairs of color fundus images, we developed a convolutional neural network with a twin structure [231]. A twin neural network consists of two copies of the same neural network, with weights shared between them. After both inputs have been passed through the neural network, the pair of intermediate results are processed to generate a final result. This approach has been used in the past for tasks such as determining whether two signatures are the same [231], determining whether two images show the same landmark or scene regardless of viewpoint or lighting [232, 233], and generally any task where the goal is to measure the similarity of a pair of inputs [234]. This makes twin neural networks a good fit for our task, as our goal is to determine whether two images of an eye show similar glaucoma severity. The twin structure with shared weights was chosen in

order to reduce the probability of overfitting, as initial experiments without shared weights failed to demonstrate good performance on the validation set.

We chose to base our network on an EfficientNetV2B0 [235] model pretrained on ImageNet [188] as transfer learning has been shown to be beneficial in other medical imaging tasks [236] and EfficientNetV2 has been reported to outperform other models for transfer learning tasks [235]. In transfer learning, a model that was originally trained for one task, such as identifying whether an image contains a cat or a dog, is repurposed for a new task. For this study we used a two phase transfer learning approach to improve performance. In the first phase, only the final layer of the model is trainable, ensuring that the steep gradients when adapting the model to an entirely new domain do not result in the weights in the pretrained model getting destroyed. Once the model’s performance stabilized, we then unlocked a number of EfficientNetV2’s six blocks, allowing them to be fine-tuned for the new task.

To complete the forward pass of the model, we first took our square-cropped input images and resized them to 224x224 with bilinear interpolation. Our model takes in the baseline and final images and passes each separately through the EfficientNetV2 backbone. The intermediate outputs are then passed through a batch normalization layer [237] before they are concatenated together. This vector is then passed through a dropout layer [238] to help prevent overfitting before being fed into a hidden layer and finally the output layer. The architecture is illustrated in Figure 7.2.

In order to determine the optimal hyperparameters for our model, we used the Keras Tuner library [128] to perform a random search over 20 different combinations of phase one and phase two learning rates, dropout rates, portions of the model to unlock for the second phase of training, and the size of the hidden layer. The objective optimized during the random search was the area under the receiver operator characteristic curve (AUC-ROC) on the validation set. The full search space is shown in Table 7.2. All models were trained using the Adam optimizer [58], a batch size of 32, and binary cross entropy loss for 50 epochs

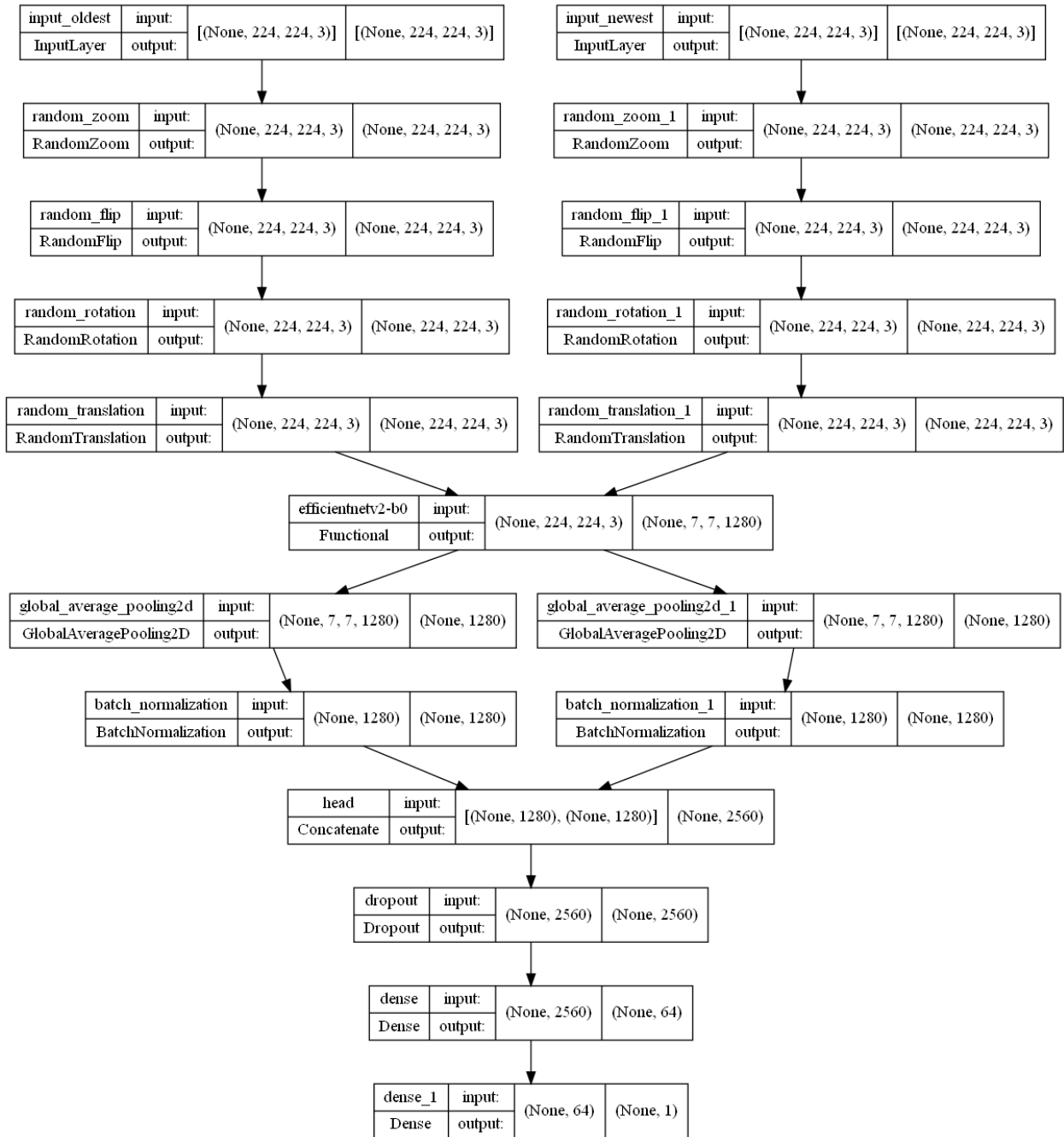


Figure 7.2: Architecture of the twin neural network

Table 7.2: Hyperparameter search space for twin neural network

Hyperparameter Name	Possible Values
Phase One Learning Rate	1^{-2} , 1^{-3} , 1^{-4}
Phase Two Learning Rate	1^{-4} , 1^{-5} , 1^{-6}
Blocks Unlocked	1, 3, 6, all layers
Dropout Rate	0.0, 0.2, 0.4, 0.6
Hidden Layer Size	0, 64, 512, 1024

in phase one and 200 epochs in phase two. Due to the imbalance between the progressing and non-progressing eyes in the dataset, we increase the weight assigned to progressing eyes when calculating loss.

Lastly, in order to further reduce the likelihood of overfitting, we applied augmentations to the images during training. Augmentations applied included random rotations of up to 36 degrees clockwise or counterclockwise, a random vertical shift of up to 10% up or down, a random horizontal shift of up to 10% left or right, a horizontal flip, a vertical flip, and a random zoom of up to 20%. Augmentations are applied independently to each image, including images belonging to the same eye. This was done to help ensure the model’s predictions are invariant to shifts in position and apparent size.

The code used to train and evaluate our model is available upon request. The model was developed with Python 3.9.7 [150]. Libraries used include TensorFlow 2.9.0 [129], Keras Tuner 1.04 [128], NumPy 1.19.5 [126], SciPy 1.7.1 [80], and scikit-learn 0.24.2 [79]. All training was performed on a single NVIDIA 2080 Ti graphics card.

7.4.1 Segmentation Model

In order to investigate the utility of attributes such as color and texture, we also investigated the performance of rim-to-disc ratio (RDR) and rim-area-to-disc-area-ratio (RADAR), two often used measures of glaucoma severity [239], for identifying progression. While RDR and RADAR are typically calculated by hand by trained glaucoma specialists, our goal was to build a high throughput automated system, and as such we calculated these measures using the deep learning segmentation model and computer vision tools described in RimNet [13]. The segmentation model accepted 224x224 images preprocessed with contrast limited adaptive histogram equalization (CLAHE).

7.5 Model Performance and Statistical Analysis

Model performance was evaluated on the test set, with AUC-ROC discriminating between progressing stable eyes as the primary metric. The ROC curve displays the trade-off between sensitivity and false positive rates ($1 - \text{specificity}$) with an AUC of 0.5 representing discrimination no better than chance and an AUC of 1 denoting perfect discrimination [240]. Additionally, we report the area under the precision-recall curve (PR-AUC), as this curve allows for the trade-off between precision and recall to be visualized, which is useful in situations such as this where the significant majority of eyes are not progressing.

We also report overall accuracy, sensitivity and specificity, and visualize model performance with a confusion matrix. These metrics were calculated using a threshold of 0.5, though the threshold can be customized to optimize for sensitivity or specificity.

Lastly, we used eXplanation with Ranked Area Integrals (XRAI) to generate saliency maps [241], allowing for clinicians to confirm that when the model correctly identifies an eye as progressing it is basing its prediction on the same features in an image that a clinician would. Our XRAI saliency maps were generated with the baseline image constant, and as such they are intended to highlight which zones of the final visit photos are most important

to the prediction. We chose to use XRAI rather than GradCAM [242] for a variety of reasons, including that XRAI’s salient regions tend to adhere more closely to the bounds of objects in the frame and that in instances where multiple relevant regions are in an image, GradCAM can tend to focus on space in between the regions rather than on the regions themselves [241]. These two attributes are particularly important for our work because we expect the model to focus on fine detail, such as the thickness of the optic disc rim or the position of a blood vessel. Additionally, markers of progression may be present in multiple regions within an image, such as in the case where broad thinning of the optic disc rim is present [243]. As such, it is important that our selected saliency method be able to adhere tightly to multiple regions of interest within a single image. We also investigated Blur IG [244], which has been used in multiple studies of the eye [244, 245]. However, we found that while Blur IG excels at identifying the small-scale pathologies in an eye with diabetic retinopathy, it is not as effective at highlighting the features indicative of glaucoma progression in an easily interpretable manner.

7.6 Results

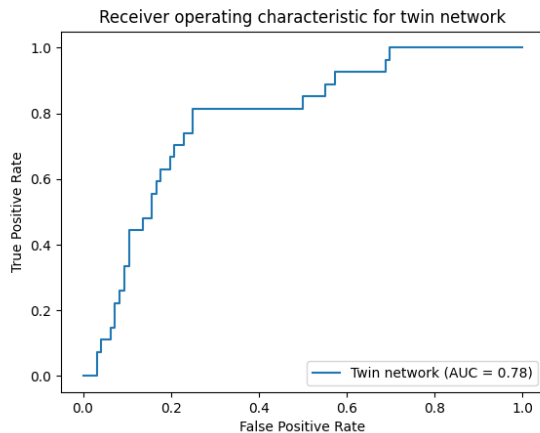


Figure 7.3: ROC curve on the test set

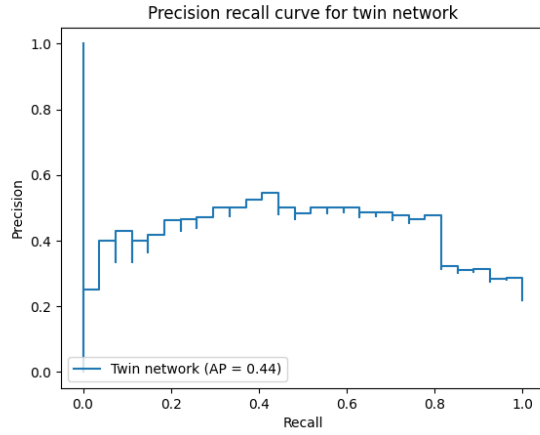


Figure 7.4: Precision recall curve on the test set

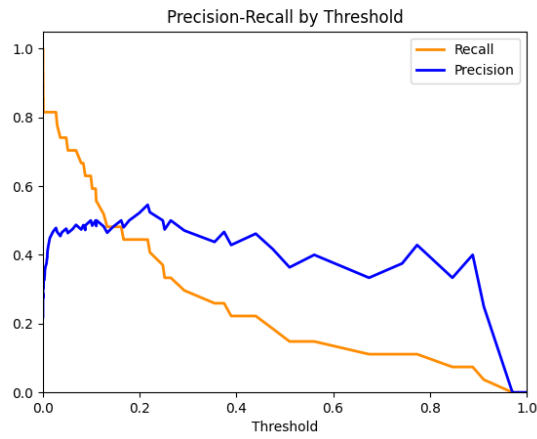


Figure 7.5: Precision and recall as a function of threshold on the test set

Table 7.3: Performance metrics for the investigated models on the test set

Model	AUC (95% CI)	PR-AUC (95% CI)
Twin Network	0.766 (± 0.0015)	0.413 (± 0.0024)
Baseline RDR	0.754 (± 0.0032)	0.424 (± 0.0055)
Baseline RADAR	0.737 (± 0.0030)	0.394 (± 0.0049)

Performance metrics obtained via 1,000 bootstraps for the three approaches investigated are shown in Table 7.3. As each eye was evaluated with four unique (baseline, final visit) image pairs, the model’s final prediction was taken the mean of all four predictions. The AUC of the deep learning model for discriminating between deteriorating and stable eyes was 0.77, shown in Figure 7.3. The RDR based model had an AUC of 0.75 and the RADAR based model had an AUC of 0.74. The PR-AUC of the twin network model was 0.435, with the curve illustrated in Figure 7.3. The effect of varying the threshold on precision and recall is shown in Figure 7.5. We then calculated the change of visual field MD from the baseline to the final visit. To this aim, available VFs within 12 months of the baseline and final DPs were included. Of the 123 eyes in the test set, 100 had matching VFs. AUC for mild (baseline MD ≥ 6 dB), moderate (baseline MD between -6 and -12 dB) and severe (baseline MD ≤ -12 dB) glaucoma were 0.81, 0.57 and 0.66, respectively.

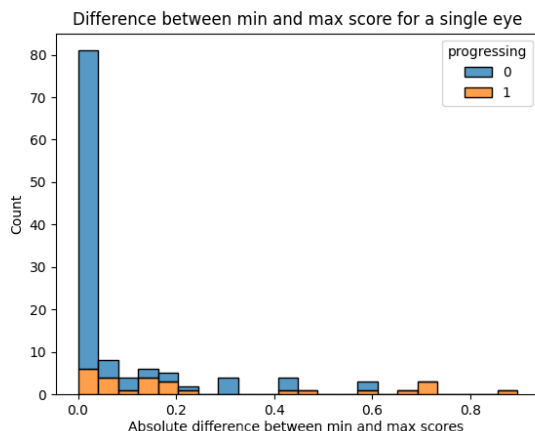


Figure 7.6: Maximum variation in scores for a single eye with twin network

In order to validate our assumption that it is valid to evaluate glaucoma progression from mismatched stereo pairs, such as a left stereo image from the baseline and a right stereo image from the final visit, we investigated the variation within the four image pairs generated for each eye. We found that the median (IQR) difference between an eye’s highest and lowest scores was 5.2×10^{-3} (0.103). The maximum score difference per eye, calculated

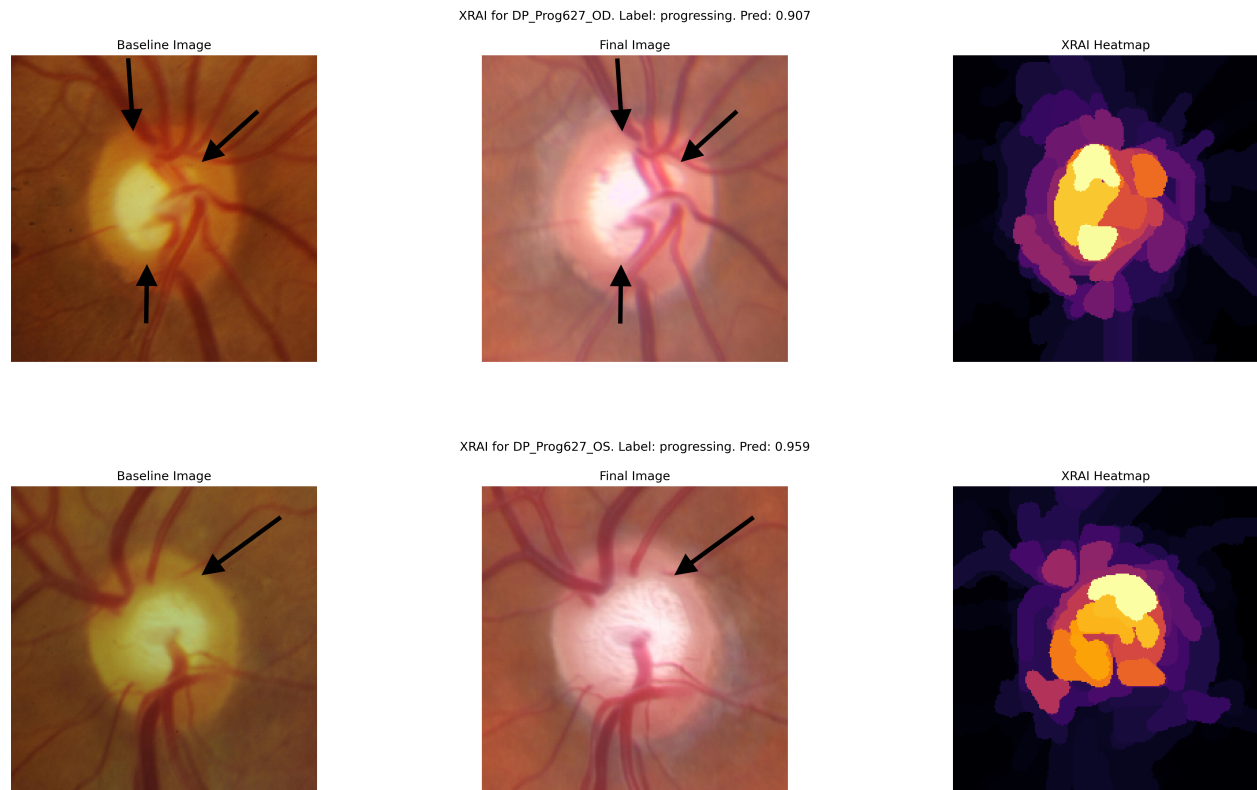


Figure 7.7: XRAI saliency maps for two eyes that were correctly classified as progressors by the deep learning model. Arrows indicate areas of rim loss.

as the difference between an eye's highest and lowest scores, is shown in a histogram in Figure 7.6.

We next inspected the saliency maps to assess whether the trained model was basing its predictions off of the same regions of the image as clinicians would. Examples of eyes classified as progressing are shown in Figure 7.7. Close inspection of the image series demonstrates that warmer colored areas of the saliency map in the right column match the area of the disc clearly showing change between the baseline and final follow-up images. In the first example (Figure 7.7, top row), discernible optic disc rim thinning (focal notch) in the 11, 2, and 6 o'clock positions can be observed at the final visit; the XRAI saliency map correctly

highlights the corresponding locations. A broad thinning of the optic disc rim developed at the 1-2:30 o'clock position during the follow-up period in the second example (bottom row) and the saliency map highlighted the same area.

7.7 Discussion

We designed and trained a deep learning model to detect structural glaucoma progression solely based on longitudinal series of optic disc photos. We used consensus derived labels from experienced clinicians as the ground truth. The final model achieved good and clinically relevant performance, distinguishing progressors with efficacy (AUC = 0.766). The kappa agreement [246] between the deep learning model and the clinician labels was 0.42 with a threshold of 0.11 (95% CI: ± 0.0027). Once validated on an external dataset, our proposed deep learning model could be deployed clinically and used as an assistive software tool to identify functionally relevant structural progression.

The twin neural network outperforming the segmentation model based approach as measured by AUC suggests that there is additional information useful for understanding glaucoma progression beyond the thickness of the optic disc rim. This is in line with recent findings that suggest that glaucoma can be identified without the use of the optic disc at all [230].

Evaluation of optic disc photos remains an essential tool for detection of structural glaucoma deterioration [247,248]. There are multiple advantages to using serial DPs for assessing glaucoma progression. First, clinically significant optic nerve head changes can often precede functional changes in eyes with ocular hypertension or early glaucoma [249,250]. Additionally, the platforms for acquiring and reviewing DPs are relatively inexpensive and have remained quite stable over a long period of time, although the quality of the fundus cameras has improved over time. However, the interobserver agreement for detection of glaucoma progression between clinicians is low [209]. Azuara-Blanco et al. reported significant in-

terobserver variability for detection of glaucoma progression with DPs [251], with a kappa value ranging from 0.34 to 0.68. Our proposed deep learning model, trained for detection of structural change on serial optic disc photos, tended to focus on the optic disc rim and the cup as the primary regions of interest, as shown by the XRAI saliency maps. The twin neural network model also agreed well with the ground truth used in this study, with a kappa value of 0.42, compared to 0.20 among clinicians in Jampel et al. [209]. One could argue that since we required agreement of two experienced glaucoma specialists on the presence or lack of progression in challenging cases, the quality of the ground truth used in this study was above any individual experienced clinician and hence the performance of the deep learning model is actually quite good. Another strength of the twin neural network model is that the series of DPs used in this study were acquired with different devices or modalities ranging from scanned slides (before 2013) to two versions of Zeiss fundus cameras providing digital DPs with 15° or 20° field of view.

Inspection of saliency maps is a very useful technique to identify specific image regions and features used by a neural network to reach a conclusion, such as presence or absence of progression in this study [252]. Therefore, saliency maps provide valuable information whether the model is focusing on the expected relevant features of the image for providing the correct classification. Thinning of optic disc rim and enlargement of the cup are the major features observed at the level of the optic nerve head with glaucoma progression [247, 253]. According to the XRAI saliency maps we generated, we could conclude that the model was focusing on changes within the optic disc rim and cup when classifying individual eyes as progressing. This finding reinforces our confidence in the proposed deep learning model with regard to assisting clinicians with the decision-making related to structural glaucoma progression.

In this study, we designed a novel CNN, which was specifically trained to provide comparison between longitudinal DPs. Although many works in this field have leveraged transfer learning previously, these models have been mainly used for classification at a single point in

time [223, 254, 255]. Therefore, we had to build a new model capable of accurately assessing change on pairs of images in order to classify glaucomatous eyes as progressing or stable. In a recent study, Medeiros et al. developed a deep learning model using transfer learning to classify glaucoma progression by predicting longitudinal changes of global retinal nerve fiber thickness (RNFL) from series of DPs with promising results [256]. Our model showed a lower AUC, with an AUC of 0.78 versus 0.86. However, while their study used machine measurements for ground truth, this study used clinical evaluation of the serial DPs and hence, our ground truth may more closely mimic the day-to-day performance of experienced clinicians managing glaucoma patients. For example, using an OCT-based ground truth may result in a model that is unable to recognize signs of progression that would not manifest on an OCT.

The reasonable and clinically relevant performance of the model (AUC = 0.784) would make this model, once externally validated, a potential candidate as an assistive tool for decision-making by optometrists or general ophthalmologists alike.

7.7.1 Limitations

At present, the techniques described in this paper have only been evaluated on the Stein Eye internal dataset. Validating against an external dataset would allow insight into the generalizability of the models to populations underrepresented in our study and to cameras besides the ones used at Stein Eye. We are currently planning to acquire data from outside our center to do this. Another limitation is that it is possible that the relatively small sample size did not allow the model to be adequately trained to detect subtle features of change in glaucoma eyes. Some glaucoma phenotypes, such as an acquired pit of the optic nerve, are relatively uncommon compared to others [243].

The performance of the techniques described in this paper were potentially limited by the approach taken to increase the number of unique training tuples. In order to increase the number of unique training tuples, the model was provided with tuples where a baseline image

was a left stereo image and the final image was a right stereo image. While this provides a degree of regularization to the system by disconnecting a change in image angle from the content of the image and the outcome label, it is possible that this may have a negative effect on test set performance, as a change in angle can alter the apparent position of structures in the eye due to parallax error. This is evidenced by the small number of eyes that had a difference of more than 0.3 between their highest and lowest scores, as shown in Figure 7.6. It is worth investigating whether better performance can be achieved by training a model on tuples where the both images come from the same side of a stereo pair, as well as whether a model can be trained with tuples consisting of stereo left and right from both the baseline and final visit, thus utilizing four images per prediction. Accomplishing this would require each image being labeled as left stereo or right stereo. Additionally, visual inspection of the dataset has shown a handful of duplicate images present with different filenames and slightly different image contents due to recompression. Removing these images from the dataset will result in having more eyes usable for evaluation, as eyes with more than two images on a given day are currently excluded.

Recent research has shown that there are statistically significant differences in glaucoma phenotypes in patients of different ages, race/ethnicity, and sex [243]. It is possible that providing such demographic information to our model may enable it to better learn how glaucoma progression manifests in different subpopulations, thus providing increased performance.

Lastly, although the dataset included multiple visits for each eye, the label of progression was only applied to the last visit in each patient's history. As the dataset does not say when exactly progression occurred between a patient's first and last visit, the twin neural network model was limited to using only images from the first and last visit. If the dataset were regraded to note exactly when in a patient's history progression occurred, an updated model could be implemented that works with the full sequence of images in a patient's history. Having an exact date on which progression occurred would also allow for the size

of the training set to be significantly expanded. Currently, additional nonprogressing pairs can be generated by pairing all images from a nonprogressing eye or by pairing images from progressing eyes taken on the same day, but there is no way to generate additional progressing pairs. Being able to generate additional image pairs beyond the $\sim 6,000$ that were used to train the model described in this paper may improve performance and generalizability.

7.8 Conclusion

In this chapter, we introduced a novel twin neural network-based system for detecting glaucoma progression based on fundus images. Eyes correctly identified by the model demonstrated clinically relevant functional deterioration. The system's accuracy is a promising first step towards the creation of an automated ancillary method for clinical evaluation of glaucoma progression. Additionally, we demonstrate that XRAI can be effectively used to evaluate optic nerve head-based models, allowing for effective debugging and the development of trust between clinicians and the model.

CHAPTER 8

Conclusion

The works described above showcase the potential for new analytical techniques to be wedded with new sensors, new information rich heterogeneous data sources, and extensive but imperfect health records in order to overcome long-standing hurdles to the adoption of artificial intelligence in medicine, including data missingness, generalizability, and heterogeneity. In overcoming these hurdles, the systems are good first steps towards gaining new insights, improving healthcare access, and aiding physicians. In summary, this dissertation described:

- HTAD, a new system that uses attention to generate target aware representations of medical record items represented as a heterogeneous network, increasing accuracy for the diagnosis prediction task.
- A novel fingertip PPG-based system for the identification of mental stress in older adults with cognitive impairment, enabling real time monitoring in the home in a way that was not previously possible.
- A large scale analysis of the CURE-CKD repository, resulting in a new model capable of predicting patients at risk for rapid kidney function decline despite high missingness. This work also resulted in new insights as to the groups most at-risk for such a decline.
- RimNet, a fully automated system for accurate segmentation of the optic disc rim, and the first study to report performance for this task on incomplete rims.
- DDLSNet, the first system for automated estimation of DDLS, enabling faster evaluation with less variability. Additionally, this is the first study to report on the problem

of determining optic disc size solely using optic disc photos without any external aids.

- The first system that evaluates pairs of images for the same markers of progression that glaucoma specialists look for. Such a system demonstrates deep learning systems may be promising add-ons for clinical decision-making regarding glaucoma progression.

8.1 Future Work

Through my work developing clinically relevant systems, I have encountered a number of challenges that I believe would be great avenues for future exploration.

First, through my collaborations with clinicians it became clear that clinicians often have access to a wealth of data, but they do not always have the expertise to leverage this data. This is especially true after the transition to EHRs and the associated massive record digitization efforts. In-house data teams and cross-disciplinary collaborations can be very valuable, but I have found that there are typically more ideas worthy of exploration than there are team members with the expertise and availability to work on them. While there may be only a handful of data analytics or machine learning experts on a team, there are typically multiple individuals comfortable with the raw data and statistics essentials. This pattern is not unique to medicine, and as such recent years have seen the rise of low-code or no-code systems for machine learning, often referred to as “AutoML”. Examples of AutoML include Google’s VertexAI and Amazon Web Service’s SageMaker platform. While an individual may not have the expertise to build a convolutional neural network that leverages transfer learning, they are more than capable of leveraging these AutoML platforms to gain insights into the datasets they already have. The models derived from these AutoML platforms could then serve as either a proof of concept to establish the feasibility of further exploring an idea, a baseline to understand the benefits of handmade models, or just a way to quickly get a system off the ground for relatively little cost. While these systems are not currently strongly marketed towards the medical community, recent work has shown that

commercial AutoML can achieve good performance in the medical domain [257]. It would be worth exploring how such AutoML platforms could be further improved for common medical tasks, such as by incorporating new medicine-specific transfer learning datasets, adding common enhancements such as contrast limited adaptive histogram equalization, providing more flexible architectures, such as systems that accept pairs or sequences of images, or by providing state-of-the-art medical segmentation models, such as some of the newest UNet derivatives [174–177].

Another clear direction for future work would be increased utilization of multimodal and multi-machine fusion when generating predictions. As described in Chapter 2, it is common to have highly heterogeneous data in a health record. This problem becomes even more difficult as technology evolves, such as when color cameras replaced black and white cameras for fundus photography, or now in modern times when new generation OCT machines generate more accurate readings and cleaner images than older machines. Clinicians need to be able to understand a patient’s condition over long periods of time, and leveraging multiple modalities from multiple generations of technology is essential to accomplishing this. While HTAD shows how time series data can be combined with an information network to generate better performance, systems capable of generating predictions leveraging all of a patient’s data, regardless of the modality or the machine that generated it, will be key to a comprehensive understanding of a patient health.

Lastly, there is still a need for tools and frameworks for model understanding and explainability. One of the most difficult parts of the project described in Chapter 4 was figuring out a way to determine what unique insights, if any, the model was generating, as well as what its predictions meant in a clinical setting. While it is great to know that a model is predicting a particular individual as being at risk for an outcome, this is not sufficient for clinicians. Clinicians want to know what populations are at risk, they want to know what type of patient they should be most concerned about, and they want to know if the model has developed insights that have not previously been reported. While there has been a

great amount of research in towards model explainability through tools like SHAP [258] and LIME [259], there is not yet a good framework that can take in a model and test dataset and automatically output something akin to a model card [260], though reporting on a model's key insights in addition to solely evaluating it for potential biases.

REFERENCES

- [1] George L. Spaeth, Jeffrey Henderer, Connie Liu, Muge Kesen, Undraa Altangerel, Atilla Bayer, L. Jay Katz, Jonathan Myers, Douglas Rhee, William Steinmann, James C. Bobrow, and Robert Ritch. The disc damage likelihood scale: reproducibility of a new method of estimating the amount of optic nerve damage caused by glaucoma. *Transactions of the American Ophthalmological Society*, 100:181, 2002.
- [2] Afolabi O. Joshua, Fulufhelo V. Nelwamondo, and Gugulethu Mabuza-Hocquet. Segmentation of optic cup and disc for diagnosis of glaucoma on retinal fundus images. pages 183–187. Institute of Electrical and Electronics Engineers Inc., 5 2019.
- [3] Julian Zilly, Joachim M. Buhmann, and Dwarikanath Mahapatra. Glaucoma detection using entropy sampling and ensemble learning for automatic optic cup and disc segmentation. *Computerized Medical Imaging and Graphics*, 55:28–41, 1 2017.
- [4] A. Sevastopolsky. Optic disc and cup segmentation methods for glaucoma detection with modification of u-net convolutional neural network. *Pattern Recognition and Image Analysis 2017 27:3*, 27:618–624, 11 2017.
- [5] Venkata Gopal Edupuganti, Akshay Chawla, and Amit Kale. Automatic optic disk and cup segmentation of fundus images using deep learning. pages 2227–2231. IEEE Computer Society, 8 2018.
- [6] Baidaa Al-Bander, Bryan M. Williams, Waleed Al-Nuaimy, Majid A. Al-Tae, Harry Pratt, and Yalin Zheng. Dense fully convolutional segmentation of the optic disc and cup in colour fundus for glaucoma diagnosis. *Symmetry*, 10, 4 2018.
- [7] Shuang Yu, Di Xiao, Shaun Frost, and Yogesan Kanagasingham. Robust optic disc and cup segmentation with deep learning for glaucoma detection. *Computerized Medical Imaging and Graphics*, 74:61–71, 6 2019.
- [8] Anahita Hosseini, Tyler Davis, and Majid Sarrafzadeh. Hierarchical target-attentive diagnosis prediction in heterogeneous information networks. In *2019 International Conference on Data Mining Workshops (ICDMW)*, pages 949–957. IEEE, 2019.
- [9] Anahita Hosseini, Ting Chen, Wenjun Wu, Yizhou Sun, and Majid Sarrafzadeh. Heteromed: Heterogeneous information network for medical diagnosis. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management, CIKM '18*, pages 763–772, New York, NY, USA, 2018. ACM.
- [10] Migyeong Gwak, Tyler Davis, Majid Sarrafzadeh, and Ellen Woo. Psychological stress detection in older adults with cognitive impairment using photoplethysmography. In *2021 IEEE 9th International Conference on Healthcare Informatics (ICHI)*, pages 209–213. IEEE, 2021.

- [11] Tyler Austin Davis, Panayiotis Petousis, Davina J Zamanzadeh, Xiaoyan Wang, Keith C Norris, Obidiugwu Duru, Katherine Tuttle, Alex Bui, Susanne B Nicholas, CURE-CKD Registry Study Team, et al. Po0528: Predicting rapid egfr decline using electronic health record (ehr) data despite high missingness in the cure-ckd registry. 2020.
- [12] Tyler Austin Davis, Panayiotis Petousis, Davina J Zamanzadeh, Keith C Norris, Obidiugwu Duru, Katherine Tuttle, Alex Bui, Susanne B Nicholas, Majid Sarrafzadeh, CURE-CKD Registry Study Team, et al. Po937: Predicting rapid egfr decline in the cure-ckd registry. 2022.
- [13] Haroon Adam Rasheed, Tyler Davis, Esteban Morales, Zhe Fei, Lourdes Grassi, Agustina De Gainza, Kouros Nouri-Mahdavi, and Joseph Caprioli. Rimnet: A deep neural network pipeline for automated identification of the optic disc rim. *Ophthalmology Science*, November 2022.
- [14] Haroon Adam Rasheed, Tyler Davis, Esteban Morales, Zhe Fei, Lourdes Grassi, Agustina De Gainza, Kouros Nouri-Mahdavi, and Joseph Caprioli. Ddlsnet: A novel deep learning-based system for grading fundusoscopic images for glaucomatous damage. *Ophthalmology Science*, 3(2):100255, 2023.
- [15] Kun-Hsing Yu, Andrew L Beam, and Isaac S Kohane. Artificial intelligence in health-care. *Nature biomedical engineering*, 2(10):719–731, 2018.
- [16] Michael Matheny, S Thadaney Israni, Mahnoor Ahmed, and Danielle Whicher. Artificial intelligence in health care: The hope, the hype, the promise, the peril. *Washington, DC: National Academy of Medicine*, 2019.
- [17] Howard L Bleich. The computer as a consultant. *New England Journal of Medicine*, 284(3):141–147, 1971.
- [18] Eta S Berner, George D Webster, Alwyn A Shugerman, James R Jackson, James Algina, Alfred L Baker, Eugene V Ball, C Glenn Cobbs, Vincent W Dennis, Eugene P Frenkel, et al. Performance of four computer-based diagnostic systems. *New England Journal of Medicine*, 330(25):1792–1796, 1994.
- [19] Peter Szolovits and Stephen G Pauker. Categorical and probabilistic reasoning in medical diagnosis. *Artificial Intelligence*, 11(1-2):115–144, 1978.
- [20] Sage Lazzaro. Machine learning’s rise, applications, and challenges, Jun 2021.
- [21] Claudine Badue, Rânik Guidolini, Raphael Vivacqua Carneiro, Pedro Azevedo, Vinicius B Cardoso, Avelino Forechi, Luan Jesus, Rodrigo Berriel, Thiago M Paixao, Filipe Mutz, et al. Self-driving cars: A survey. *Expert Systems with Applications*, 165:113816, 2021.

- [22] 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.
- [23] Mauricio Delbracio, Damien Kelly, Michael S Brown, and Peyman Milanfar. Mobile computational photography: A tour. *Annual review of vision science*, 7:571–604, 2021.
- [24] Michael E Matheny, Danielle Whicher, and Sonoo Thadaney Israni. Artificial intelligence in health care: a report from the national academy of medicine. *Jama*, 323(6):509–510, 2020.
- [25] Ziad Obermeyer and Ezekiel J Emanuel. Predicting the future—big data, machine learning, and clinical medicine. *The New England journal of medicine*, 375(13):1216, 2016.
- [26] US Department of Labor and Statistics. Cpi inflation calculator, September 2021.
- [27] Katharine Levit, Cynthia Smith, Cathy Cowan, Helen Lazenby, Art Sensenig, and Aaron Catlin. Trends in us health care spending, 2001. *Health Affairs*, 22(1):154–164, 2003.
- [28] Anne B Martin, Micah Hartman, David Lassman, Aaron Catlin, and National Health Expenditure Accounts Team. National health care spending in 2019: Steady growth for the fourth consecutive year: Study examines national health care spending for 2019. *Health Affairs*, 40(1):14–24, 2021.
- [29] Sean P Keehan, Gigi A Cuckler, John A Poisal, Andrea M Sisko, Sheila D Smith, Andrew J Madison, Kathryn E Rennie, Jacqueline A Fiore, and James C Hardesty. National health expenditure projections, 2019–28: Expected rebound in prices drives rising spending growth: National health expenditure projections for the period 2019–2028. *Health Affairs*, 39(4):704–714, 2020.
- [30] IHS Markit. The complexities of physician supply and demand: Projections from 2015 to 2030. *Assoc. Amer. Med. Colleges*, 2017.
- [31] Emily Gudbranson, Aaron Glickman, and Ezekiel J. Emanuel. Reassessing the Data on Whether a Physician Shortage Exists. *JAMA*, 317(19):1945–1946, 05 2017.
- [32] Darrell G. Kirch and Kate Petelle. Addressing the Physician Shortage: The Peril of Ignoring Demography. *JAMA*, 317(19):1947–1948, 05 2017.
- [33] United States Department of Health The Office of the National Coordinator for Health Information Technology (ONC) Office of the Secretary and Human Services. 2016 report to congress on health it progress. 2016.

- [34] United States Department of Health The Office of the National Coordinator for Health Information Technology (ONC) Office of the Secretary and Human Services (HHS). 2018 report to congress, annual update on the adoption of a nationwide system for the use and exchange of health information. 2018.
- [35] Marcus A Banks. Sizing up big data. *Nature medicine*, 26(1):5–7, 2020.
- [36] Hrayr Harutyunyan, Hrant Khachatrian, David C Kale, Greg Ver Steeg, and Aram Galstyan. Multitask learning and benchmarking with clinical time series data. *arXiv preprint arXiv:1703.07771*, 2017.
- [37] Sanjay Purushotham, Chuizheng Meng, Zhengping Che, and Yan Liu. Benchmark of deep learning models on large healthcare mimic datasets. *arXiv preprint arXiv:1710.08531*, 2017.
- [38] Edward Choi, Mohammad Taha Bahadori, Elizabeth Searles, Catherine Coffey, Michael Thompson, James Bost, Javier Tejedor-Sojo, and Jimeng Sun. Multi-layer representation learning for medical concepts. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1495–1504. ACM, 2016.
- [39] Edward Choi, Mohammad Taha Bahadori, Le Song, Walter F Stewart, and Jimeng Sun. Gram: graph-based attention model for healthcare representation learning. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 787–795. ACM, 2017.
- [40] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119, 2013.
- [41] Sheng Zhou, Jiajun Bu, Xin Wang, Jiawei Chen, Bingbing Hu, Defang Chen, and Can Wang. Hahe: Hierarchical attentive heterogeneous information network embedding. *arXiv preprint arXiv:1902.01475*, 2019.
- [42] Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Lio, and Yoshua Bengio. Graph attention networks. *arXiv preprint arXiv:1710.10903*, 2017.
- [43] Alistair EW Johnson, Tom J Pollard, Lu Shen, H Lehman Li-wei, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. MIMIC-III, a freely accessible critical care database. *Scientific data*, 3:160035, 2016.
- [44] Fenglong Ma, Radha Chitta, Jing Zhou, Quanzeng You, Tong Sun, and Jing Gao. Dipole: Diagnosis prediction in healthcare via attention-based bidirectional recurrent

- neural networks. In *Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1903–1911. ACM, 2017.
- [45] Edward Choi, Mohammad Taha Bahadori, Jimeng Sun, Joshua Kulas, Andy Schuetz, and Walter Stewart. Retain: An interpretable predictive model for healthcare using reverse time attention mechanism. In *Advances in Neural Information Processing Systems*, pages 3504–3512, 2016.
- [46] Bryan Perozzi, Rami Al-Rfou, and Steven Skiena. Deepwalk: Online learning of social representations. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 701–710. ACM, 2014.
- [47] Aditya Grover and Jure Leskovec. node2vec: Scalable feature learning for networks. In *Proceedings of the 22nd ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 855–864. ACM, 2016.
- [48] Yuxiao Dong, Nitesh V Chawla, and Ananthram Swami. metapath2vec: Scalable representation learning for heterogeneous networks. In *Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 135–144. ACM, 2017.
- [49] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*, 2014.
- [50] Yizhou Sun and Jiawei Han. Mining heterogeneous information networks: principles and methodologies. *Synthesis Lectures on Data Mining and Knowledge Discovery*, 3(2):1–159, 2012.
- [51] American Medical Association. *International classification of diseases, 9th revision, clinical modification: physician ICD-9-CM, 2005: volumes 1 and 2, color-coded, illustrated*, volume 1. Amer Medical Assn, 2004.
- [52] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017.
- [53] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.
- [54] Ting Chen and Yizhou Sun. Task-guided and path-augmented heterogeneous network embedding for author identification. In *Proceedings of the Tenth ACM International Conference on Web Search and Data Mining, WSDM '17*, pages 295–304, New York, NY, USA, 2017. ACM.
- [55] Christopher Manning, Prabhakar Raghavan, and Hinrich Schütze. Introduction to information retrieval. *Natural Language Engineering*, 16(1):100–103, 2010.

- [56] Huan Song, Deepta Rajan, Jayaraman J Thiagarajan, and Andreas Spanias. Attend and diagnose: Clinical time series analysis using attention models. In *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.
- [57] Martín Abadi, Paul Barham, Jianmin Chen, Zhifeng Chen, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Geoffrey Irving, Michael Isard, et al. Tensorflow: A system for large-scale machine learning. In *12th {USENIX} Symposium on Operating Systems Design and Implementation ({OSDI} 16)*, pages 265–283, 2016.
- [58] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.
- [59] Healthcare Cost, Utilization Project (HCUP), et al. Beta clinical classifications software (ccs) for icd-10-cm/pcs.
- [60] World Health Organization. Global health estimates 2016: Disease burden by cause, age, sex, by country and by region, 2000-2016, 2018.
- [61] Alzheimer’s Association. 2019 alzheimer’s disease facts and figures. *Alzheimer’s & Dementia*, 15(3):321–387, 2019.
- [62] James R Knickman and Emily K Snell. The 2030 problem: caring for aging baby boomers. *Health services research*, 37(4):849–884, 2002.
- [63] Stacey B Scott, Jennifer E Graham-Engeland, Christopher G Engeland, Joshua M Smyth, David M Almeida, Mindy J Katz, Richard B Lipton, Jacqueline A Mogle, Elizabeth Munoz, Nilam Ram, et al. The effects of stress on cognitive aging, physiology and emotion (escape) project. *BMC psychiatry*, 15(1):1–14, 2015.
- [64] Sonia Lupien, AndréRoch Lecours, George Schwartz, Shakti Sharma, Richard L Hauger, Michael J Meaney, and NPV Nair. Longitudinal study of basal cortisol levels in healthy elderly subjects: evidence for subgroups. *Neurobiology of Aging*, 17(1):95–105, 1996.
- [65] Marie-France Marin, Catherine Lord, Julie Andrews, Robert-Paul Juster, Shireen Sindi, Geneviève Arsenault-Lapierre, Alexandra J Fiocco, and Sonia J Lupien. Chronic stress, cognitive functioning and mental health. *Neurobiology of learning and memory*, 96(4):583–595, 2011.
- [66] Dror Ben-Zeev, Emily A Scherer, Rui Wang, Haiyi Xie, and Andrew T Campbell. Next-generation psychiatric assessment: Using smartphone sensors to monitor behavior and mental health. *Psychiatric rehabilitation journal*, 38(3):218, 2015.
- [67] Tian Hao, Kimberly N Walter, Marion J Ball, Hung-Yang Chang, Si Sun, and Xinxin Zhu. Stresshacker: towards practical stress monitoring in the wild with smartwatches. In *AMIA Annual Symposium Proceedings*, volume 2017, page 830. American Medical Informatics Association, 2017.

- [68] Yekta Said Can, Bert Arnrich, and Cem Ersoy. Stress detection in daily life scenarios using smart phones and wearable sensors: A survey. *Journal of biomedical informatics*, page 103139, 2019.
- [69] Denisse Castaneda, Aibhlin Esparza, Mohammad Ghamari, Cinna Soltanpur, and Homer Nazeran. A review on wearable photoplethysmography sensors and their potential future applications in health care. *International journal of biosensors & bioelectronics*, 4(4):195, 2018.
- [70] Lucio Ciabattoni, Francesco Ferracuti, Sauro Longhi, Lucia Pepa, Luca Romeo, and Federica Verdini. Real-time mental stress detection based on smartwatch. In *2017 IEEE International Conference on Consumer Electronics (ICCE)*, pages 110–111. IEEE, 2017.
- [71] Brigitte M Kudielka, Angelika Buske-Kirschbaum, Dirk H Hellhammer, and Clemens Kirschbaum. Differential heart rate reactivity and recovery after psychosocial stress (tsst) in healthy children, younger adults, and elderly adults: the impact of age and gender. *International journal of behavioral medicine*, 11(2):116–121, 2004.
- [72] Migyeong Gwak, Ellen Woo, and Majid Sarrafzadeh. The role of ppg in identification of mild cognitive impairment. In *Proceedings of the 12th ACM International Conference on Pervasive Technologies Related to Assistive Environments*, pages 32–35, 2019.
- [73] Jacqueline Wijsman, Bernard Grundlehner, Hao Liu, Hermie Hermens, and Julien Penders. Towards mental stress detection using wearable physiological sensors. In *2011 Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, pages 1798–1801. IEEE, 2011.
- [74] Nonin Medical Inc. Onyx® ii model 9550 finger pulse oximeter, 2019.
- [75] Dean C Delis. California verbal learning test. *Adult version. Manual. Psychological Corporation*, 2000.
- [76] Kyle Brauer Boone, Bruce L Miller, Ira M Lesser, Elizabeth Hill, and Lou D’Elia. Performance on frontal lobe tests in healthy, older individuals. *Developmental Neuropsychology*, 6(3):215–223, 1990.
- [77] Charles J Golden and Shawna M Freshwater. Stroop color and word test. 1978.
- [78] M. Duarte and R.N. Watanabe. Notes on scientific computing for biomechanics and motor control. <https://github.com/BMCLab/BMC>, 2018.
- [79] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

- [80] 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, CJ 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.
- [81] Giorgio Biagetti, Paolo Crippa, Laura Falaschetti, Simone Orcioni, and Claudio Turchetti. Motion artifact reduction in photoplethysmography using bayesian classification for physical exercise identification. In *International Conference on Pattern Recognition Applications and Methods*, volume 2, pages 467–474. SCITEPRESS, 2016.
- [82] Yifan Zhang, Shuang Song, Rik Vullings, Dwaipayan Biswas, Neide Simões-Capela, Nick Van Helleputte, Chris Van Hoof, and Willemijn Groenendaal. Motion artifact reduction for wrist-worn photoplethysmograph sensors based on different wavelengths. *Sensors*, 19(3):673, 2019.
- [83] Josef Coresh. Update on the burden of ckd. *Journal of the American Society of Nephrology*, 28(4):1020–1022, 2017.
- [84] Ron T Gansevoort and Luuk B Hilbrands. Ckd is a key risk factor for covid-19 mortality. *Nature Reviews Nephrology*, 16(12):705–706, 2020.
- [85] Andrew S Levey, Josef Coresh, Kline Bolton, Bruce Culeton, Kathy Schiro Harvey, T Alp Ikizler, Cynda Ann Johnson, Annamaria Kausz, Paul L Kimmel, John Kusek, et al. K/doqi clinical practice guidelines for chronic kidney disease: evaluation, classification, and stratification. *American Journal of Kidney Diseases*, 39(2 SUPPL. 1):i–ii+, 2002.
- [86] Kyeong Min Kim, Hyung Jung Oh, Hyung Yun Choi, Hajeong Lee, and Dong-Ryeol Ryu. Impact of chronic kidney disease on mortality: A nationwide cohort study. *Kidney research and clinical practice*, 38(3):382, 2019.
- [87] Marie Evans, Morgan E Grams, Yingying Sang, Brad C Astor, Peter J Blankestijn, Nigel J Brunskill, John F Collins, Philip A Kalra, Csaba P Kovesdy, Adeera Levin, et al. Risk factors for prognosis in patients with severely decreased gfr. *Kidney international reports*, 3(3):625–637, 2018.
- [88] Morgan E Grams, Yingying Sang, Shoshana H Ballew, Juan Jesus Carrero, Ognjenka Djurdjev, Hiddo JL Heerspink, Kevin Ho, Sadayoshi Ito, Angharad Marks, David

- Naimark, et al. Predicting timing of clinical outcomes in patients with chronic kidney disease and severely decreased glomerular filtration rate. *Kidney international*, 93(6):1442–1451, 2018.
- [89] Andrew S Levey, Kai-Uwe Eckardt, Yusuke Tsukamoto, Adeera Levin, Josef Coresh, Jerome Rossert, Dick DE Zeeuw, Thomas H Hostetter, Norbert Lameire, and Garabed Eknoyan. Definition and classification of chronic kidney disease: a position statement from kidney disease: Improving global outcomes (kdigo). *Kidney international*, 67(6):2089–2100, 2005.
- [90] Philipp Burckhardt, Daniel S Nagin, and Rema Padman. Multi-trajectory models of chronic kidney disease progression. In *AMIA Annual Symposium Proceedings*, volume 2016, page 1737. American Medical Informatics Association, 2016.
- [91] Jianyong Zhong, Hai-Chun Yang, and Agnes B Fogo. A perspective on chronic kidney disease progression. *American Journal of Physiology-Renal Physiology*, 312(3):F375–F384, 2017.
- [92] Macaulay AC Onuigbo. Syndrome of rapid-onset end-stage renal disease: a new unrecognized pattern of ckd progression to esrd. *Renal failure*, 32(8):954–958, 2010.
- [93] Liang Li, Brad C Astor, Julia Lewis, Bo Hu, Lawrence J Appel, Michael S Lipkowitz, Robert D Toto, Xuelei Wang, Jackson T Wright Jr, and Tom H Greene. Longitudinal progression trajectory of gfr among patients with ckd. *American journal of kidney diseases*, 59(4):504–512, 2012.
- [94] Hidjo J Lambers Heerspink, Hocine Tighiouart, Yingying Sang, Shoshana Ballew, Hasi Mondal, Kunihiro Matsushita, Josef Coresh, Andrew S Levey, and Lesley A Inker. Gfr decline and subsequent risk of established kidney outcomes: a meta-analysis of 37 randomized controlled trials. *American journal of kidney diseases*, 64(6):860–866, 2014.
- [95] Keith C Norris, O Kenrik Duru, Radica Z Alicic, Kenn B Daratha, Susanne B Nicholas, Sterling M McPherson, Douglas S Bell, Jenny I Shen, Cami R Jones, Tannaz Moin, et al. Rationale and design of a multicenter chronic kidney disease (ckd) and at-risk for ckd electronic health records-based registry: Cure-ckd. *BMC nephrology*, 20(1):1–9, 2019.
- [96] Katherine R Tuttle, Radica Z Alicic, O Kenrik Duru, Cami R Jones, Kenn B Daratha, Susanne B Nicholas, Sterling M McPherson, Joshua J Neumiller, Douglas S Bell, Carol M Mangione, et al. Clinical characteristics of and risk factors for chronic kidney disease among adults and children: an analysis of the cure-ckd registry. *JAMA network open*, 2(12):e1918169–e1918169, 2019.

- [97] Josef Coresh, Tanvir Chowdhury Turin, Kunihiro Matsushita, Yingying Sang, Shoshana H Ballew, Lawrence J Appel, Hisatomi Arima, Steven J Chadban, Massimo Cirillo, Ognjenka Djurdjev, et al. Decline in estimated glomerular filtration rate and subsequent risk of end-stage renal disease and mortality. *Jama*, 311(24):2518–2531, 2014.
- [98] Chava L Ramspek, Ype de Jong, Friedo W Dekker, and Merel van Diepen. Towards the best kidney failure prediction tool: a systematic review and selection aid. *Nephrology Dialysis Transplantation*, 35(9):1527–1538, 03 2019.
- [99] Justin B Echouffo-Tcheugui and Andre P Kengne. Risk models to predict chronic kidney disease and its progression: a systematic review. *PLoS medicine*, 9(11):e1001344, 2012.
- [100] Navdeep Tangri, Lesley A Stevens, John Griffith, Hocine Tighiouart, Ognjenka Djurdjev, David Naimark, Adeera Levin, and Andrew S Levey. A predictive model for progression of chronic kidney disease to kidney failure. *Jama*, 305(15):1553–1559, 2011.
- [101] Navdeep Tangri, Morgan E Grams, Andrew S Levey, Josef Coresh, Lawrence J Appel, Brad C Astor, Gabriel Chodick, Allan J Collins, Ognjenka Djurdjev, C Raina Elley, et al. Multinational assessment of accuracy of equations for predicting risk of kidney failure: a meta-analysis. *Jama*, 315(2):164–174, 2016.
- [102] Andreas Heinzl, Michael Kammer, Gert Mayer, Roman Reindl-Schwaighofer, Karin Hu, Paul Perco, Susanne Eder, Laszlo Rosivall, Patrick B Mark, Wenjun Ju, et al. Validation of plasma biomarker candidates for the prediction of egfr decline in patients with type 2 diabetes. *Diabetes care*, 41(9):1947–1954, 2018.
- [103] Chava L Ramspek, Marie Evans, Christoph Wanner, Christiane Drechsler, Nicholas C Chesnaye, Maciej Szymczak, Magdalena Krajewska, Claudia Torino, Gaetana Porto, Samantha Hayward, et al. Kidney failure prediction models: a comprehensive external validation study in patients with advanced ckd. *Journal of the American Society of Nephrology*, 32(5):1174–1186, 2021.
- [104] Kerstin Folkerts, Natalia Petruski-Ivleva, Erin Comerford, Michael Blankenburg, Thomas Evers, Alain Gay, Linda Fried, and Csaba P Kovcsdy. Adherence to chronic kidney disease screening guidelines among patients with type 2 diabetes in a us administrative claims database. In *Mayo Clinic Proceedings*, volume 96, pages 975–986. Elsevier, 2021.
- [105] Shweta Bansal, Michael Mader, and Jacqueline A Pugh. Screening and recognition of chronic kidney disease in va health care system primary care clinics. *Kidney360*, 1(9):904, 2020.

- [106] Centers for Disease Control and Prevention. Chronic kidney disease surveillance system—united states. indicator details: Percentage of patients with urine albumin laboratory results. <http://www.cdc.gov/ckd>, 2019.
- [107] Caroline S Fox, Philimon Gona, Martin G Larson, Jacob Selhub, Geoffrey Tofler, Shih-Jen Hwang, James B Meigs, Daniel Levy, Thomas J Wang, Paul F Jacques, et al. A multi-marker approach to predict incident ckd and microalbuminuria. *Journal of the American Society of Nephrology*, 21(12):2143–2149, 2010.
- [108] Kuo-Liong Chien, Hung-Ju Lin, Bai-Chin Lee, Hsiu-Ching Hsu, Yuan-Teh Lee, and Ming-Fong Chen. A prediction model for the risk of incident chronic kidney disease. *The American journal of medicine*, 123(9):836–846, 2010.
- [109] Conall M O’Seaghdha, Asya Lyass, Joseph M Massaro, James B Meigs, Josef Coresh, Ralph B D’Agostino Sr, Brad C Astor, and Caroline S Fox. A risk score for chronic kidney disease in the general population. *The American journal of medicine*, 125(3):270–277, 2012.
- [110] Meg J Jardine, Jun Hata, Mark Woodward, Vlado Perkovic, Toshiharu Ninomiya, Hisatomi Arima, Sophia Zoungas, Alan Cass, Anushka Patel, Michel Marre, et al. Prediction of kidney-related outcomes in patients with type 2 diabetes. *American journal of kidney diseases*, 60(5):770–778, 2012.
- [111] Donald B Rubin. Inference and missing data. *Biometrika*, 63(3):581–592, 1976.
- [112] Judi Scheffer. Dealing with missing data. 2002.
- [113] Petter Bjornstad, David Z Cherney, Janet K Snell-Bergeon, Laura Pyle, Marian Rewers, Richard J Johnson, and David M Maahs. Rapid gfr decline is associated with renal hyperfiltration and impaired gfr in adults with type 1 diabetes. *Nephrology Dialysis Transplantation*, 30(10):1706–1711, 2015.
- [114] Piero Ruggenenti, Esteban L Porrini, Flavio Gaspari, Nicola Motterlini, Antonio Canana, Fabiola Carrara, Claudia Cella, Silvia Ferrari, Nadia Stucchi, Aneliya Parvanova, et al. Glomerular hyperfiltration and renal disease progression in type 2 diabetes. *Diabetes care*, 35(10):2061–2068, 2012.
- [115] Gianpaolo Reboldi, Paolo Verdecchia, Gioia Fiorucci, Lawrence J Beilin, Kazuo Eguchi, Yutaka Imai, Kazuomi Kario, Takayoshi Ohkubo, Sante D Pierdomenico, Joseph E Schwartz, et al. Glomerular hyperfiltration is a predictor of adverse cardiovascular outcomes. *Kidney international*, 93(1):195–203, 2018.
- [116] Paolo Palatini, Lucio Mos, Pierferruccio Ballerini, Adriano Mazzer, Francesca Saladini, Alessandra Bortolazzi, Susanna Cozzio, Edoardo Casiglia, HARVEST Investigators, et al. Relationship between gfr and albuminuria in stage 1 hypertension. *Clinical Journal of the American Society of Nephrology*, 8(1):59–66, 2013.

- [117] Svenskt njurregister. <https://www.medscinet.net/snr/>.
- [118] Ann M O’Hare, Adam Batten, Nilka Ríos Burrows, Meda E Pavkov, Leslie Taylor, Indra Gupta, Jeff Todd-Stenberg, Charles Maynard, Rudolph A Rodriguez, Fliss EM Murtagh, et al. Trajectories of kidney function decline in the 2 years before initiation of long-term dialysis. *American Journal of Kidney Diseases*, 59(4):513–522, 2012.
- [119] Vlado Perkovic, Audrey Koitka-Weber, Mark E Cooper, Guntram Schernthaner, Egon Pfarr, Hans J Woerle, Maximilian von Eynatten, and Christoph Wanner. Choice of endpoint in kidney outcome trials: considerations from the empa-reg outcome® trial. *Nephrology Dialysis Transplantation*, 2020.
- [120] Lesley A Inker, Hiddo J Lambers Heerspink, Hasi Mondal, Christopher H Schmid, Hocine Tighiouart, Farzad Noubary, Josef Coresh, Tom Greene, and Andrew S Levey. Gfr decline as an alternative end point to kidney failure in clinical trials: a meta-analysis of treatment effects from 37 randomized trials. *American journal of kidney diseases*, 64(6):848–859, 2014.
- [121] Hiddo J Lambers Heerspink, Misghina Weldegiorgis, Lesley A Inker, Ron Gansevoort, Hans-Henrik Parving, Jamie P Dwyer, Hasi Mondal, Josef Coresh, Tom Greene, Andrew S Levey, et al. Estimated gfr decline as a surrogate end point for kidney failure: a post hoc analysis from the reduction of end points in non-insulin-dependent diabetes with the angiotensin ii antagonist losartan (renaal) study and irbesartan diabetic nephropathy trial (idnt). *American journal of kidney diseases*, 63(2):244–250, 2014.
- [122] Girish N Nadkarni, Fergus Fleming, James R McCullough, Kinsuk Chauhan, Divya A Verghese, John C He, John Quackenbush, Joseph V Bonventre, Barbara Murphy, Chirag R Parikh, et al. Prediction of rapid kidney function decline using machine learning combining blood biomarkers and electronic health record data. *BioRxiv*, page 587774, 2019.
- [123] Anima Singh, Girish Nadkarni, Omri Gottesman, Stephen B Ellis, Erwin P Bottinger, and John V Guttag. Incorporating temporal ehr data in predictive models for risk stratification of renal function deterioration. *Journal of biomedical informatics*, 53:220–228, 2015.
- [124] Rolf HH Groenwold, Ian R White, A Rogier T Donders, James R Carpenter, Douglas G Altman, and Karel GM Moons. Missing covariate data in clinical research: when and when not to use the missing-indicator method for analysis. *Cmaj*, 184(11):1265–1269, 2012.
- [125] The pandas development team. pandas-dev/pandas: Pandas, February 2020.
- [126] Charles R. Harris, K. Jarrod Millman, Stéfan J. van der Walt, Ralf Gommers, Pauli Virtanen, David Cournapeau, Eric Wieser, Julian Taylor, Sebastian Berg, Nathaniel J.

- Smith, Robert Kern, Matti Picus, Stephan Hoyer, Marten H. van Kerkwijk, Matthew Brett, Allan Haldane, Jaime Fernández del Río, Mark Wiebe, Pearu Peterson, Pierre Gérard-Marchant, Kevin Sheppard, Tyler Reddy, Warren Weckesser, Hameer Abbasi, Christoph Gohlke, and Travis E. Oliphant. Array programming with NumPy. *Nature*, 585(7825):357–362, September 2020.
- [127] Lisha Li, Kevin Jamieson, Giulia DeSalvo, Afshin Rostamizadeh, and Ameet Talwalkar. Hyperband: A novel bandit-based approach to hyperparameter optimization. *The Journal of Machine Learning Research*, 18(1):6765–6816, 2017.
- [128] Tom O’Malley, Elie Bursztein, James Long, François Chollet, Haifeng Jin, Luca Invernizzi, et al. Keras Tuner. <https://github.com/keras-team/keras-tuner>, 2019.
- [129] Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S. Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Ian Goodfellow, Andrew Harp, Geoffrey Irving, Michael Isard, Yangqing Jia, Rafal Jozefowicz, Lukasz Kaiser, Manjunath Kudlur, Josh Levenberg, Dandelion Mané, Rajat Monga, Sherry Moore, Derek Murray, Chris Olah, Mike Schuster, Jonathon Shlens, Benoit Steiner, Ilya Sutskever, Kunal Talwar, Paul Tucker, Vincent Vanhoucke, Vijay Vasudevan, Fernanda Viégas, Oriol Vinyals, Pete Warden, Martin Wattenberg, Martin Wicke, Yuan Yu, and Xiaoqiang Zheng. TensorFlow: Large-scale machine learning on heterogeneous systems, 2015. Software available from tensorflow.org.
- [130] Aaron Defazio, Francis Bach, and Simon Lacoste-Julien. Saga: A fast incremental gradient method with support for non-strongly convex composite objectives. *Advances in neural information processing systems*, 27, 2014.
- [131] Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. *CoRR*, abs/1603.02754, 2016.
- [132] Omer Sagi and Lior Rokach. Ensemble learning: A survey. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 8(4):e1249, 2018.
- [133] Yawen Xiao, Jun Wu, Zongli Lin, and Xiaodong Zhao. A deep learning-based multi-model ensemble method for cancer prediction. *Computer methods and programs in biomedicine*, 153:1–9, 2018.
- [134] Guillaume Lemaître, Fernando Nogueira, and Christos K. Aridas. Imbalanced-learn: A python toolbox to tackle the curse of imbalanced datasets in machine learning. *Journal of Machine Learning Research*, 18(17):1–5, 2017.
- [135] Frank J Massey Jr. The kolmogorov-smirnov test for goodness of fit. *Journal of the American statistical Association*, 46(253):68–78, 1951.

- [136] 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, Aditya Vijaykumar, Alessandro Pietro Bardelli, Alex Rothberg, Andreas Hilboll, Andreas Kloeckner, Anthony Scopatz, Antony Lee, Ariel Rokem, C. Nathan Woods, Chad Fulton, Charles Masson, Christian Häggström, Clark Fitzgerald, David A. Nicholson, David R. Hagen, Dmitrii V. Pasechnik, Emanuele Olivetti, Eric Martin, Eric Wieser, Fabrice Silva, Felix Lenders, Florian Wilhelm, G. Young, Gavin A. Price, Gert Ludwig Ingold, Gregory E. Allen, Gregory R. Lee, Hervé Audren, Irvin Probst, Jörg P. Dietrich, Jacob Silterra, James T. Webber, Janko Slavič, Joel Nothman, Johannes Buchner, Johannes Kulick, Johannes L. Schönberger, José Vinícius de Miranda Cardoso, Joscha Reimer, Joseph Harrington, Juan Luis Cano Rodríguez, Juan Nunez-Iglesias, Justin Kuczynski, Kevin Tritz, Martin Thoma, Matthew Newville, Matthias Kümmerer, Maximilian Bolingbroke, Michael Tartre, Mikhail Pak, Nathaniel J. Smith, Nikolai Nowaczyk, Nikolay Shebanov, Oleksandr Pavlyk, Per A. Brodtkorb, Perry Lee, Robert T. McGibbon, Roman Feldbauer, Sam Lewis, Sam Tygier, Scott Sievert, Sebastiano Vigna, Stefan Peterson, Surhud More, Tadeusz Pudlik, Takuya Oshima, Thomas J. Pingel, Thomas P. Robitaille, Thomas Spura, Thouis R. Jones, Tim Cera, Tim Leslie, Tiziano Zito, Tom Krauss, Utkarsh Upadhyay, Yaroslav O. Halchenko, and Yoshiki Vázquez-Baeza. Scipy 1.0: fundamental algorithms for scientific computing in python. *Nature Methods* 2020 17:3, 17:261–272, 2 2020.
- [137] Sture Holm. A simple sequentially rejective multiple test procedure. *Scandinavian journal of statistics*, pages 65–70, 1979.
- [138] United States Renal Data System. 2021 USRDS annual data report: Epidemiology of kidney disease in the united states. <https://adr.usrds.org/2021>, 2021.
- [139] Adeera Levin, Paul E Stevens, Rudy W Bilous, Josef Coresh, Angel LM De Francisco, Paul E De Jong, Kathryn E Griffith, Brenda R Hemmelgarn, Kunitoshi Iseki, Edmund J Lamb, et al. Kidney disease: Improving global outcomes (kdigo) ckd work group. kdigo 2012 clinical practice guideline for the evaluation and management of chronic kidney disease. *Kidney international supplements*, 3(1):1–150, 2013.
- [140] William T Cefalu, Erika Gebel Berg, Mindy Saraco, Matthew P Petersen, Sacha Uelmen, and Shamera Robinson. Microvascular complications and foot care: standards of medical care in diabetes-2019. *Diabetes Care*, 42:S124–S138, 2019.
- [141] Annette Giangiacomo and Anne Louise Coleman. The epidemiology of glaucoma. In *Glaucoma*, pages 13–21. Springer, 2009.

- [142] Georg Michelson, Joachim Hornegger, Simone Warntges, and Berthold Lausen. The papilla as screening parameter for early diagnosis of glaucoma. *Deutsches Arzteblatt International*, 105:583, 8 2008.
- [143] Yahya Shaikh, Fei Yu, and Anne L. Coleman. Burden of undetected and untreated glaucoma in the united states. *American journal of ophthalmology*, 158:1121–1129.e1, 12 2014.
- [144] M Cristina Leske, Anders Heijl, Mohamed Hussein, Bo Bengtsson, Leslie Hyman, Eugene Komaroff, Early Manifest Glaucoma Trial Group, et al. Factors for glaucoma progression and the effect of treatment: the early manifest glaucoma trial. *Archives of ophthalmology*, 121(1):48–56, 2003.
- [145] J. R. Harish Kumar, Chandra Sekhar Seelamantula, Yogish Subraya Kamath, and Rajani Jampala. Rim-to-disc ratio outperforms cup-to-disc ratio for glaucoma prescreening. *Scientific Reports 2019 9:1*, 9:1–9, 5 2019.
- [146] Paolo Formichella, Roxanne Annoh, Fabrizio Zeri, and Andrew J. Tatham. The role of the disc damage likelihood scale in glaucoma detection by community optometrists. *Ophthalmic and Physiological Optics*, 40:752–759, 11 2020.
- [147] Weihan Tong, Maryanne Romero, Vivien Lim, Seng Chee Loon, Maya E Suwandono, Yu Shuang, Xiao Di, Yogi Kanagasingam, and Victor Koh. Reliability of graders and comparison with an automated algorithm for vertical cup-disc ratio grading in fundus photographs. *Annals of the Academy of Medicine, Singapore*, 48:282–289, 9 2019.
- [148] D. R. Sarvamangala and Raghavendra V. Kulkarni. Convolutional neural networks in medical image understanding: a survey. *Evolutionary Intelligence*, 15:1–22, 1 2021.
- [149] G. Cybenko. Approximation by superpositions of a sigmoidal function. *Mathematics of Control, Signals and Systems 1989 2:4*, 2:303–314, 12 1989.
- [150] G Van Rossum and F L Drake. Python 3 reference manual; createspace. *Scotts Valley, CA*, page 242, 2009.
- [151] G. Bradski. The OpenCV Library. *Dr. Dobb’s Journal of Software Tools*, 2000.
- [152] Thomas Elsken, Jan Hendrik Metzen, and Frank Hutter. Neural architecture search: A survey. *Journal of Machine Learning Research*, 20:1–21, 2019.
- [153] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2016-December:2818–2826, 12 2015.

- [154] Muhammad Shoaib and Nasir Sayed. Yolo object detector and inception-v3 convolutional neural network for improved brain tumor segmentation. *Traitement du Signal*, 39(1), 2022.
- [155] Wessam M. Salama and Moustafa H. Aly. Deep learning in mammography images segmentation and classification: Automated cnn approach. *Alexandria Engineering Journal*, 60:4701–4709, 10 2021.
- [156] Jayanthi Sivaswamy, Arunava Chakravarty, Gopal Datt Joshi, and Tabish Abbas Syed. A comprehensive retinal image dataset for the assessment of glaucoma from the optic nerve head analysis. *JSM Biomed Imaging Data Pap*, 2:1004, 2015.
- [157] Jayanthi Sivaswamy, S. R. Krishnadas, Gopal Datt Joshi, Madhulika Jain Ujjwal, and Syed Tabish. Drishti-gs: Retinal image dataset for optic nerve head(onh) segmentation. pages 53–56. Institute of Electrical and Electronics Engineers Inc., 7 2014.
- [158] Shie Mannor, Bori Peleg, and Reuven Rubinstein. The cross entropy method for classification. *ICML 2005 - Proceedings of the 22nd International Conference on Machine Learning*, pages 561–568, 2005.
- [159] Abhishek Chaurasia and Eugenio Culurciello. Linknet: Exploiting encoder representations for efficient semantic segmentation. *2017 IEEE Visual Communications and Image Processing, VCIP 2017*, 2018-January:1–4, 6 2017.
- [160] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang Chieh Chen. Mobilenetv2: Inverted residuals and linear bottlenecks. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 4510–4520, 1 2018.
- [161] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2016-December:770–778, 12 2015.
- [162] Mingxing Tan and Quoc V. Le. Efficientnet: Rethinking model scaling for convolutional neural networks. *36th International Conference on Machine Learning, ICML 2019*, 2019-June:10691–10700, 5 2019.
- [163] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings*, 9 2014.
- [164] Weihao Weng and Xin Zhu. U-net: Convolutional networks for biomedical image segmentation. *IEEE Access*, 9:16591–16603, 5 2015.
- [165] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. 12 2016.

- [166] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. Pyramid scene parsing network. *Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017*, 2017-January:6230–6239, 12 2016.
- [167] Sebastian Ruder. An overview of gradient descent optimization algorithms. 9 2016.
- [168] Niharika Thakur and Mamta Juneja. Survey on segmentation and classification approaches of optic cup and optic disc for diagnosis of glaucoma. *Biomedical Signal Processing and Control*, 42:162–189, 4 2018.
- [169] R. Chrástek, M. Wolf, K. Donath, H. Niemann, D. Paulus, T. Hothorn, B. Lausen, R. Lämmer, C. Y. Mardin, and G. Michelson. Automated segmentation of the optic nerve head for diagnosis of glaucoma. *Medical image analysis*, 9:297–314, 2005.
- [170] J. Liu, D. W.K. Wong, J. H. Lim, X. Jia, F. Yin, H. Li, W. Xiong, and T. Y. Wong. Optic cup and disk extraction from retinal fundus images for determination of cup-to-disc ratio. *2008 3rd IEEE Conference on Industrial Electronics and Applications, ICIEA 2008*, pages 1828–1832, 2008.
- [171] Megha Lotankar, Kevin Noronha, and Jayasudha Koti. Detection of optic disc and cup from color retinal images for automated diagnosis of glaucoma. *2015 IEEE UP Section Conference on Electrical Computer and Electronics, UPCON 2015*, 4 2016.
- [172] José Martins, Jaime S. Cardoso, and Filipe Soares. Offline computer-aided diagnosis for glaucoma detection using fundus images targeted at mobile devices. *Computer Methods and Programs in Biomedicine*, 192:105341, 8 2020.
- [173] Samiksha Pachade, Prasanna Porwal, Manesh Kokare, Luca Giancardo, and Fabrice Mériaudeau. Nenet: Nested efficientnet and adversarial learning for joint optic disc and cup segmentation. *Medical Image Analysis*, 74:102253, 12 2021.
- [174] Zongwei Zhou, Md Mahfuzur Rahman Siddiquee, Nima Tajbakhsh, and Jianming Liang. Unet++: A nested u-net architecture for medical image segmentation. In *Deep learning in medical image analysis and multimodal learning for clinical decision support*, pages 3–11. Springer, 2018.
- [175] Huimin Huang, Lanfen Lin, Ruofeng Tong, Hongjie Hu, Qiaowei Zhang, Yutaro Iwamoto, Xianhua Han, Yen-Wei Chen, and Jian Wu. Unet 3+: A full-scale connected unet for medical image segmentation. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1055–1059. IEEE, 2020.
- [176] Jieneng Chen, Yongyi Lu, Qihang Yu, Xiangde Luo, Ehsan Adeli, Yan Wang, Le Lu, Alan L Yuille, and Yuyin Zhou. Transunet: Transformers make strong encoders for medical image segmentation. *arXiv preprint arXiv:2102.04306*, 2021.

- [177] Hu Cao, Yueyue Wang, Joy Chen, Dongsheng Jiang, Xiaopeng Zhang, Qi Tian, and Manning Wang. Swin-unet: Unet-like pure transformer for medical image segmentation. *arXiv preprint arXiv:2105.05537*, 2021.
- [178] Nishtha Panwar, Philemon Huang, Jiaying Lee, Pearse A. Keane, Tjin Swee Chuan, Ashutosh Richhariya, Stephen Teoh, Tock Han Lim, and Rupesh Agrawal. Fundus photography in the 21st century—a review of recent technological advances and their implications for worldwide healthcare. *Telemedicine Journal and e-Health*, 22:198, 3 2016.
- [179] Hossein Nazari Khanamiri, Austin Nakatsuka, and Jaafar El-Annan. Smartphone fundus photography. *JoVE (Journal of Visualized Experiments)*, page e55958, 7 2017.
- [180] Yih Chung Tham, Xiang Li, Tien Y. Wong, Harry A. Quigley, Tin Aung, and Ching Yu Cheng. Global prevalence of glaucoma and projections of glaucoma burden through 2040: A systematic review and meta-analysis. *Ophthalmology*, 121:2081–2090, 11 2014.
- [181] Nicholas YQ Tan, David S Friedman, Ingeborg Stalmans, Iqbal Ike K Ahmed, and Chelvin CA Sng. Glaucoma screening: where are we and where do we need to go? *Current opinion in ophthalmology*, 31(2):91–100, 2020.
- [182] Ivan Goldberg, Colin I Clement, Tina H Chiang, John G Walt, Lauren J Lee, Stuart Graham, and Paul R Healey. Assessing quality of life in patients with glaucoma using the glaucoma quality of life-15 (gql-15) questionnaire. *Journal of glaucoma*, 18(1):6–12, 2009.
- [183] Ryan L. Shelton, Woonggyu Jung, Samir I. Sayegh, Daniel T. McCormick, Jeehyun Kim, and Stephen A. Boppart. Optical coherence tomography for advanced screening in the primary care office. *Journal of Biophotonics*, 7:525–533, 7 2014.
- [184] Sanghoon Kim, Michael Crose, Will J. Eldridge, Brian Cox, William J. Brown, and Adam Wax. Design and implementation of a low-cost, portable oct system. *Biomedical Optics Express*, 9:1232, 3 2018.
- [185] Paul R. Healey and Fotis Topouzis Robert N. Weinreb. Glaucoma screening: The 5th consensus report of the world glaucoma association. 2008.
- [186] Jeffrey D Henderer, Connie Liu, Muge Kesen, Undraa Altangerel, Atilla Bayer, William C Steinmann, and George L Spaeth. Reliability of the disk damage likelihood scale. *American journal of ophthalmology*, 135(1):44–48, 2003.
- [187] Diego Ardila, Atilla P. Kiraly, Sujeeth Bharadwaj, Bokyung Choi, Joshua J. Reicher, Lily Peng, Daniel Tse, Mozziyar Etemadi, Wenxing Ye, Greg Corrado, David P. Naidich, and Shravya Shetty. End-to-end lung cancer screening with three-dimensional deep learning on low-dose chest computed tomography. *Nature Medicine 2019 25:6*, 25:954–961, 5 2019.

- [188] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. pages 248–255, 3 2010.
- [189] Geoffrey Hinton, Nitish Srivastava, and Kevin Swersky. Neural networks for machine learning lecture 6a overview of mini-batch gradient descent. *Cited on*, 14(8):2, 2012.
- [190] Ilya Sutskever, James Martens, George Dahl, and Geoffrey Hinton. On the importance of initialization and momentum in deep learning. In *International conference on machine learning*, pages 1139–1147. PMLR, 2013.
- [191] Sakshi Ahuja, B. K. Panigrahi, and Tapan Gandhi. Transfer learning based brain tumor detection and segmentation using superpixel technique. *2020 International Conference on Contemporary Computing and Applications, IC3A 2020*, pages 244–249, 2 2020.
- [192] Muhammad Mateen, Junhao Wen, Nasrullah, Sun Song, and Zhouping Huang. Fundus image classification using vgg-19 architecture with pca and svd. *Symmetry 2019, Vol. 11, Page 1*, 11:1, 12 2018.
- [193] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 9351:234–241, 5 2015.
- [194] E. Sudheer Kumar and C. Shoba Bindu. Two-stage framework for optic disc segmentation and estimation of cup-to-disc ratio using deep learning technique. *Journal of Ambient Intelligence and Humanized Computing*, 2021.
- [195] Kelvin K.W. Cheng and Andrew J. Tatham. Spotlight on the disc-damage likelihood scale (ddls). *Clinical Ophthalmology (Auckland, N.Z.)*, 15:4059, 2021.
- [196] Muhammad Salman Haleem, Liangxiu Han, Jano van Hemert, and Baihua Li. Automatic extraction of retinal features from colour retinal images for glaucoma diagnosis: a review. *Computerized medical imaging and graphics : the official journal of the Computerized Medical Imaging Society*, 37:581–596, 10 2013.
- [197] Kurnika Choudhary and Shamik Tiwari. Ann glaucoma detection using cup-to-disk ratio and neuroretinal rim. *International Journal of Computer Applications*, 111:975–8887, 2015.
- [198] Ashish Issac, M. Partha Sarathi, and Malay Kishore Dutta. An adaptive threshold based image processing technique for improved glaucoma detection and classification. *Computer methods and programs in biomedicine*, 122:229–244, 11 2015.

- [199] Premnath Gnaneswaran, Sathi Devi, Ramgopal Balu, Dhanraj Rao, Narendra Puttaiah, Rohit Shetty, and Rajesh Sasikumar. Agreement between clinical versus automated disc damage likelihood scalw (ddls) staging in asian indian eyes. *Investigative Ophthalmology & Visual Science*, 54(15):4806–4806, 2013.
- [200] Jae Wook Han, Soon Young Cho, and Kui Dong Kang. Correlation between optic nerve parameters obtained using 3d nonmydriatic retinal camera and optical coherence tomography: Interobserver agreement on the disc damage likelihood scale. *Journal of Ophthalmology*, 2014, 2014.
- [201] DIGvijay Singh, Srilathaa Gunasekaran, Maya Hada, and Varun Gogia. Clinical validation of ria-g, an automated optic nerve head analysis software. *Indian Journal of Ophthalmology*, 67:1089–1094, 7 2019.
- [202] Harry A Quigley and Aimee T Broman. The number of people with glaucoma worldwide in 2010 and 2020. *British journal of ophthalmology*, 90(3):262–267, 2006.
- [203] Harry A Quigley, Gregory R Dunkelberger, and W Richard Green. Retinal ganglion cell atrophy correlated with automated perimetry in human eyes with glaucoma. *American journal of ophthalmology*, 107(5):453–464, 1989.
- [204] Harry A Quigley, Robert W Nickells, Lisa A Kerrigan, Mary E Pease, Diane J Thibault, and Donald J Zack. Retinal ganglion cell death in experimental glaucoma and after axotomy occurs by apoptosis. *Investigative ophthalmology & visual science*, 36(5):774–786, 1995.
- [205] Navid Amini, Reza Alizadeh, Nucharee Parivisutt, EunAh Kim, Kouros Nouri-Mahdavi, and Joseph Caprioli. Optic disc image subtraction as an aid to detect glaucoma progression. *Translational vision science & technology*, 6(5):14–14, 2017.
- [206] HannaMaria Öhnel, Anders Heijl, Harald Anderson, and Boel Bengtsson. Detection of glaucoma progression by perimetry and optic disc photography at different stages of the disease: results from the early manifest glaucoma trial. *Acta Ophthalmologica*, 95(3):281–287, 2017.
- [207] J Ahn, IS Yun, HG Yoo, JJ Choi, and M Lee. Developing new automated alternation flicker using optic disc photography for the detection of glaucoma progression. *Eye*, 31(1):119–126, 2017.
- [208] Christophe Breusegem, Steffen Fieuws, Ingeborg Stalmans, and Thierry Zeyen. Agreement and accuracy of non-expert ophthalmologists in assessing glaucomatous changes in serial stereo optic disc photographs. *Ophthalmology*, 118(4):742–746, 2011.
- [209] Henry D Jampel, David Friedman, Harry Quigley, Susan Vitale, Rhonda Miller, Frederick Knezevich, and Yulan Ding. Agreement among glaucoma specialists in assessing

- progressive disc changes from photographs in open-angle glaucoma patients. *American journal of ophthalmology*, 147(1):39–44, 2009.
- [210] Rohit Varma, William C Steinmann, and Ingrid U Scott. Expert agreement in evaluating the optic disc for glaucoma. *Ophthalmology*, 99(2):215–221, 1992.
- [211] James M Tielsch, Joanne Katz, Harry A Quigley, Neil R Miller, and Alfred Sommer. Intraobserver and interobserver agreement in measurement of optic disc characteristics. *Ophthalmology*, 95(3):350–356, 1988.
- [212] Marcelo T Nicolela, Stephen M Drance, David C Broadway, Balwantray C Chauhan, Terry A McCormick, and Raymond P LeBlanc. Agreement among clinicians in the recognition of patterns of optic disk damage in glaucoma. *American journal of ophthalmology*, 132(6):836–844, 2001.
- [213] GN Shuttleworth, CH Khong, and JP Diamond. A new digital optic disc stereo camera: intraobserver and interobserver repeatability of optic disc measurements. *British journal of ophthalmology*, 84(4):403–407, 2000.
- [214] Daniel Shu Wei Ting, Louis R Pasquale, Lily Peng, John Peter Campbell, Aaron Y Lee, Rajiv Raman, Gavin Siew Wei Tan, Leopold Schmetterer, Pearse A Keane, and Tien Yin Wong. Artificial intelligence and deep learning in ophthalmology. *British Journal of Ophthalmology*, 103(2):167–175, 2019.
- [215] Abhimanyu S Ahuja. The impact of artificial intelligence in medicine on the future role of the physician. *PeerJ*, 7:e7702, 2019.
- [216] Thomas Davenport and Ravi Kalakota. The potential for artificial intelligence in healthcare. *Future healthcare journal*, 6(2):94, 2019.
- [217] Waseem Rawat and Zenghui Wang. Deep convolutional neural networks for image classification: A comprehensive review. *Neural computation*, 29(9):2352–2449, 2017.
- [218] Alessandro A Jammal, Atalie C Thompson, Eduardo B Mariottoni, Samuel I Berchuck, Carla N Urata, Tais Estrela, Susan M Wakil, Vital P Costa, and Felipe A Medeiros. Human versus machine: comparing a deep learning algorithm to human gradings for detecting glaucoma on fundus photographs. *American journal of ophthalmology*, 211:123–131, 2020.
- [219] Atalie C Thompson, Alessandro A Jammal, and Felipe A Medeiros. A deep learning algorithm to quantify neuroretinal rim loss from optic disc photographs. *American journal of ophthalmology*, 201:9–18, 2019.
- [220] Naoto Shibata, Masaki Tanito, Keita Mitsuhashi, Yuri Fujino, Masato Matsuura, Hiroshi Murata, and Ryo Asaoka. Development of a deep residual learning algorithm to screen for glaucoma from fundus photography. *Scientific reports*, 8(1):1–9, 2018.

- [221] Anirban Mitra, Priya Shankar Banerjee, Sudipta Roy, Somasis Roy, and Sanjit Kumar Setua. The region of interest localization for glaucoma analysis from retinal fundus image using deep learning. *Computer methods and programs in biomedicine*, 165:25–35, 2018.
- [222] Zhixi Li, Yifan He, Stuart Keel, Wei Meng, Robert T Chang, and Mingguang He. Efficacy of a deep learning system for detecting glaucomatous optic neuropathy based on color fundus photographs. *Ophthalmology*, 125(8):1199–1206, 2018.
- [223] Mark Christopher, Akram Belghith, Christopher Bowd, James A Proudfoot, Michael H Goldbaum, Robert N Weinreb, Christopher A Girkin, Jeffrey M Liebmann, and Linda M Zangwill. Performance of deep learning architectures and transfer learning for detecting glaucomatous optic neuropathy in fundus photographs. *Scientific reports*, 8(1):1–13, 2018.
- [224] Daniel Shu Wei Ting, Carol Yim-Lui Cheung, Gilbert Lim, Gavin Siew Wei Tan, Nguyen D Quang, Alfred Gan, Haslina Hamzah, Renata Garcia-Franco, Ian Yew San Yeo, Shu Yen Lee, et al. Development and validation of a deep learning system for diabetic retinopathy and related eye diseases using retinal images from multiethnic populations with diabetes. *Jama*, 318(22):2211–2223, 2017.
- [225] Michael David Abramoff, Yiyue Lou, Ali Erginay, Warren Clarida, Ryan Amelon, James C Folk, and Meindert Niemeijer. Improved automated detection of diabetic retinopathy on a publicly available dataset through integration of deep learning. *Investigative ophthalmology & visual science*, 57(13):5200–5206, 2016.
- [226] Atalie C Thompson, Alessandro A Jammal, and Felipe A Medeiros. A review of deep learning for screening, diagnosis, and detection of glaucoma progression. *Translational Vision Science & Technology*, 9(2):42–42, 2020.
- [227] Siamak Yousefi, Taichi Kiwaki, Yuhui Zheng, Hiroki Sugiura, Ryo Asaoka, Hiroshi Murata, Hans Lemij, and Kenji Yamanishi. Detection of longitudinal visual field progression in glaucoma using machine learning. *American journal of ophthalmology*, 193:71–79, 2018.
- [228] Siamak Yousefi, Michael H Goldbaum, Madhusudhanan Balasubramanian, Tzyy-Ping Jung, Robert N Weinreb, Felipe A Medeiros, Linda M Zangwill, Jeffrey M Liebmann, Christopher A Girkin, and Christopher Bowd. Glaucoma progression detection using structural retinal nerve fiber layer measurements and functional visual field points. *IEEE Transactions on Biomedical Engineering*, 61(4):1143–1154, 2013.
- [229] Christopher Bowd, Intae Lee, Michael H Goldbaum, Madhusudhanan Balasubramanian, Felipe A Medeiros, Linda M Zangwill, Christopher A Girkin, Jeffrey M Liebmann, and Robert N Weinreb. Predicting glaucomatous progression in glaucoma suspect eyes

- using relevance vector machine classifiers for combined structural and functional measurements. *Investigative ophthalmology & visual science*, 53(4):2382–2389, 2012.
- [230] Ruben Hemelings, Bart Elen, João Barbosa-Breda, Matthew B Blaschko, Patrick De Boever, and Ingeborg Stalmans. Deep learning on fundus images detects glaucoma beyond the optic disc. *Scientific Reports*, 11(1):1–12, 2021.
- [231] Jane Bromley, Isabelle Guyon, Yann LeCun, Eduard Säckinger, and Roopak Shah. Signature verification using a “siamese” time delay neural network. *Advances in neural information processing systems*, 6, 1993.
- [232] Iaroslav Melekhov, Juho Kannala, and Esa Rahtu. Siamese network features for image matching. In *2016 23rd international conference on pattern recognition (ICPR)*, pages 378–383. IEEE, 2016.
- [233] Xuning Liu, Yong Zhou, Jiaqi Zhao, Rui Yao, Bing Liu, and Yi Zheng. Siamese convolutional neural networks for remote sensing scene classification. *IEEE Geoscience and Remote Sensing Letters*, 16(8):1200–1204, 2019.
- [234] Davide Chicco. Siamese neural networks: An overview. *Artificial Neural Networks*, pages 73–94, 2021.
- [235] Mingxing Tan and Quoc Le. Efficientnetv2: Smaller models and faster training. In *International Conference on Machine Learning*, pages 10096–10106. PMLR, 2021.
- [236] Mohammad Amin Morid, Alireza Borjali, and Guilherme Del Fiol. A scoping review of transfer learning research on medical image analysis using imagenet. *Computers in biology and medicine*, 128:104115, 2021.
- [237] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International conference on machine learning*, pages 448–456. PMLR, 2015.
- [238] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research*, 15(1):1929–1958, 2014.
- [239] Jost B Jonas, Antonio Bergua, Paul Schmitz-Valckenberg, Konstantinos I Papatathopoulos, and Wido M Budde. Ranking of optic disc variables for detection of glaucomatous optic nerve damage. *Investigative Ophthalmology & Visual Science*, 41(7):1764–1773, 2000.
- [240] Karimollah Hajian-Tilaki. Receiver operating characteristic (roc) curve analysis for medical diagnostic test evaluation. *Caspian journal of internal medicine*, 4(2):627, 2013.

- [241] Andrei Kapishnikov, Tolga Bolukbasi, Fernanda Viégas, and Michael Terry. Xrai: Better attributions through regions. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4948–4957, 2019.
- [242] 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.
- [243] Lourdes Grassi, DIANA SALAZAR, Agustina De Gainza, Ella Bouris, Esteban Morales, and Joseph Caprioli. Phenotypic expression of the optic disc in primary open angle glaucoma. *Investigative Ophthalmology & Visual Science*, 63(7):1645–A0140, 2022.
- [244] Shawn Xu, Subhashini Venugopalan, and Mukund Sundararajan. Attribution in scale and space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9680–9689, 2020.
- [245] Ashish Bora, Siva Balasubramanian, Boris Babenko, Sunny Virmani, Subhashini Venugopalan, Akinori Mitani, Guilherme de Oliveira Marinho, Jorge Cuadros, Paisan Ruamviboonsuk, Greg S Corrado, et al. Predicting the risk of developing diabetic retinopathy using deep learning. *The Lancet Digital Health*, 3(1):e10–e19, 2021.
- [246] Mary L McHugh. Interrater reliability: the kappa statistic. *Biochemia medica*, 22(3):276–282, 2012.
- [247] Michael J Lloyd, Steven L Mansberger, Brad A Fortune, Hau Nguyen, Rodrigo Torres, Shaban Demirel, Stuart K Gardiner, Chris A Johnson, and George A Cioffi. Features of optic disc progression in patients with ocular hypertension and early glaucoma. *Journal of glaucoma*, 22(5):343, 2013.
- [248] Remo Susanna Jr and Roberto M Vessani. New findings in the evaluation of the optic disc in glaucoma diagnosis. *Current Opinion in Ophthalmology*, 18(2):122–128, 2007.
- [249] Robert N Weinreb and Peng Tee Khaw. Primary open-angle glaucoma. *The lancet*, 363(9422):1711–1720, 2004.
- [250] Michael A Kass, Dale K Heuer, Eve J Higginbotham, Chris A Johnson, John L Keltner, J Philip Miller, Richard K Parrish, M Roy Wilson, Mae O Gordon, Ocular Hypertension Treatment Study Group, et al. The ocular hypertension treatment study: a randomized trial determines that topical ocular hypotensive medication delays or prevents the onset of primary open-angle glaucoma. *Archives of ophthalmology*, 120(6):701–713, 2002.

- [251] Augusto Azuara-Blanco, L Jay Katz, George L Spaeth, Stephen A Vernon, Fiona Spencer, and Ines M Lanzl. Clinical agreement among glaucoma experts in the detection of glaucomatous changes of the optic disk using simultaneous stereoscopic photographs. *American journal of ophthalmology*, 136(5):949–950, 2003.
- [252] 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.
- [253] Jost B Jonas, Martín C Fernández, and Jörg Stürmer. Pattern of glaucomatous neuroretinal rim loss. *Ophthalmology*, 100(1):63–68, 1993.
- [254] Ryo Asaoka, Hiroshi Murata, Kazunori Hirasawa, Yuri Fujino, Masato Matsuura, Atsuya Miki, Takashi Kanamoto, Yoko Ikeda, Kazuhiko Mori, Aiko Iwase, et al. Using deep learning and transfer learning to accurately diagnose early-onset glaucoma from macular optical coherence tomography images. *American journal of ophthalmology*, 198:136–145, 2019.
- [255] Aydin Kaya, Ali Seydi Keceli, Cagatay Catal, Hamdi Yalin Yalic, Huseyin Temucin, and Bedir Tekinerdogan. Analysis of transfer learning for deep neural network based plant classification models. *Computers and electronics in agriculture*, 158:20–29, 2019.
- [256] Felipe A Medeiros, Alessandro A Jammal, and Eduardo B Mariottoni. Detection of progressive glaucomatous optic nerve damage on fundus photographs with deep learning. *Ophthalmology*, 128(3):383–392, 2021.
- [257] Edward Korot, Nikolas Pontikos, Xiaoxuan Liu, Siegfried K Wagner, Livia Faes, Josef Huemer, Konstantinos Balaskas, Alastair K Denniston, Anthony Khawaja, and Pearse A Keane. Predicting sex from retinal fundus photographs using automated deep learning. *Scientific reports*, 11(1):1–8, 2021.
- [258] Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 4765–4774. Curran Associates, Inc., 2017.
- [259] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. “why should I trust you?”: Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, August 13-17, 2016*, pages 1135–1144, 2016.
- [260] Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. Model cards for model reporting. In *Proceedings of the conference on fairness, accountability, and transparency*, pages 220–229, 2019.