Advancing Clinical Outcome Prediction through Innovative Multimodal and Domain-Generalized AI that Accommodates Limited Data

by

Elisa Villaflores Warner

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Bioinformatics)
in the University of Michigan
2024

Doctoral Committee:

Professor Arvind Rao, Chair
Professor Alla Karnovsky
Professor Nambi Nallasamy
Professor Karandeep Singh
Professor Ashok Srinivasan
Professor Ji Zhu

Elisa Villaflores Warner
elisawa@umich.edu
ORCID iD: 0000-0001-6694-2701

DEDICATION

I dedicate this thesis to my loving husband Dongxiao, who saw every phase of me during this turbulent time. Thank you for your endless support and your faith in me. I'd like to also dedicate this to my Mom, who provided a warm place to visit every Saturday as I toiled through. You won't have to deal with me being in school anymore! Lastly, I'd like to dedicate this to my Dad, who was my biggest cheerleader and comforter through school. When nobody believed in me, not even myself, Dad was out there firmly convinced that I could do anything. I know he would have been very proud of me. Rest in peace, Daddykins.

ACKNOWLEDGEMENTS

There are several key people who I would like to acknowledge. First off, I'd like to acknowledge my collaborators who have given me the chance to work on their projects and did nothing but trust and believe in me. Thank you so much for making me and my work feel valued! Special thanks to Drs. Nambi Nallasamy, Lucia Cevidanes, Najla Al-Turkestani, and Jayapalli Bapuraj. Dr. Nallasamy, in addition to serving as a committee member, provides me with unlimited knowledge of the eye and he is just as kind as he is knowledgable. Thank you Dr. Nallasamy for your mentorship and for making me as excited about eyes as you are. And thank you to Mrs. Nallasamy as well! The Nallasamys, as it turns out, are great conversationalists who have excellent culinary tastes. I also appreciate their interest in exotic birds. Dr. Cevidanes is an inspiration to me and she gave me just the right amount of push to help me reach my goals. Dr. Al-Turkestani is a sweet and intelligent woman who entrusted me with her work and encouraged me to complete my PhD. Dr. Bapuraj deserves acknowledgment for his professionalism, as he is seemingly available anytime anywhere. His input on the MRI paper was absolutely elemental and I appreciate his contribution very much.

Thank you to the committee members Drs. Alla Karnovsky, Ashok Srinivisan, Karandeep Singh, and Ji Zhu for assisting me in this journey from start to end. I'd like to acknowledge my advisor, Dr. Arvind Rao, for taking me under his wing at a critical time. He gave me the space to pursue a topic that I was passionate about and supported my wish to do an internship. He also trusted me with lab management and our website development, enabling me to learn more skills. Lastly, Professor Rao let me travel the world for science, and I really appreciate that. Additionally, I'd like to acknowledge the staff and the hard-working coordinators Kati Ellis and Julia Eussen for their kindness and patience with me during my entire PhD. I relied on them for many things.

I thank my lab members and all the senior students who inspired me. I'd especially like to thank Dr. Santhoshi Krishnan for serving as a lab whip so I stay focused on the goals. I'd also like to thank Dr. Joonsang Lee for serving as a friend and confidant for a significant portion of my PhD. I'd like to acknowledge both of their incredible abilities as scientists in this field as well.

Thank you to all of my friends for keeping me sane. Besides Dr. Krishnan and Dr. Lee, I'd like to thank my undergraduate besties Allysha Choudhury, Dr. Marina Haque and Daran He for listening to me complain about PhD life and always providing a home where I need it. Thank you to Drs. Grace Yuen and Jan Santiago for inspiration and helping me move forward in difficult times. Thank you to Yahui Luo for encouraging me endlessly and believing in me. Thank you to Winda Jeon, Jinwoong Kim and the Kim family for always being a family to me through every hard time. They always provide a home away from home for me and remind me of what is most important in life.

I want to also acknowledge my family. My father, David Warner, passed away during my PhD in March 2022, and it was earth shattering. It left a hole in my heart that will never be repaired. He was the one person in the world who I could always call about my school or life troubles, and he would comfort me and believe without any doubt that I was academically capable of whatever I put my mind to. I thank my father for his unwavering belief in me in that critical time growing up, and for keeping that faith in me until the day he died. May he be always remembered and given God's blessings.

I'd like to thank my mother, Elisa Warner, as well. I've always been proud that my mom is an engineer and I thank her for setting the example that women CAN be scientists and have a career. She is hardworking but also absolutely loving and I appreciate her for everything — not just the support she gave for my studies, but for bringing me into this world and showering me with love and understanding.

Thank you to my younger brother Alan Warner. My selfish pursuance of a PhD meant in reality some sacrifice for my brother during the time that my father was sick and I cannot end this section without acknowledging how he stepped forward to take care of my parents while I stepped back to toil away at research. He is a supportive brother too, and I appreciate that he's been so understanding of my situation.

Lastly, I want to acknowledge my most important supporter, my loving husband, Dongxiao Yan. Dongxiao served as a de facto co-advisor and set the standard for what a good scientist looks like. He is detail-oriented, reliable, professional, consistent, confident, and personable, and anyone who meets him can see that. He is also above all an understanding and supportive husband who listened to all my PhD woes, comforted me when I cried, and washed dishes and picked up food when I couldn't do my fair share around the house. I thank Dongxiao for his unwavering faith in me, much like my father, and for his belief that women can kick butt in the STEM fields just as much as men.

Thank you everyone!

PREFACE

I have been studying Bioinformatics at the University of Michigan for the last six years in pursuance of my doctorate, and in the midst of studying, met with some very personal tragedies. In my second year, an unprecedented worldwide pandemic of Covid-19 shattered standard norms for everyday life and hospitals became closed off to the world. It was in that year, 2020, that my father discovered he had congestive heart failure and a life expectancy of less than two years. After several days in the hospital where we were not allowed to visit, he was sent home to live out the rest of his term. As he was unable to do much else, he sat on the couch every day watching television and chatting with anyone who was nearby. As my father's daughter, it was very difficult for me to watch this man who I had always seen as a strong and stubborn individual become so physically weak and given up to his destiny.

In early 2022, my father could no longer breathe comfortably and we started to worry if the fate we were fearing was finally coming to pass. In the thick of this, my mother was told that because of her own personal health issues, she too would die if she did not get a kidney transplant in the coming months or undergo dialysis. Suddenly my two strong parents were under the whims of the US healthcare system and ultimately at the hands of God.

Prior to these events, I had been academically very interested in clinical decision support systems and multimodal/multidomain learning as a concept, but was relatively far removed from the actual circumstances in which they would be applied. Now, in the last two weeks of my father's life, I was coming in to the hospital every day to find my dad's room in the fourth floor patient ward and sit with him as multiple sensors connected to his fingers and his chest beeped and monitored.

Unfortunately, my father passed away before my mother's transplant. I became her supporter through it. We spent hours at the clinic, where my mother underwent scores of tests to prove that she was worthy of a transplant. There were so many tests needed to determine her qualification. I started to understand that clinicians had their own models in their head to determine how long a patient has to live, whether or not they can go home, or if they qualify for a transplant. I called these models in their heads "mental models." I started to understand that the mental models are intrinsically multimodal and multidomain as well. Clinicians need to collect as much data as possible from patients through multiple modalities and machines so they can understand a patient's needs.

After my mother's transplant, I discovered a new sense of purpose. I understood the importance of my work in a way that I hadn't before. I saw the usefulness of clinical decision support in a new light and how it had a potential to improve chaotic clinical settings, if done right. The work of my thesis is an attempt to "do clinical decision support right." It is an attempt to construct proof-of-concept multimodal and multidomain models which can make life easier for clinicians — allowing certain rare but high-quality data to assist a predictive model built on routinely-collected variables, leverage multimodal MRI to automatically detect differences between tumor growth and "fake tumor growth", and to adapt eye models for different populations with different needs.

At the end of my PhD, I look back at the years of incredibly painful events at home, and I try to think what I have learned from it. These dark times gave me a new understanding of the science that I was so passionate about before and it gave me a new reason to finish the work that I started. Well, I am proud that I stayed here and that I finished this work. I'm proud of the work that was done in this thesis. And I come out of the entire six years with a new-found appreciation of life itself.

I hope the reader can also come to appreciate the motivation behind this dissertation and the importance of overcoming these unique challenges in clinical settings. Thank you all for reading.

TABLE OF CONTENTS

LIST OF FIGURES

FIGURE

LIST OF TABLES

TABLE

LIST OF APPENDICES

LIST OF ACRONYMS

ACD Anterior Chamber Depth

ADC Apparent Diffusion Coefficient

AE Absolute Error

AL Axial Length

AI Artificial Intelligence

ANN Artificial Neural Network

AUC Area Under the Receiver Operating Curve

AUROC Area Under the Receiver Operating Characteristic Curve

BoW Bag-of-Words

CBCT Cone-beam Computed Tomography

CCT Central Corneal Thickness

CDS Clinical Decision Support

CIR Correct IOL Rate

CLIP Constrastive Language-Image Pretraining

CR Corneal Radius

ctDNA circulating tumor DNA

CT Computed Tomography

CNN Convolutional Neural Network

DC/TMD Diagnostic Criteria for Temporomandibular Disorders

DR Diabetic Retinopathy

DWI Diffusion Weighted Imaging

ECG Electrocardiogram

FDA Food and Drug Administration

FLAIR Fluid-Attentuated Inversion Recovery

FPI Formula Performance Index

GAN Generative Adversarial Network

GCN Graph Convolutional Network

GWR Geographically-Weighted Regression

HGG High-Grade Glioma

HITECH Health Information Technology for Economic and Clinical Health Act

ICE Individual Component Expectation

IOL Intraocular Lens

K1 Keratometry of the Left Eye

K2 Keratometry of the Right Eye

LOOCV Leave-One-Out Cross Validation

LT Lens Thickness

MAE Mean Absolute Error

MAEPI Mean Absolute Error in Prediction of Intraocular Lens

ME Mean Error

MedAE Median Absolute Error

ML Machine Learning

NPV Negative Predictive Value

MRI Magnetic Resonance Imaging

OOB Out-of-Bag

NCI National Cancer Institute

NLP Natural Language Processing

PET Positron Emission Tomography

PHI Patient Health Information

PMMA Polymethyl methacrylate

PPG Photoplethysmogram

PPV Positive Predictive Value

PCA Principal Component Analysis

PsP Pseudoprogression

PWI Perfusion Weighted Imaging

RCP Radius of Peripheral Cornea

RF Random Forest

RMSE Root Mean Squared Error

SIFT Shift-Invariant Feature Transforms

STD Standard Deviation

SVM Support Vector Machine

SVM$^+$ Support Vector Machine for Privileged Learning

T1 T1-weighted

T2 T2-weighted

TCIA The Cancer Imaging Archive

TCGA The Cancer Genome Atlas

TMD Temporomandibular Disorders

TMJ Temporomandibular Joint

TMJ OA Temporomandibular Joint Osteoarthritis

TP True Progression

TRIPOD Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis

ULIB User Group for Laser Interference Biometry

WHO World Health Organization

WTW White-to-white

ABSTRACT

Clinical decision support systems are computer-based systems developed with the goal of assisting health care providers in arduous clinical tasks or improving decision-making. In routine clinical care, medical practices tend to be dynamic and must account for diversity of data. In this thesis, we focus on developing innovative multimodal and multidomain AI models for clinical decision support, with a focus on applications with limited data availability. We start with a survey chapter followed by three case studies of multimodal/multidomain proof-of-concept Clinical Decision Support (CDS) models that accommodate limited data. Our research seeks to address questions regarding constructing machine-learning-based models that mimic real-world mental models and bridge domain gaps in cases of limited data. In the first chapter, we explore a survey of state-of-the-art methods in multimodal machine learning applied to biomedicine, highlighting how these models address five challenges of multimodal machine learning: representation, fusion, translation, alignment and co-learning. Next, we tackle a case study where we develop a low-parameter model to discriminate pseudoprogression and true progression in glioblastoma using a small sample of MRI images. Then, we develop a clinically-informed privileged learning model which leverages both routine clinical data and privileged information (CBCT and protein serum/saliva tests) to detect Temporomandibular Joint Osteoarthritis (TMJ OA). Finally, we present a case of domain generalization to allow a model trained on one Alcon SN60WF lens to predict post-operative refraction in patients implanted with other lenses in cataract surgery, with an attempt to adapt to other populations and "A-constants" as well. We present these three case studies as examples of informed models that accommodate diverse data types, as real-world clinical practice is intrinsically multimodal and multidomain. We hope these models provide inspiration for additional models outside of the provided use cases and assert that methodologies can be combined and adapted as needed.

CHAPTER 1

Introduction

1.1 Clinical Decision Support Systems

Clinical decision support systems are computer-based systems developed with the goal of assisting health care providers in arduous clinical tasks or improving decision-making. They are also tools for implementing "precision medicine," which aims to understand a patient's specific health profile in order to prescribe a unique treatment customized to patient needs. CDS systems are on the rise, with financial support from the US Health and Medicare acts to implement them with electronic health records [184]. Studies show that CDS provides positive clinical benefit in prescribing treatments and facilitating preventive care services, among other things [26]. Accordingly, CDS has received growing interest in both private and public ventures.

In order to be useful, CDS systems must be developed to be adaptive and robust. This is because in routine clinical care, medical practices tend to be dynamic and must account for diversity of data. Healthcare providers depend on their own "mental models" shaped by years of education and practice to reach clinical decisions such as diagnosis or prognosis of a patient. These mental models typically involve interpreting a patient's clinical chart, discussing symptoms and medical history with the patient, and ordering tests or procedures as needed such as medical imaging, blood, or urine tests. All of this information is gathered because it is intrinsically understood that the inclusion of a wide variety of data is necessary for a holistic understanding of a patient's body and their physical health.

Moreover, health care provider mental models are also trained to be adaptive to changes in data caused by use of different instruments, workflows, or changes to patient demographic. For instance, patient demographic adjustments to mental models for diagnosing kidney failure may require recalibration when working in environments with more African Americans, because estimated glomerular filtration rates, which measure kidney function, are so different for African Americans that they are assessed with their own race-specific metric [44].

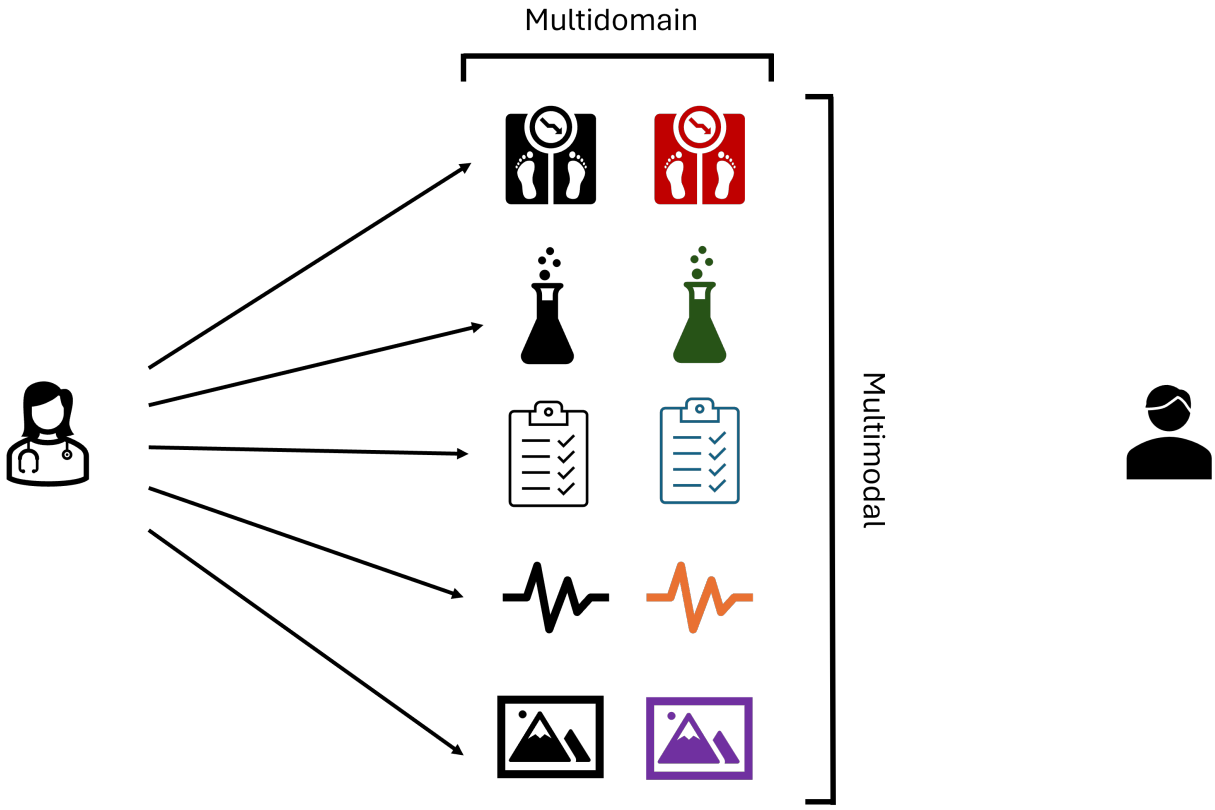

Figure 1.1: In a dynamic clinical environment, health care providers build mental models constructed of diverse data inputs to understand a patients physical health holistically.

Instrument and workflow adjustments for mental models may be required as hospitals and healthcare centers continue to upgrade equipment and services. However, computer-based CDS models tend to struggle in this department due to a lack of data variability. In model training, this results in "batch effects", whereby patterns emerge in a dataset that are attributable to non-biological factors such as workflow or instrumentation rather than true biological patterns. While including more data is desirable to mitigate bias due to batch effects, this is often not economically feasible to attain. However, the rise in publicly-available datasets offer potential solutions to these challenges.

Despite the challenges, there is merit to building models that accommodate diverse data, even if it is difficult to execute. Diverse data sources enrich clinical decision-making in myriad ways. Multimodal data, which can involve an amalgamation of imaging, patient history, and lab tests, among many other kinds of data, offer a more holistic scope of patient health to drive decision making, while multidomain data enable versatility in models so that decision-making can extend to data from which models weren't explicitly trained, even if data collection methods or distributions are not exactly the same.

Artificial Intelligence (AI), powered by machine learning (ML) models, has been essential to developing rapid and effective CDS solutions. These systems can leverage diverse types of data to enrich decision-making, considering information captured from various modalities such as medical imaging or lab tests. Artificial Intelligence (AI) and Machine Learning (ML) development in the biomedical sector is due in large part to improvements in image processing due to the rise of big data. A small summary of the history of image processing and the rise of big data is given in Appendix A.

1.2 Workflow of ML-based CDS Models

While this work presents a diverse collection of ML-based CDS models, we begin by briefly discussing general ML-based CDS workflows, as they can be summarized in four distinct stages. Note that in many workflows there are minor strategies to mitigate bias due to lack of data diversity. While there may be several additional details involved in each step, we present the following general workflow anatomy:

1. **Data acquisition**

   In machine learning/deep learning, data acquisition is one of the most critical steps. The purpose of data acquisition is to find or create datasets that can be used to build models. Data acquisition starts at the study design phase, relying on proper record-keeping, unbiased reporting, and fairly complete records. Acquisition can also involve data harmonization, which attempts to integrate different datasets into a single set. The amount and quality of medical input data for artificial intelligence applications are critical factors which influence a model's performance, sometimes implicitly. Even if a model performs well in metrics such as accuracy or Area Under the Receiver Operating Characteristic Curve (AUROC) (sometimes abbreviated as Area Under the Receiver Operating Curve (AUC)), the model may still perform poorly in real-life clinical settings if the data the model was trained on was significantly biased or based on poor study design.

   If working with images, additional methods such as augmentation of images can be implemented once the data are obtained. This is very useful in increasing the diversity of data available for training models. Image data augmentation such as cropping, padding, rotation, transformation, or horizontal flipping are commonly used to artificially expand the size of a training dataset to build models. Data augmentation methods also can be considered as a preprocessing method. However, any synthetically-created perturbations should match the physics of the system at hand, as changing the

brightness of a picture of a dog doesn't make it not a dog but changing the intensity on a Computed Tomography (CT) scan would change a structure from muscle to bone. These adversarial perturbations can also be performed after training a model to assess the robustness and quality of that model.

Finally, data acquisition involves ground truth labeling. In order to train supervised models, a training set of labeled data must be provided to the machine. Therefore, data must be labeled by an expert based on the desired outcome. For example, in segmentation problems for tumor detection, a radiologist can be asked to designate the location of a tumor in an image. In discrimination problems, a physician may be asked to look at lab vitals, biomarker information, or images, and determine whether or not a disease is present in a given patient. It is common practice to assign more than one expert labeler to reduce potential bias in labeling. In order to provide the basis for a good model, each data observation should contain an accurately-labeled ground truth.

2. **Preprocessing**

Next, data need to be preprocessed before feeding them into the model. This is especially important for images, where specific image processing techniques exist to normalize data and reduce artifacts in the images. The purpose of image preprocessing is an improvement of the image data that suppresses unwanted distortions or enhances some image features important for further processing. Several preprocessing algorithms have been studied for accuracy, variability, and reproducibility [145]. Image preprocessing typically consists of image scaling, intensity normalization, dimensionality reduction, adding/reducing noise, etc. Image scaling refers to the resizing of a digital image, resulting in a higher or lower number of pixels per image. This is useful because some images that feed into an AI algorithm vary in size; therefore, a base size for all images must be established before feeding them into the algorithm. In deep neural networks, intensity rescaling is commonly employed to restrict the image to the range of 0 to 1 due to possible overflow, optimization, stability issues, and so on. Gray scaling is another type of transformation which turns a color RGB image into an image with only shades of gray representing colors. Gray scaling is commonly used in the preprocessing step in machine learning, especially in radiomics [51]. Image normalization in deep learning refers to intensity rescaling, gray scaling, centering, and standard deviation normalization.

For non-imaging data, normalization depends on the data type. In genetics data, where counts often differ dramatically, transformations via the use of log counts is a

common strategy. In others, normalization via using z-scores or rescaling from 0 to 1 are common strategies. For other data such as 1-D Electrocardiogram (ECG) or Photoplethysmogram (PPG) signals, low- or high-pass filters may be required to filter out signals unrelated to the heart.

In AI classification problems, there are often too many correlated and/or redundant features, which increase computational requirements but provide no new information. They may also bias evaluation metrics, as they arbitrarily increase the dimensionality of the prediction space and likely pull data points further from each other in space. Therefore, conventional practice aims to reduce dimensionality through linear algebra techniques such as projection or factorization, or via feature selection. Dimensionality reduction is the process of reducing the number of random variables or features under consideration by obtaining a set of principal variables or features.

The choice of whether or not to conduct the next step of data preparation, feature extraction, is contingent upon whether or not the model will automatically extract features from its input signal. In the case of many Convolutional Neural Network (CNN)s, the feature extraction step is generally skipped, because the model itself determines features of importance directly from the pixel intensities of the image. In most other models, however, feature extraction must be implemented, whereby quantitative features are pulled from an existing image as representative summaries then fed into the model.

3. **Model building and evaluation**

Once the data are collected and preprocessed, the data need to be split into at least two groups; a training set and a test set. The training data will be used to train a model and the test data will be used to evaluate the trained model. In other cases, another subset of the training set will be removed as a validation set. The validation set is used to assess the performance of the model built in the training phase. In this situation, k fold cross validation method is one of the most popular methods in machine learning models to estimate how accurately a predictive model will perform. In k fold cross validation, the data is divided into k subsets. Then, one of the k subsets is used as the test dataset and the other $k-1$ subsets are a training set to train a model. This method will be repeated k times. Other common forms of validation include LOOCV and bootstrapping. After validation, the test set will be used to evaluate the performance of the trained model. This process is the final step, to be conducted after validation is completed.

There are several metrics to evaluate model's performance. The choice of evaluation

metrics depends on a given machine learning task such as regression, classification, or clustering. In general, the Root Mean Squared Error (RMSE) is commonly used in regression problems and accuracy, precision, and recall are commonly used in classification problems. Cross-validation techniques are also used to compare the performance of different machine learning models on the same dataset.

4. **Inference**

   While the terms "inference" and "prediction" are sometimes muddled in the machine learning community, inference has traditionally referred to understanding the factors that influence the distribution of the data [168]. Prediction on the other hand is the forward-looking notion of taking the data inputs and using them to evaluate new examples. Statistical inference techniques were developed on much smaller datasets but provide some insight into the data generation process even in a big data world. Typically, statistical models will provide their coefficients and use them to interpret and understand how it came to predictions. Machine learning techniques tend to focus on prediction and aren't focused on creating a parsimonious and interpretable model of the world. That said, research is being conducted into interpretability of deep learning models, such as work in adversarial networks and network dissection [35, 18]. Both predictive power and the inferences a model is making are important, but that can vary depending on the application and its goals.

The workflow shown here describes a typical process for developing a CDS. Crucially, data and their labels should be trustworthy and of high quality. In order to boost a model's robustness, various strategies like augmentation of data and preprocessing to reduce batch effects are necessary. Lastly, good cross-validation, evaluation, and an ability to interpret results is desired. Although CDS models are constructed to handle a bit of versatility in data inputs, they are not usually conditioned to handle data from entirely different modalities or domains based on different distributions of data or acquisition changes. Therefore, the discussion continues about how to incorporate CDS systems for diverse data.

1.3 CDS Systems for Diverse Data

CDS systems that intake diverse types may consider leveraging information from multiple modalities to enrich decision-making through a more holistic lens. A modality may describe something observed in the natural world — something seen or heard, for example. These things seen or heard are often measured via an instrument and subsequently captured in digital form for computers to render or decipher patterns from. An example of this is a

camera, which measures intensity of light bouncing off of objects and portrays this in a form of pixel intensities in the red, blue and green spectrums. Another example of this is sound, which can be captured through recording of wavelengths bouncing off of a vibrating sensor in a microphone. In other aspects of medicine, X-rays can measure pixel intensities based off the deflection of X-rays in tissue, and MRI can measure spatial distributions of hydrogen nuclei in the body.

Implementing a multimodal learning approach in CDS systems comes with a set of unique challenges. An in-depth discussion of these challenges in multimodal learning as well as a survey of state-of-the-art methods of approaching each challenge is discussed in length in Chapter 2. Briefly, challenges include how to represent multimodal data in ways that can preserve relationships between modalities (representation), how to fuse multimodal data into a single discriminatory model in a way that best leverages each modality (fusion), how to map or translate one modality to another modality (translation), how to align modalities together (alignment), or how to distill knowledge learned from one modality to a model for another modality (co-learning).

Another kind of diverse data we focus on is data from different domains. It is important to consider this kind of data because clinics are dynamic environments where data collections and populations can change. Different institutions also collect the same data in different ways such as with different devices, manufacturers, techniques and workflows. For example, while MRI under different acquisition parameters may change pixel intensity values in an image, humans are very good at finding salient patterns in images from all kinds of settings and normally small changes in parameters may not affect the overall ability to detect the patterns. Likewise, different staining protocols for histopathology tissue may change contrast of tissue, but many human mental models can still distinguish patterns in tissue without much difficulty.

By contrast, small changes in a computational model may be considered a "domain shift", whereby the model is no longer able to make an accurate prediction for the data. In the case of lens implants in cataract surgery, for example, implantation of different lens models from different manufacturers can cause a critical need for adjustment in prediction models determining patients refraction after surgery. In such cases, simple models are domain-adapted through the use of experimentally-derived variables called "A-constants," which can shift the model as needed [173]. In these cases, accounting for domain differences is a critical part of CDS and can determine whether a patient requires a second corrective surgery or not. Thus, building models that incorporate diversity of domains is imperative to constructing informed, real-world CDS models.

Although it is understood that including diverse data such as multimodal and multido-

main data is helpful for mimicking real-world clinical scenarios, many institutions and clinics face practical limitations in executing such strategies. Although the recorded number of health center sites has been increasing in the United States since 2010 [180], 83.6% of US medical practices contain less than 100 physicians and 55.9% of institutions contain less than 10 physicians [179]. As a result, practices with fewer physicians likely can gather less patient data, contain a lower probability of complete datasets for multimodal processing, and hold less opportunities for collaboration with other physicians to increase dataset size and provide data from different domains. Therefore, we recognize the importance of adapting innovative multimodal and multidomain AI models to cases of limited data for inclusivity of both small and large data sources. Such models for limited data can also be scaled up to larger datasets should more data be obtained.

1.4 In This Work

This dissertation will focus on developing innovative multimodal and multidomain AI models for clinical decision support, particularly in scenarios with limited data availability. Our research seeks to address questions regarding constructing machine-learning-based models that mimic real-world mental models and bridge domain gaps in cases of limited data.

We begin with a comprehensive survey of state-of-the-art multimodal learning methods, including a discussion on domain adaptation and generalization. Although this dissertation concentrates on implementation given limited data, it is imperative to begin with a well-rounded understanding of the field and state-of-the-art methodologies. Then we can understand that many of these approaches are based on deep learning strategies and are not suitable for limited data cases.

This leads to our experimental work, where we collect three specific use cases demonstrating clinical challenges with limited sample sizes where a multimodal or multidomain approach is leveraged despite size limitations. In our three exemplary case studies, we present three specific use cases illustrating clinical challenges with limited sample sizes where multimodal or multidomain approaches are leveraged. These include distinguishing pseudoprogression and true progression in glioblastoma using multimodal MRI (first case), detecting temporomandibular joint osteoarthritis with privileged information (second case), and achieving domain generalization in predicting postoperative refraction in cataract surgery (third case).

In the first case, we attempt to leverage multimodality in a small sample of MRI images to distinguish pseudoprogression and true progression in glioblastoma. While most modern imaging techniques are based on leveraging deep neural network architectures, it may not be possible to apply such methods when datasets are small and no appropriate pretrained

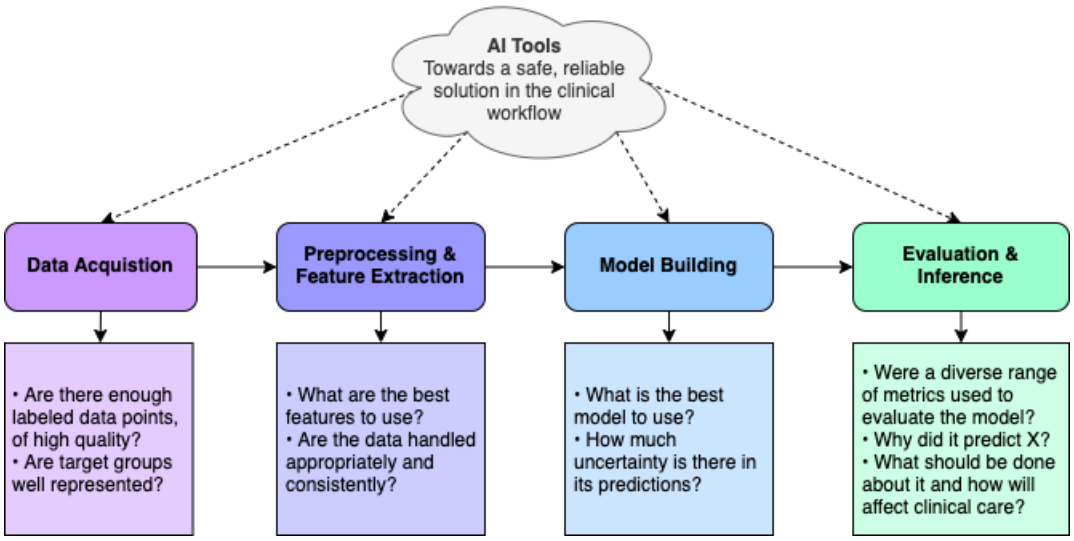

Figure 1.2: During each phase of the machine learning workflow, model builders can ask themselves the above questions to ensure that their model is being incorporated meaningfully: designed robustly with minimized bias and attention to methodological detail.

models exist. In this chapter, we demonstrate a low-parameter approach designed to examine unique properties in multimodal MRI even with small sample sizes.

In the second case, we consider the clinical paradigm where routinely-collected variables in the clinic can provide low-signal information towards prediction of a disease, but where non-routine imaging or lab tests can provide better information. We attempt to leverage these additional modalities using a knowledge-distillation paradigm called "privileged information" and introduce the application of our "Random Forest+" (RF+) model to detect temporomandibular joint osteoarthritis with privileged information. While it is desirable to have high-signal features available all the time, it is important to develop real-world application models that do not require the high-signal data to function but can still leverage their discriminatory abilities. The RF+ model attempts to build a discriminatory model that leverages all information based on these realistic constraints.

The last case presented describes a specific multidomain scenario of applying machine learning to postoperative refraction in cataract surgery. In this case, limited data prevented training on multiple lens types. Therefore, the study targets domain generalization of a machine learning model trained on patients implanted with one lens type, with the goal of adapting the model to provide accurate predictions for other lens types. The study demonstrates an approach of applying innovative feature engineering to reduce error in data from different domains.

1.5 Commitment to Meaningful AI-Based Models

The projects in this thesis present proof-of-concept models for CDS in specific diagnostic scenarios. How does one move beyond creation of the model? Although many proof-of-concept models exist, many ML methods used for clinical decision support face difficulties which prevent real-world deployment. One reason for this is the FDA approval process, as some clinical decision support systems are categorized as a "medical device" under the Food and Drug Administration (FDA) and must therefore pass an approval process before serving patients in the public domain [55]. According to Van Norman [197], the average time for a medical device to make it to market (pass the FDA approval process and present a finished product which is widely available) is on the range of 3 to 7 years. Another reason, however, is often the lack of understanding between decision-support developers and the target clinical environment, leading to impractical and often biased models. To this end, we focus on this latter point, describing what we believe "meaningful" incorporation of AI to entail. In Figure 1.2, we present our ML workflow along with possible critical questions that model-builders can ask when assessing their own or others' models in order to ensure that the model is meaningfully executed with attention to the below considerations. This section assumes that input data provided to model-builders have been provided which meet the criteria for proper record-keeping, unbiased reporting, and fairly complete records and speaks directly to the stages of preprocessing, model-building, and inference.

The wide-scale availability of software such as TensorFlow and Pytorch, as well as packages like scikit-learn have resulted in an eruption of candidate AI-based academic decision-support systems. With the potential for so many choices of support systems in the future, it will become important to standardize reporting and approaches to data acquisition, preprocessing, and model building, so that models can be compared effectively. Standards such as the Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD) statement [127], which provides a checklist of information that should be reported on a model, help to build trust in clinical models. On a similar note, it is also critical to ensure that models are safe, effective, and are as minimally-biased against any particular race, age, or social class as possible.

With the proliferation of AI systems, it is easy to miss the fundamental principles of statistical evaluation. As models become more complex, overfitting on a training set becomes easier and easier. Dividing a dataset into training, validation, and testing sets is a critical part of evaluating the performance of any model. A training set is useful for training model parameters which are learned, while a validation set, in concert with a training set, can be used to view the model's response to changes in hyperparameters like number of layers in a

model. A testing dataset, sometimes denoted a holdout set, should be completely held aside to provide an unbiased estimate of predictive power, and tested only after the final model is conjured. Likewise, Russell and Norvig [166] assert that peeking at testing data is an easy way to go wrong, particularly in repeating an experiment with the same testing dataset until the results are better.

Beyond multiple testing, another source of overfitting is data leakage, wherein information from the testing set influences the training data. Data leakage can occur in seemingly innocuous ways, such as normalizing features or doing principal component analysis using the entire dataset. Furthermore, it's inappropriate to normalize the testing dataset based on itself, rather the normalization from the training set, as the former would push the feature distributions to be more similar than they are in reality. Proper technique would require saving normalization or Principal Component Analysis (PCA) parameters from the training set for use in normalization or PCA of the test set, respectively. Another pitfall is splitting an individual subject across training/testing sets or cross-validation folds, so that the model has already seen the individual its trying to predict. A simple way to correct this is through subject-wise cross-validation, which simply places all records of a patient in the same dataset. Additionally, the population from the original dataset may be unrepresentative of the general population, or may be missing subtypes, which can't be easily detected within a single dataset. For these reasons, we suggest additional validation of a model's performance through conducting one or more external validation studies whenever possible.

Likewise, any AI system targeted at clinical practice should recognize that publishing a good predictive result on a single dataset is only the start. Factors beyond a single standard predictive power measurement are equal determinants of a model's true performance clinically. Oakden-Rayner and Palmer [134] provide a comprehensive summarization of both the validation and study design process that should be undergone by groups who intend to implement their decision-support systems clinically. They begin with the distinction between safety and efficacy, stressing that model performance does not equate to patient safety, and that efficacy, if based on a concept such as saving lives, needs to be quantified as such by lives saved and then compared to some gold standard which has a firm scientific basis.

Further cautions should be heeded to ensure the best ground truth labeling, a critical issue in the data acquisition phase. Supervised models (most discriminatory models) are measured and trained based on the assumption that labels are correct. However, this assumption is hardly met in the clinical realm because physicians are not often certain what condition a patient has. In one case, mammography, radiologists agreed with their colleagues only 78% of the time (inter-rater reliability), while they agreed with themselves only 84% of the time [52]. This issue can be further exacerbated for early detection problems where physicians

must determine when a patient begins exhibiting early signs of disease outcome. Cases such as heart failure or arrhythmia may be easier for physicians to detect early as opposed to a slow and ambiguous condition such as liver cancer or sepsis. In these cases, Oakden-Rayner and Palmer suggest using as many physicians as possible for ground truth labeling, casting doubt on the common practice of assigning only two to three physicians as potentially adding "significant bias" to the overall model [134]. Adamson and Welch [5] state their concerns with the inter-rater reliability problem and their possible solution from a pathologist's standpoint. In another approach to handling disagreement in labeling, Reamaroon et al attempt to factor in physician confidence into SVM model labels, attributing higher weight to physician labels which are presented with higher confidence of correct attribution [157]. However, Friedman [56] shows that confident physicians do not always equate to the best physicians. This may also be problematic, as the "overconfident but incorrect" physician may erroneously bias the model away from correct discrimination. An additional method to address the issue of label uncertainty may involve the use of fuzzy networks or Bayesian-based methods [86, 101].

Lastly, in the model building phase, one must be careful to choose the correct reporting statistics. It is common practice to report AUC for machine learning models, but we further suggest incorporating a "panel" of measures such as sensitivity and specificity or other task-specific metrics, since each of these measures help reveal potential weaknesses in the model. For example, if AUC and accuracy differ by a sufficient amount, one can ascertain that the model itself is likely biased toward the outcome group with the most examples in the dataset. Furthermore, although F1 score and AUC can often differ when sensitivity and specificity are imbalanced, F1 score is thought to be more robust than AUC with highly imbalanced datasets. Brier score and the no information error rate are also good score options for imbalanced data. Another common practice is to include sensitivity and specificity of the model in reporting. Oakden-Rayner and Palmer suggest instead to incorporate Positive Predictive Value (PPV) and Negative Predictive Value (NPV) [134]. From an epidemiological standpoint, this can be more representative of the effectiveness of the model in a population where the condition of interest is rare, as is the case for many cancer subtypes. We end our discussion with a caution towards the common theme of building models which only optimize AUC by referring to Goodhart's law: "When a measure becomes a target, it ceases to be a good measure."

Although our work only presents proof-of-concept models, we focus on several scientific practices mentioned: good cross-validation, multiple evaluation measures, tests to assess the validity of the model, and prevention of data leakage. Models were run multiple times to achieve robust estimates of performance, univariate analyses were used to corroborate feature importance, code was assessed multiple times, and work was documented carefully.

This dissertation is based on the belief that good science comes from an attention to detail so that these models can prove useful to readers and move beyond proof-of-concept in the future.

Note that although the work presents three individual models which solve different challenges in incorporating multimodal or multidomain data, that the approaches can be applied to other scenarios with other data and combined as needed.

1.6 Summary

In conclusion, we have shed light on the critical role of clinical decision support systems (CDSS) in modern healthcare. With the increasing complexity and variability of patient data, there is a pressing need for innovative AI-driven models that can effectively integrate diverse data sources to enhance clinical decision-making. Multimodal and multidomain approaches offer promising avenues for addressing these challenges, providing a more comprehensive understanding of patient health and enabling more personalized and precise interventions.

The subsequent chapters of this thesis will delve deeper into the methodologies and strategies for developing advanced AI models for clinical decision support. We will explore the intricacies of multimodal learning, discussing challenges such as representation, fusion, translation, alignment, and co-learning, and surveying state-of-the-art methods for addressing these challenges. Additionally, we will investigate domain adaptation and generalization techniques to bridge the gap between different data domains and ensure robust performance across diverse clinical scenarios. Through specific case studies, we aim to demonstrate the feasibility and efficacy of multimodal and multidomain AI models in improving clinical decision-making processes. By the end of this thesis, we present paths to moving forward in the same research directions with the hope of contributing valuable insights and advancements to the field of meaningful and informed AI-driven clinical decision support.

1.7 Publication and Acknowledgement

This Introduction includes passages from a published work [212]: Elisa Warner, Nicholas Wang, Joonsang Lee, and Arvind Rao. *Meaningful incorporation of artificial intelligence for personalized patient management during cancer: Quantitative imaging, risk assessment, and therapeutic outcomes*, page 339–359. Elsevier, 2021.

CHAPTER 2

Multimodal Machine Learning in Image-Based and Clinical Biomedicine: Survey and Prospects

2.1 Abstract

ML applications in medical AI systems have shifted from traditional and statistical methods to increasing application of deep learning models. This survey navigates the current landscape of multimodal ML, focusing on its profound impact on medical image analysis and clinical decision support systems. Emphasizing challenges and innovations in addressing multimodal representation, fusion, translation, alignment, and co-learning, the paper explores the transformative potential of multimodal models for clinical predictions. It also questions practical implementation of such models, bringing attention to the dynamics between decision support systems and healthcare providers. Despite advancements, challenges such as data biases and the scarcity of "big data" in many biomedical domains persist. We conclude with a discussion on effective innovation and collaborative efforts to further the mission of seamless integration of multimodal ML models into biomedical practice.

2.2 Introduction

ML, the process of leveraging algorithms and optimization to infer strategies for solving learning tasks, has enabled some of the greatest developments in AI in the last decade, enabling the automated segmentation or class identification of images, the ability to answer nearly any text-based question, and the ability to generate images never seen before. In biomedical research, many of these ML models are quickly being applied to medical images and decision support systems in conjunction with a significant shift from traditional and statistical methods to increasing application of deep learning models. At the same time, the

importance of both plentiful and well-curated data has become better understood, coinciding as of the time of writing this article with the incredible premise of OpenAI's ChatGPT and GPT-4 engines as well as other generative AI models which are trained on a vast, well-curated, and diverse array of content from across the internet [139].

As more data has become available, a wider selection of datasets containing more than one modality has also enabled growth in the multimodal research sphere. Multimodal data is intrinsic to biomedical research and clinical care. While data belonging to a single modality can be conceptualized as a way in which something is perceived or captured in the world into an abstract digitized representation such as a waveform or image, multimodal data aggregates multiple modalities and thus consists of several intrinsically different representation spaces (and potentially even different data geometries). CT and Positron Emission Tomography (PET) are specific examples of single imaging modalities, while MRI is an example itself of multimodal data, as its component sequences T1-weighted, T2-weighted, and Fluid-Attentuated Inversion Recovery (FLAIR) can each be considered their own unique modalities, since each of the MR sequences measure some different biophysical or biological property. Laboratory blood tests, patient demographics, ECG and genetic expression values are also common modalities in clinical decision models. This work discusses unique ways that differences between modalities have been addressed and mitigated to improve accuracy of AI models in similar ways to which a human would naturally be able to re-calibrate to these differences.

There is conceptual value to building multimodal models. Outside of the biomedical sphere, many have already witnessed the sheer power of multimodal AI in text-to-image generators such as DALL·E 2, DALL·E 3 or Midjourney [153, 21, 140], some of whose artful creations have won competitions competing against humans [122]. In the biomedical sphere, multimodal models provide potentially more robust and generalizable AI predictions as well as a more holistic approach to diagnosis or prognosis of patients, akin to a more human-like approach to medicine. While a plethora of biomedical AI publications based on unimodal data exist, fewer multimodal models exist due to cost and availability constraints of obtaining multimodal data. However, since patient imaging and lab measurements are decreasing in cost and increasing in availability, the case for building multimodal biomedical AI is becoming increasingly compelling.

With the emergence of readily-available multimodal data comes new challenges and responsibilities for those who use them. The survey and taxonomy from [15] presents an organized description of these new challenges, which can be summarized in Figure 2.1: 1) Representation, 2) Fusion, 3) Alignment, 4) Translation, 5) Co-learning. **Representation** often condenses a single modality such as audio or an image to a machine-readable data

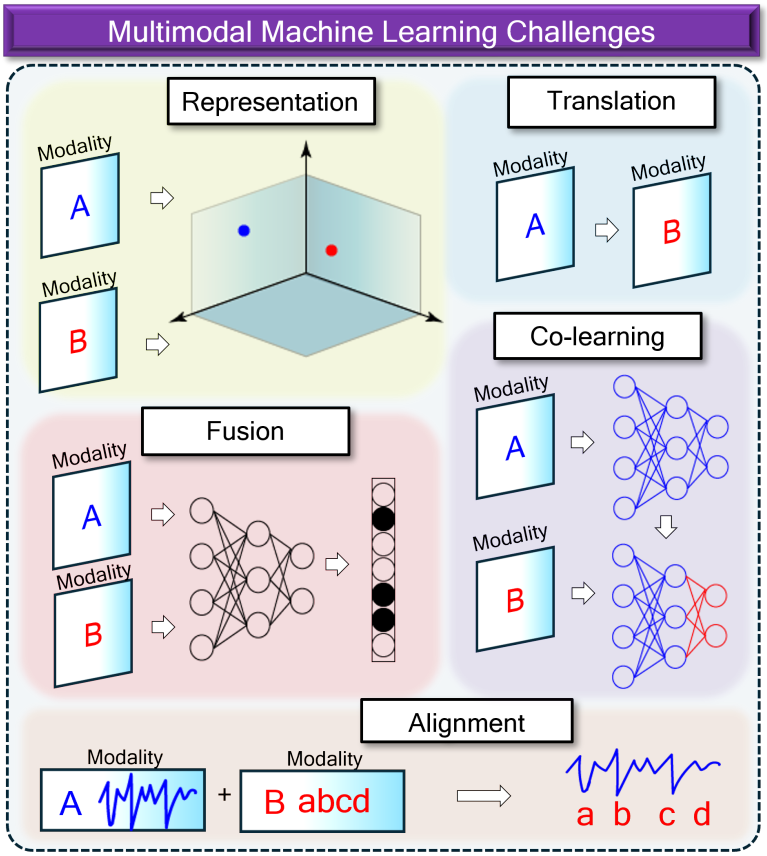

Figure 2.1: Challenges in multimodal learning: 1) Representation, which concerns how multiple modalities will be geometrically represented and how to represent intrinsic relationships between them; 2) Fusion, the challenge of combining multiple modalities into a predictive model; 3) Translation, involving the mapping of one modality to another; 4) Alignment, which attempts to align two separate modalities spatially or temporally; and 5) Co-learning, which involves using one modality to assist the learning of another modality.

structure such as a vector, matrix, tensor object, or other geometric form, and is concerned with ways to combine more than one modality into the same representation space. Good multimodal representations are constructed in ways in which relationships and context can be preserved between modalities. Multimodal **fusion** relates to the challenge of how to properly combine multimodal data into a predictive model. In multimodal **alignment**, models attempt to automatically align one modality to another. In a simple case, models could be constructed to align PPG signals taken at a 60Hz sampling frequency with a 240Hz ECG signal. In a more challenging case, video of colonoscopy could be aligned to an image representing the camera's location in the colon. Multimodal **translation** consists of mapping one modality to another. For example, several popular Natural Language Processing (NLP) models attempt to map an image to a description of the image, switching from the imag-

ing domain to a text domain. In translational medicine, image-to-image translation tends to be the most common method, whereby one easily-obtained imaging domain such as CT is converted to a harder-to-obtain domain such as T1-weighted MRI. Lastly, multimodal **co-learning** involves the practice of transferring knowledge learned from one modality to a model or data from a different modality.

In this paper, we use the taxonomical framework from [15] to survey current methods which address each of the five challenges of multimodal learning with a novel focus on addressing these challenges in medical image-based clinical decision support. The aim of this work is to introduce both current and new approaches for addressing each multimodal challenge. We conclude with a discussion on the future of AI in biomedicine and what steps we anticipate could further progress in the field.

2.3 Multimodal Learning in Medical Applications

In the following section, we reintroduce the five common challenges in multimodal ML addressed in Section 1 and discuss modern approaches to each challenge as applied to image-based biomedicine. The taxonomical subcategories of Representation and Fusion are summarized in Figure 2.2, while those for Translation, Alignment and Co-learning are summarized in Figure 2.3. A table of relevant works by the challenge addressed and data types used are given in Table 2.1.

2.3.1 Representation

Representation in machine learning typically entails the challenge of transferring contextual knowledge of a complex entity such as an image or sound to a mathematically-interpretable or machine-readable format such as a vector or a matrix. Prior to the rise of deep learning, features were engineered in images using techniques such as the aforementioned Shift-Invariant Feature Transforms (SIFT) transforms or through methods such as edge detection. Features in audio or other waveform signals such as ECG could be extracted utilizing wavelets or Fourier transform to isolate latent properties of signals and then quantitative values could be derived from morphological patterns in the extracted signal. Multimodal representation challenges venture a step further, consisting of the ability to translate similarities and differences from one modality's representation to another modality's representation. For example, when representing both medical text and CT images, if the vector representations for "skull" and "brain" in medical *text* are closer than those for "skull" and "pancreas", then in a good CT representation, such relationships between vector representations of these

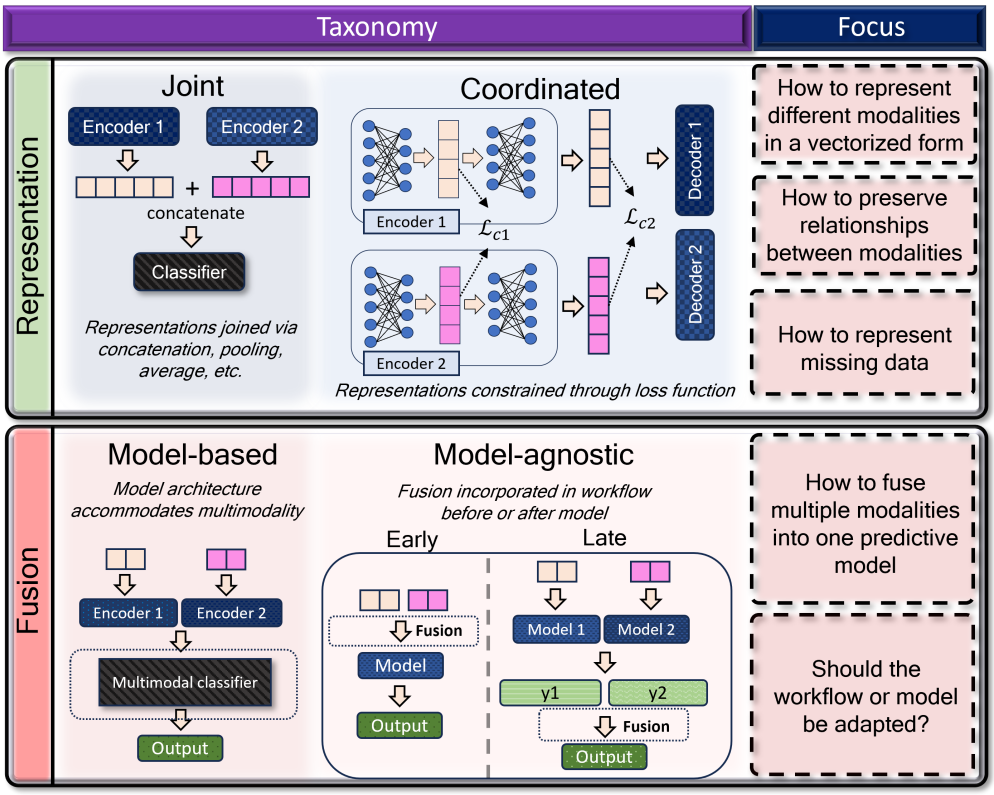

Figure 2.2: A graphical representation of the taxonomical sublevels of multimodal representation and fusion, and the focus of each challenge. Multimodal representation can be categorized into whether the representations are joined into a single vector (*joint*) or separately constructed to be influenced by each other (*coordinated*). Multimodal fusion can be distinguished by whether a model is uniquely constructed to fuse the modalities (*model-based*), or whether fusion occurs before or after the model step (*model-agnostic*).

structures in the *image* should remain preserved. The derivation of "good" representations in multimodal settings have been outlined in Bengio et al [20] and extended by Srivastava and Salakhutdinov [178].

It is crucial to acknowledge that representation becomes notably challenging when dealing with more abstract concepts. In a unimodal context, consider the task of crafting representations from an image. Beyond pixel intensities, these representations must encapsulate contextual and semantically-proximate information from the image. A simplistic model may fail to encode context adequately, discerning insufficient distinctions between a foreground and background to represent nuanced visual-semantic concepts. Achieving such subtleties in representations, particularly in abstract contexts, poses increased challenges compared to quantifying similarities and differences in less-nuanced data such as cell counts or gene expression.

Prior to delving into multimodal representations, it is instructive to elucidate strategies for crafting proficient unimodal representations, as multimodal approaches often involve combining or adapting multiple unimodal methods. For images, *pretrained networks* are a common approach for transforming images into good vector representations. Another approach is use of autoencoders, which condense image representations into lower-dimensional context vectors that can be decoded to reconstruct the original image. *Multimodal autoencoders* have been applied to MRI modalities in [68] and in this example were also utilized to impute representations for missing modalities.

Another approach for multimodal representation could be through the use of *disentanglement networks*, which can separate latent properties of an image into separate vectors. In such cases, an image is given as input and the autoencoder is often split in such a way that two vectors are produced as intermediate pathways, where joining the intermediate vectors should result in the original input. Each intermediate pathway is often constrained by a separate loss function term to encourage separation of each pathway into the desired latent characteristics. In this way, one input image can be represented by two separate vectors, each representing a disjointed characteristic of the image. This disentanglement method has been applied in [85] to separate context in CT and MRI from their style so that one modality can be converted in to the other. It was also applied for a single modality in [25] to separate "shape" and "appearance" representations of an input, which could potentially be applied to different imaging modalities to extract only similar shapes from each.

When two or more vectorized modalities are combined into a model, they are typically combined in one of two ways: 1) joint, or 2) coordinated representations. A **joint representation** is characterized by aggregation of the vectors at some point in the process, whereby vector representations from two separate modalities are joined together into a single vector

form through methods such as aggregation, concatenation or summation. Joint representation is both a common and effective strategy for representation; however, a joint strategy such as concatenation is often constricted to serving in situations where both modalities are available at train- and test-time (one exception using Boltzmann Machines can be found in [178]). If a modality has the potential to be missing, a joint strategy such as aggregation via weighted means could be a better option [108, 34, 233, 41]. Using mathematical notation from [15], we can denote joint representations x_m as the following:

$$x_m = f(x_1, ..., x_n) \tag{2.1}$$

This denotes that feature vectors $x_i, i = 1...n$ are combined in some way through a function f to create a new representation space x_m. By the contrary, **coordinated representations** are represented as the following:

$$f(x_1) \sim g(x_2), \tag{2.2}$$

whereby a function designed to create representations for one modality may be constrained (represented by $\sim$) by a similar function from another modality, with the assumption that relationships between data points in the first modality should be relatively well-preserved in the second modality.

Joint representations tend to be the most common approach to representing two or more modalities together in a model because it is perhaps the most straightforward approach. For example, joining vectorized multimodal data together through concatenation before entering a model tends to be one of the most direct approaches to joint representation. In [199], for example, chest x-rays are combined with text data from electronic health records in a vectorized form using a pretrained model first. Then, the vectors from each modality are sent individually through two attention-based blocks, then concatenated into a joint feature space to predict a possible cardiovascular disease and generate a free-text "impression" of the condition. Other joint representation models follow simpler methods, simply extracting baseline features from a pretrained model and concatenating them [42, 220].

Although coordinated representations have traditionally tended to be more challenging to implement, the convenience of neural network architectural and loss adjustments have resulted in increased traction in publications embodying coordinated representations [216, 206, 31, 152, 231, 22]. One of the most notable of these in recent AI approaches is OpenAI's Constrastive Language-Image Pretraining (CLIP) model, which forms representations for OpenAI's DALL·E 2 [152, 153] and uses a contrastive-learning approach to shape both image embeddings of entire images to match text embeddings of entire captions describing those

images. The representations learned from CLIP were demonstrated to not only perform well in zero-shot image-to-text or text-to-image models, but also to produce representations that could outpace baseline supervised learning methods. In a biomedical context, similar models abound, including ConVIRT, a predecessor and forerunner for CLIP [231], and related works [22].

Coordinated approaches are especially useful in co-learning. In [31], which employs a subset of co-learning called privileged information, the geometric forms of each modality are not joined into a single vector representation. Instead, network weights are encouraged to produce similar output vectors for each modality and ultimately the same classifications. This constraint warps the space of chest x-ray representations closer to the space of text representations, with the assumption that this coordinated strategy provides chest x-ray representations more useful information because of the text modality. For more on privileged information, see the Section 2.3.5.1 below.

In the biomedical sphere, where models are built to prioritize biologically- or clinically-relevant outcomes, quality of representations may often be overlooked or overshadowed by emphasis on optimization of prediction accuracy. However, there is conceptual value in building good multimodal representations. If models are constructed to ensure that similar concepts in different modalities also demonstrate cross-modal similarity, then there is greater confidence that an accurate model is understanding cross-modal relationships. While building good cross-modal representations for indexing images on the Internet like in the CLIP model is a digestible challenge, building similar cross-modal representations for medical data presents a far more formidable challenge due to data paucity. OpenAI's proprietary WebTextImage dataset, used for CLIP, contains 400 million examples, a sample size that is as of yet unheard of for any kind of biomedical imaging data. Until such a dataset is released, bioinformaticians must often rely on the ability to leverage pretrained models and transfer learning strategies for optimal results amidst low resources to leverage big data for good representations on smaller data.

2.3.2 Fusion

Next, we discuss challenges in multimodal fusion. This topic is a natural segue from the discussion of representation because many multimodal representations are subsequently fed into a discriminatory model. Multimodal fusion entails the utilization of methods to combine representations from more than one modality into a classification, regression, or segmentation model. According to [15], fusion models can be classified into two subcategories: model-agnostic and model-based approaches. The term **"model-agnostic"** refers to methods for

multimodal fusion occurring either before or after the model execution and typically does not involve altering the prediction model itself. Model-agnostic approaches can further be delineated by the stage at which the fusion of modalities occurs, either early in the model (prior to output generation) or late in the model (such as ensemble models, where outputs from multiple models are combined). Additionally, hybrid models, incorporating a blend of both early and late fusion, have been proposed [29]. In contrast, a **model-based** approach entails special adjustments to the predictive model to ensure it handles each modality uniquely.

While model-agnostic methods remain pertinent as useful strategies for multimodal fusion, the overwhelming popularity of neural networks has led to a predominant shift towards model-based methods in recent years. These model-based methods involve innovative loss functions and architectures designed to handle each modality differently. One common model-based fusion strategy is multimodal *multiple instance learning (MIL)*, where multiple context vectors for each modality are generated and subsequently aggregated into a single representation leading to the output classification. The method for aggregation varies across studies, with attention-based approaches, emphasizing specific characteristics of each modality, being a common choice [108, 34, 233, 41].

The construction of a good model architecture is crucial; however, challenges associated with fusion are often highly contextual, and thus it is important to understand what kinds of data are being utilized in recent models and what problems they try to solve. Most multimodal models understandably incorporate MRI modalities, given that MR images are a natural multimodal medium. Consequently, studies incorporating MRI such as [11], which aims to classify Alzheimer's Disease severity, and [232], predicting overall survival in brain tumor patients, exemplify the type of research often prevalent in multimodal image-based clinical application publications. Brain-based ML studies are also popular because of the wide availability of brain images and a strong interest in applying ML models in clinical neuroradiology. However, recent models encompass a myriad of other clinical scenarios predicting lung cancer presence [42], segmenting soft tissue sarcomas [132], classifying breast lesions [65], and predicting therapy response [220], among others, by amalgamating and cross-referencing modalities such as CT images [42, 132], blood tests [220], electronic health record (EHR) data [220, 199, 42], mammography images [65], and ultrasound [65].

Multimodal fusion models are emerging as the gold standard for clinical-assisted interventions due to the recognition that diagnosis and prognosis in real-world clinical settings are often multimodal problems. However, these models are not without limitations. For one, standardization across equipment manufacturers or measurement protocols can affect model performance dramatically, and this issue becomes more pronounced as more modalities are

incorporated into a model. Second, while fusion models attempt to mimic real-world clinical practice, they face practical challenges that can limit their utility. For instance, physicians may face various roadblocks to obtaining all model input variables due to a lack of permission from insurance companies to perform all needed tests or time constraints. These issues underscore challenges associated with missing modalities, and several studies have attempted to address this concern [29, 230, 41, 206, 112]. However, incorporating mechanisms to account for missing modalities in a model is not yet a common practice for most multimodal biomedical models.

Lastly, many models are not configured to make predictions that adapt with additional variables. Most models necessitate all variables to be present at the time of operation, meaning that, even if all tests are conducted, the model can only make a decision once all test results have been obtained. In conclusion, in the dynamic and fast-paced environment of hospitals and other care centers, even accurate models may not be suitable for practical use, unless also coupled with mechanisms to handle missing data.

2.3.3 Translation

In multimodal translation, a model is devised to operate as a mapping entity facilitating the transformation from one modality to another. This involves the conversion of input contextual data, such as CT scans, into an alternative contextual data format, such as MRI scans. Before the rise of modern **generative** methods leveraging multimodal Generative Adversarial Network (GAN)s or diffusion models to generate one modality from another, translation via **dictionary-based** methods was common, which typically involved a bimodal dictionary whereby a single entry would contain a key belonging to one modality and a corresponding value belonging to the other modality. Dictionary-based translation was uncommon in biomedical research but popular in NLP fields as a way to convert images into text and vice versa [109, 159]. The current ascendancy of generative models and the availability of associated coding packages have since catalyzed the growth of innovative translational studies applying generative approaches.

Presently, generative models encompass a broad spectrum of potential applications both within and beyond the biomedical domain. Outside the medical sphere, generative models find utility in NLP settings, particularly in text-to-image models like DALL·E 2 and Midjourney [109, 153, 140]. Additionally, they are employed in style transfer and other aesthetic computer vision techniques [79, 28, 234, 111, 142, 227]. Within the biomedical realm, generative models have proven efficacious in creating virtual stains for unstained histopathological tissues which would typically undergo hemotoxylin/eosin staining [113]. Furthermore, these

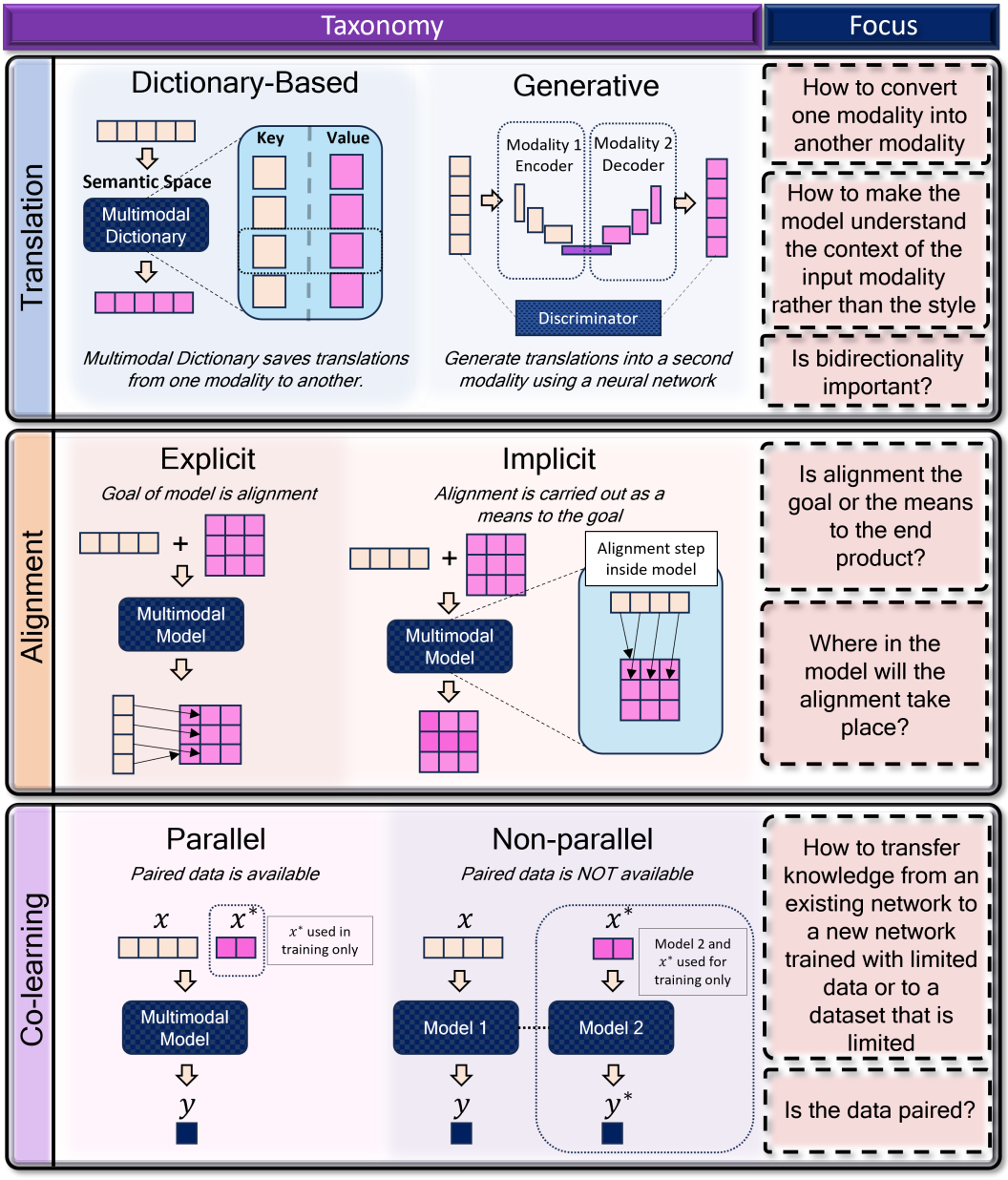

Figure 2.3: A graphical representation of the taxonomical sublevels of multimodal translation, alignment and co-learning, and the focus of each challenge. In **translation**, models are distinguished based on whether they require use of a dictionary to save associations between modalities (*dictionary-based*), or if the associations are learned in a multimodal network (*generative*). In **alignment**, distinction is made depending on the *purpose* of the alignment, whether as the goal (*explicit*) or as an intermediate step towards the goal output (*implicit*). In **co-learning**, a distinction is made between the use of *parallel* (paired) multimodal data, or *non-parallel* (unpaired) multimodal data. In co-learning models, one of the modalities is only used in training but does not appear in testing.

models serve as prominent tools for sample generation [191, 148, 39], particularly in scenarios with limited sample sizes [33]. Despite the potential diversity of multimodal translation involving any two modalities, predominant translational efforts in the biomedical realm currently revolve around mapping one imaging modality to another, a paradigm recognized as image-to-image translation.

In the contemporary landscape, the integration of simplistic generative models into a clinical context are declining in visibility, while methods employing specialized architectures tailored to the involved modalities are acknowledged for advancing the state-of-the-art in translational work. Within this context, two notable generative translation paradigms for biomedicine are explored: 1) medical image generation models, and 2) segmentation mask models. In the former, many studies attempt to form models that are bidirectional, whereby the intended output can be placed back as input and return an image similar to the original input image. In [27], this is resolved by generating deformation fields that map changes in the T1-weighted sequence modality of MRI to the T2-weighted sequence modality. In [78], separate forward and backward training processes are defined whereby an encoder representing PET images is utilized to understand the underlying distribution of that modality, allowing for more realistic synthetic generated images from MRI. In one unidirectional example, [174] modifies a pix2pix conditional GAN network to allow Alzheimer's disease classification to influence synthetic PET image generation. In another interesting example, [186] use functional MRI (fMRI) scans and diffusion models to attempt to recreate images of what their subjects had seen. Similarly, diffusion models and magnetoencephalography (MEG) are utilized by Meta for real-time prediction from brain activity of what patients had seen visually [19].

In the second potential application, image segmentation models in multimodal image-to-image translation must handle additional challenges, creating both a way to generate the output modality as well as a way to segment it. In [85], a generative model converts CT to MRI segmentation. In a reverse problem to image segmentation, [63] attempts to synthesize multimodal MRI examples of lesions with only a binary lesion mask and a multimodal MRI Atlas. In this study, six CNN-based discriminators are utilized to ensure the authentic appearance of background, brain, and lesion, respectively, in synthesized images.

Multimodal translation still remains an exciting but formidable challenge. In NLP and beyond, there have been remarkable successes observed in new image generation within text-to-image models beyond the biomedical sphere. However, the adoption of translation models in biomedical work is evolving at a more measured pace, with applications extending beyond demonstrative feasibility to practical utility remaining limited. Arguments in favor of biomedical translation models are predominantly centered around sample generation for datasets with limited sizes, as the generated medical images must adhere to stringent accu-

racy requirements. Similar to other challenges in multimodal research, translation models would greatly benefit from training on more expansive and diverse datasets. However, with the increasing digitization of medical records and a refined understanding of de-identification protocols and data sharing rights, the evolution of this field holds considerable promise.

2.3.4 Alignment

Multimodal alignment involves aligning two related modalities, often in either a spatial or temporal way. Multimodal alignment can be conducted either *explicitly* as a direct end goal, or *implicitly*, as a means to the end goal, which could be translation or classification of an input. One example of **explicit alignment** in a biomedical context is image registration. [102] highlights one approach to multimodal image registration, where histopathology slides are aligned to their (x, y, z) coordinates in a three-dimensional CT volume. Another is in [36], where surgical video was aligned to a text description of what is happening in the video. On the other hand, an example of multimodal **implicit alignment** could be the temporal alignment of multiple clinical tests to understand a patient's progress over time. Such an analysis was conducted in [220], where the authors built a customized multi-layer perceptron (MLP) called SimTA to predict response to therapy intervention at a future time step based on results from previous tests and interventions.

Literature surrounding alignment has increased since the rise of attention-based models in 2016. The concept of "attention," which relates to aligning representations in a way that is contextually relevant, is a unimodal alignment paradigm with origins in machine translation and NLP [12]. An example use of attention in NLP could be models which try to learn, based on order and word choice of an input sentence, where the subject of the sentence is so that the response can address the input topic. In imaging, attention can be used to highlight important parts of an image that are most likely to contribute to a class prediction. In 2017, Vaswani et al [202], introduced a more sophisticated attention network, named transformers, an encoder-decoder-style architecture based on repeated projection heads where attention learning takes place. Transformers and attention were originally applied to natural language [202, 12, 46] but have since been applied to images [144, 50], including histopathology slides [114, 34] and protein prediction [192]. *Multimodal transformers* were introduced in 2019, also developed for the natural language community [190]. While these multimodal transformers do not contain the same encoder-decoder structure of a traditional transformer architecture, they are hallmarked by crossmodal attention heads, where one modality's sequences intermingle with another modality's sequences.

Although typical transformers themselves are not multimodal, they often constitute in

multimodal models. The SimTA network mentioned above borrowed the positional encoding property of transformers to align multimodal inputs in time to predict therapy response [220]. Many models taking advantage of visual transformers (ViT) have also utilized pretrained transformers trained on images for multimodal fusion models. In both the TransBTS [207] and mmFormer models [230], a transformer is utilized on a vector composed of an amalgamation of information from multiple modalities of MRI, which may imply that the transformer attention heads here are aligning information from multiple modalities represented via aggregate latent vectors. The ultimate function of transformers is a form of implicit alignment, and it can be assumed here that this alignment is multimodal.

Transformer models have brought a new and largely successful approach to alignment, sparking widespread interest in their applications in biomedical use. Transformers for NLP have also engendered new interest in Large Language Models (LLMs), which are already being applied to biomedical contexts [189] and probing new questions about its potential use as a knowledge base for biomedical questions [183].

2.3.5 Co-learning

In this last section exploring recent research in multimodal machine learning, the area of co-learning is examined, a field which has recently garnered a strong momentum in both unimodal and multimodal domains. In multimodal co-learning, knowledge learned from one modality is often used to assist learning of a second modality. This first modality which transfers knowledge is often leveraged only at train-time but is not required at test-time. Co-learning is classified in [15] as either parallel or non-parallel. In **parallel** co-learning, paired samples of modalities which share the same instance are fed into a co-learning model. By contrast, in **non-parallel** co-learning, both modalities are included in a model but are not required to be paired.

While co-learning can embody a variety of topics such as conceptual grounding and zero-shot learning, this work focuses on the use of transfer learning in biomedicine. In *multimodal transfer learning*, a model trained on a higher quality or more plentiful modality is employed to assist in the training of a model designed for a second modality which is often noisier or smaller in sample size. Transfer learning can be conducted in both parallel and non-parallel paradigms. This work focuses on one parallel form of transfer learning called privileged learning, and one non-parallel form of transfer learning called domain adaptation. A visual representation of these approaches be seen in Figure 2.4.

2.3.5.1 Privileged Learning

Privileged learning originates from the mathematician Vladmir Vapnik and his ideas of knowledge transfer with the support vector machine for privileged learning (SVM+) model [200]. The concept of privileged learning introduces the idea that predictions for a low-signal, low-cost modality can be assisted by incorporating a high-signal, high-cost modality (privileged information) in training only, while at test-time only the low-cost modality is needed. In [200], Vapnik illustrates this concept through the analogy of a teacher (privileged information) distilling knowledge to a student (low-cost modality) before the student takes a test. Although a useful concept, the field is relatively under-explored compared to other areas of co-learning. One challenge to applying privileged learning models was that Vapnik's SVM+ model was one of few available before the widespread use of neural networks. Furthermore, it demands that the modality deemed "privileged" must confer high accuracy on its own in order to ensure that its contribution to the model is positive. Since then, neural networks have encouraged newer renditions of privileged information models that allow more flexibility of use [98, 172, 167].

Recently, privileged learning has emerged as a growing subset of biomedical literature, and understandably so. Many multimodal models today require health care professionals to gather a slew of patient information and are not trained to handle missing data. Therefore, the ability to minimize the number of required input data while still utilizing the predictive power of multiple modalities can be useful in real-world clinical settings. In [76] for example, the authors attempt to train a segmentation network where at train-time the "teacher network" contains four MR image modalities, but at test-time the "student network" contains only T1-weighted images, the standard modality used in preoperative neurosurgery and radiology. In [31], chest x-rays and written text from their respective radiology reports are used to train a model where only chest x-rays are available at test-time.

In privileged models based on traditional approaches (before deep neural networks), privileged information can be embedded in the model either through an alteration of allowable error ("slack variables" from SVM+) [200], or through decision trees constructed with non-privileged features to mimic the discriminative ability of privileged features (Random Forest+) [209, 128]. In a deep learning model, privileged learning is often achieved through the use of additional loss functions which attempt to constrain latent and output vectors from the non-privileged modality to mimic those from the combined privileged and non-privileged models [76, 216]. For example, in [31], encoders for each modality are compared and cross entropy loss is calculated for each modality separately. The sum of these allows the chest x-ray network to freely train for only the chest x-ray modality while being constrained through the overall loss function to borrow encoding methods from the text network, which also strives

to build an accurate model.

While privileged learning models can be applied where data is missing, users should heed caution when applying models in situations where there is systematic bias in reporting. Those who train privileged models without considering subject matter may inadvertently be choosing to include all their complete data in training and their incomplete data in testing. However, in clinical scenarios, data are often incomplete because a patient either did not qualify for a test (perhaps their condition was seen as not "dire enough" to warrant a test) or their situation was too serious to require a test (for example, a patient in septic shock may not pause to undergo a chest x-ray because they are in the middle of a medical emergency). Therefore, while applying data to highly complex models is a common approach in computer science, the context of the data and potential underlying biases need to be considered first to ensure a practical and well-developed model.

2.3.5.2 Domain Adaptation

Domain adaptation has been shown to be useful in biomedical data science applications where a provided dataset may be too small or costly to utilize for more advanced methods such as deep learning, but where a somewhat similar (albeit larger) dataset can be trained by such methods. The smaller dataset for which we want to train the model is called the "target" dataset and the larger dataset which will be used to assist the model with the learning task and provide better contextualization is called the "source" dataset. Domain adaptation strategies are often tailored to single modalities such as camera imaging or MRI, where measurements of an observed variable differ based on an instrument's post-processing techniques or acquisition parameters [217, 201, 222]. However, the distinct characteristics arising from disparate instruments or acquisition settings can lead to considerable shifts in data distribution and feature representations, mirroring the challenges faced in true multimodal contexts. Therefore, the discussion of uni-modal domain adaptation is a relevant starting point for multimodal domain adaptation, as it covers approaches to mitigate significant deviations within data that may seem similar but are represented differently. Additionally, understanding how to mitigate the impact of such variations helps one to understand ways to construct multimodal machine learning systems that confront similar challenges. We also discuss relevant multimodal domain adaptation approaches in biomedicine, which have typically consisted of applying CT images as a source domain to train an MRI target model or vice versa [38, 218, 146, 83, 48].

One way to train a model to adapt to different domains is through augmentation of the input data, which "generalizes" the model to interpret outside of the domain of the original data. In [217], a data augmentation framework for fundus images in diabetic retinopathy

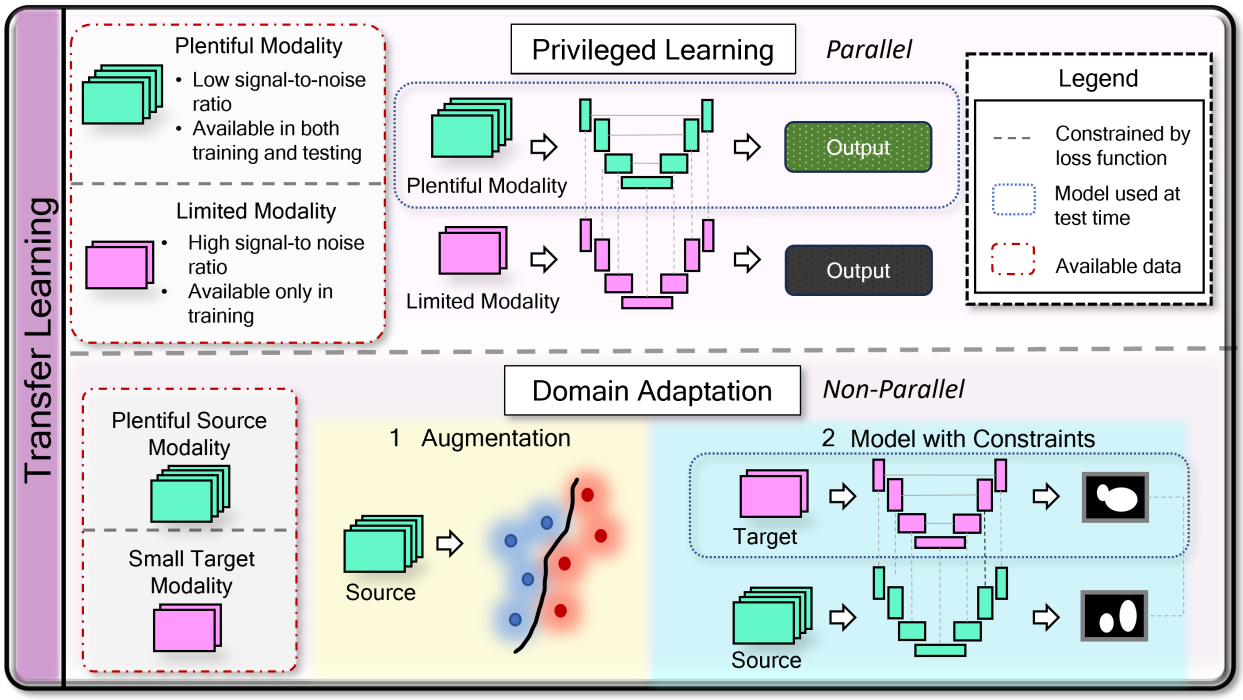

Figure 2.4: Two types of transfer learning described in this work are privileged learning (top) and domain adaptation (bottom). In privileged learning, a plentiful set consisting of data which is normally of low cost but also low signal-to-noise ratio is available in both training and testing, while a limited gold-standard quality set is used for training only. In this example, the plentiful set is used to train the target model, while the limited set constrains the model parameters to increase the model's ability to associate the low-cost modality with the ground truth. In domain adaptation, there is a target dataset which consists of a few samples and a source dataset consisting of plenty of samples. If the target data is too small to build a reliable model in training, source data can be augmented to make the model more robust. Else, the target model could be trained with few examples, while a second source model is used to help make the target model more generalizable.

(DR) is proposed to offset the domain differences of utilizing different cameras. The authors show that subtracting local average color, blurring, adaptive local contrast enhancement, and a specialized PCA strategy can increase both R^2 values for age prediction and Diabetic Retinopathy (DR) classification AUROC or AUC on test sets where either some domain information is known *a priori* and also where no information is known, respectively. In another method which attempts to augment the source domain into more examples in the target style, [38] split the source image into latent content and style vectors, using the content vectors in a style-transfer model reminiscent of cycleGAN to feed as examples with the target domain into a segmentation network [235]. In other applications, data augmentation for domain generalization may be executed utilizing simpler affine transformations [201]. This

demonstrates the utility of data augmentation strategies in more broadly defining decision boundaries where target domains differ from the source.

A second strategy for domain adaptation involves constraining neural network functions trained on a target domain by creating loss functions which require alignment with a source domain model. In [201], a framework for adapting segmentation models at test-time is proposed, whereby an adversarial loss trains a target-based U-Net to be as similar to a source-based U-Net as possible. Then a paired-consistency loss with adversarial examples is utilized to fine-tune the decision boundary to include morphologically similar data points. In a specificially multimodal segmentation-based model, [218] attempts to create two side-by-side networks, a segmenter and an edge generator, which both encourage the source and target output to be as similar as possible to each other. In the final loss function, the edge generator is used to constrain the segmenter in such a way as to promote better edge consistency in the target domain. In yet another, simpler example, domain adaptation to a target domain is performed in [77] by taking a network trained on the source domain and simply adjusting the parameters of the batch normalization layer.

Domain adaptation in biomedicine can be a common problem where instrument models or parameters change. Among multimodal co-learning methods, most networks are constructed as segmentation networks for MRI and CT because they are similar imaging domains, although measuring different things. While CT carries distinct meaning in its pixels (measured in Hounsfield Units), MRI pixel intensities are not standardized and usually require normalization, which could pose challenges to this multimodal problem. Additionally, MRI carries much more detail than CT scans, which necessitates the model to understand contextual boundaries of objects much more than a unimodal case with only CT or MRI.

2.4 Discussion

The rapidly evolving landscape of AI both within the biomedical field and beyond has posed a substantial challenge in composing this survey. Our aim is to provide the reader with a comprehensive overview of the challenges and contemporary approaches to multimodal machine learning in image-based, clinically relevant biomedicine. However, it is essential to acknowledge that our endeavor cannot be fully comprehensive due to the dynamic nature of the field and the sheer volume of emerging literature within the biomedical domain and its periphery. This robust growth has led to a race among industry and research institutions to integrate the latest cutting-edge models into the healthcare sector, with a particular emphasis on the introduction of "large language models" (LLMs). In recent years, there has been an emergence of market-level insights into the future of healthcare and machine learning, as

exemplified by the incorporation of machine learning models into wearable devices such as the Apple Watch and Fitbit devices for the detection of atrial fibrillation [147, 115]. This begs the question: *where does this transformative journey lead us?*

Healthcare professionals and physicians already embrace the concept of multimodal cognitive models in their diagnostic and prognostic practices, signaling that such computer models based on multimodal frameworks are likely to endure within the biomedical landscape. However, for these models to be effectively integrated into clinical settings, they must exhibit flexibility that aligns with the clinical environment. If the ultimate goal is to seamlessly incorporate these AI advancements into clinical practice, a fundamental question arises: how can these models be practically implemented on-site? Presently, most available software tools for clinicians are intended as auxiliary aids, but healthcare professionals have voiced concerns regarding the potential for increased computational workload, alert fatigue, and the limitations imposed by Electronic Health Record (EHR) interfaces [43, 9]. Therefore, it is paramount to ensure that any additional software introduced into clinical settings serves as an asset rather than a hindrance.

Another pertinent issue emerging from these discussions pertains to the dynamics between clinical decision support CDS systems and healthcare providers. What occurs when a computer-generated recommendation contradicts a physician's judgment? This dilemma is not new, as evidenced by a classic case recounted by [53], where physicians were granted the choice to either follow or disregard a CDS system for antibiotic prescription. Intriguingly, the group provided with the choice exhibited suboptimal performance compared to both the physician-only and computer-only groups. Consequently, it is unsurprising that some healthcare professionals maintain a cautious approach to computer decision support systems [5, 175]. Questions arise regarding the accountability of physicians if they ignore a correct computer-generated decision and the responsibility of software developers if a physician follows an erroneous computer-generated recommendation.

A pivotal ingredient notably under-represented in many CDS models, which could help alleviate discrepancies between computer-generated and human decisions, is the incorporation of uncertainty quantification, grounded calibration, interpretability and explainability. These factors have been discussed in previous literature, underscoring the critical role of explainability in ensuring the long-term success of CDS-related endeavors [158, 91, 96, 3].

The domain of multimodal machine learning for medically oriented image-based clinical support has garnered increasing attention in recent years. This interest has been stimulated by advances in computer science architecture and computing hardware, the availability of vast and publicly accessible data, innovative model architectures tailored for limited datasets, and the growing demand for applications in clinical and biomedical contexts. Recent studies

have showcased the ability to generate synthetic images in one modality based on another (as outlined in Section 2.3.3), align multiple modalities (Section 2.3.4), and transfer latent features from one modality to train another (Section 2.3.5), among other advancements. These developments offer a promising outlook for a field that is still relatively new. However, it is also imperative to remain vigilant regarding the prevention of data biases and under-representation in ML models to maximize the potential of these technologies.

Despite these promising developments, the field faces significant hurdles, notably the lack of readily available "big data" in the medical domain. For instance, the routine digitization of histopathology slides remains a challenging goal in many healthcare facilities. Data sharing among medical institutions is fraught with challenges around appropriate procedures for the responsible sharing of patient data under institutional, national and international patient privacy regulations.

Advancing the field will likely entail overcoming these hurdles, ensuring more extensive sharing of de-identified data from research publications and greater participation in establishment of standardized public repositories for data. Dissemination of both code and pretrained model weights would also enable greater knowledge-sharing and repeatability. Models that incorporate uncertainty quantification, explainability, and strategies to account for missing data are particularly advantageous. For more guidance on building appropriate multimodal AI models in healthcare, one can refer to the World Health Organization's new ethics and governance guidelines for large multimodal models [214].

In conclusion, the field of multimodal machine learning in biomedicine has experienced rapid growth in each of its challenge areas of representation, fusion, translation, alignment, and co-learning. Given the recent advancements in deep learning models, escalating interest in multimodality, and the necessity for multimodal applications in healthcare, it is likely that the field will continue to mature and broaden its clinical applications. In this ever-evolving intersection of AI and healthcare, the imperative for responsible innovation resonates strongly. The future of multimodal machine learning in the biomedical sphere presents immense potential but also mandates a dedication to ethical principles encompassing data privacy, accountability, and transparent collaboration between human professionals and AI systems. As we navigate this transformative journey, the collective effort, ethical stewardship, and adherence to best practices will ensure the realization of the benefits of AI and multimodal machine learning, making healthcare more efficient, accurate, and accessible, all while safeguarding the well-being of patients and upholding the procedural and ethical standards of clinical practice.

Table 2.1: Literature relating to the five challenges of multimodal machine learning by the datatype analyzed.

	Datatype			
Challenge	MRI	CT	PET	EHR
Representation	[68, 230]	[42, 233]		[42, 199, 228, 206]
Fusion	[11, 132, 232, 207, 230, 164, 112, 228, 233]	[42, 132, 220, 22, 29]	[132]	[42, 220, 199, 232, 204, 108, 22, 41, 91]
Translation	[85, 78, 63, 174, 186]	[85, 236]	[78, 174]	
Alignment	[207, 230]	[220, 102, 233, 105]		[220, 105]
Co-learning	[201, 222, 78, 27, 77, 146, 83, 218, 48]	[218, 77, 146, 83, 48]		[216]

	Datatype				
Challenge	Hist.	Ultrasound	Genomic	X-Ray	Fundus
Representation				[199]	[233]
Fusion	[34, 108, 41]	[65]	[34, 41]	[199, 65, 41, 29, 91]	[233]
Translation					
Alignment	[102]				[233]
Co-learning		[217]	[31]		

2.5 Publication and Acknowledgement

This chapter has been accepted for publication in International Journal of Computer Vision [210]: Elisa Warner, Joonsang Lee, William Hsu, Tanveer Syeda-Mahmood, Charles Kahn, Olivier Gevaert, and Arvind Rao. Multimodal machine learning in image-based and clinical biomedicine: Survey and prospects, 2023.

CHAPTER 3

Multimodal Fusion in MRI: Low-Parameter Supervised Learning Models Can Discriminate Pseudoprogression and True Progression in Non-Perfusion-Based MRI

3.1 Abstract

Discrimination of Pseudoprogression (PsP) and True Progression (TP) is one challenge to the treatment of malignant gliomas. Although some techniques such as circulating tumor DNA (ctDNA) and Perfusion Weighted Imaging (PWI) demonstrate promise in distinguishing PsP from TP, we investigate robust and replicable alternatives to distinguish the two entities based on more widely-available media. In this study, we use low-parametric supervised learning techniques based on Geographically-Weighted Regression (GWR) to investigate the utility of both conventional MRI sequences as well as a diffusion-weighted sequence (apparent diffusion coefficient or ADC) in the discrimination of PsP v TP. GWR applied to MRI modality pairs is a unique approach for small sample sizes and is a novel approach in this arena. From our analysis, all modality pairs involving Apparent Diffusion Coefficient (ADC) maps, and those involving post-contrast T1-weighted (T1) regressed onto T2-weighted (T2) showed the best potential promise of predictability with all AUCs > 0.60. This work on ADC data adds to a growing body of research suggesting the predictive benefits of ADC, and suggests further research on the relationships between post-contrast T1 and T2.

3.2 Introduction

Glioma is a life-threatening condition characterized by neoplastic tumor growth in the brain. Malignant gliomas are primarily diagnosed based on the imaging features on conventional

MRI sequences. The primary approach to treatment of malignant gliomas is surgical resection within the limits of patients' safety, followed by radiation and temezolomide for six weeks. This treatment cycle is then followed by adjuvant temozolomide therapy for six more weeks. Because of the infiltrative nature of the tumors, surgical treatment is usually never curative since the tumors can extend well beyond the margins demonstrated by MR imaging. Therefore, follow-up imaging often shows tumor recurrence. True tumor recurrence needs to be differentiated from the condition of pseudoprogression which is defined as the appearance of a new lesion or increase in constrast-enhancing areas, but with changes that gradually fade or stabilize while treatment stays the same [108]. Since PsP resembles TP, treatment needs to be instituted at the earliest for the latter while waiting and watching is the approach to the former. Additionally, because of the similarity of these two conditions, patients are at risk of being started on alternative treatments prematurely or erroneously withdrawn from treatment altogether [117, 95]. Thus, there is a need to distinguish PsP from TP.

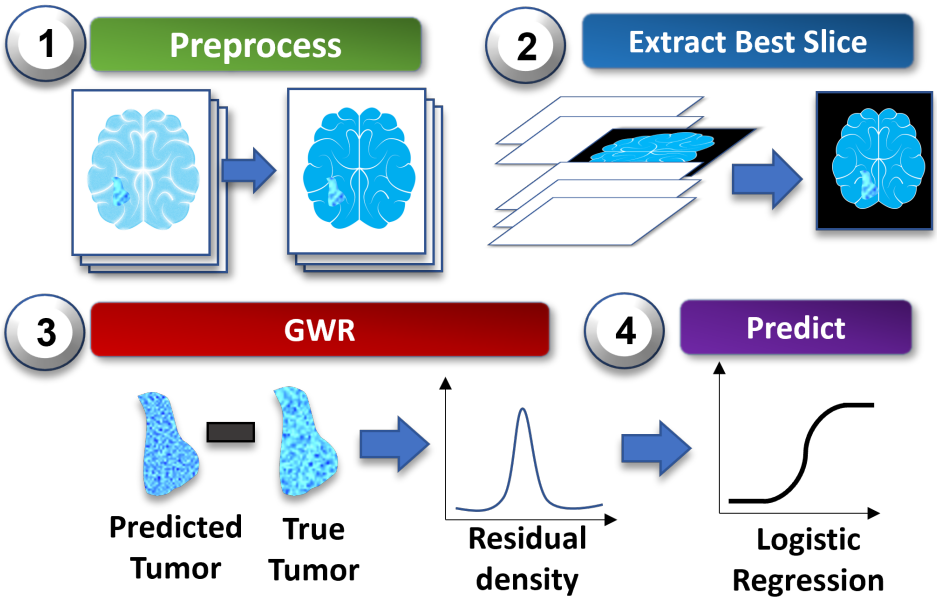

Figure 3.1: An illustrated depiction of the methods used in this study. First, (1) preprocessing on the MR images included registration and whitestripe normalization. Then, (2) a single slice is extracted from the MRI volume, and (3) GWR is applied to the tumor region of the extracted slice. Finally, (4) select characteristics of the residual density curve output from GWR are entered into a logistic regression model.

Previous studies have investigated possible methods for distinguishing PsP from TP using bloodstream biomarkers or medical imaging. In one method, evaluation of chromosomal instability using ctDNA reported promising results in PsP v TP distinction in a small trial group [62]. However, further analysis on larger cohorts is still needed, and ctDNA extraction

may not be available to all clinicians. Another common marker is PWI, which is current practice in some clinics [95]. However, this advanced imaging technique requires use of advanced software that are not widely available for all radiologists.

In the absence of perfusion imaging, some attempts have been made to exploit classical sequences of MRI in PsP studies using visual characteristics of the images determined by subject matter experts [225, 130], often finding these sequences to confer little help. A supervised machine learning approach using CNNs applied to conventional MRI sequences was able to exploit multimodal relationships between these MRI sequences for PsP distinction [100]. However, there are limitations to large multi-parametric models with small sample sizes because they carry potential to overfit the training data. A lower-complexity multimodal supervised model using geographically-weighted regression (GWR) for MRI by Mohammed et al [123] identified IDH-mutant 1p19q non-codeleted patients who exhibited a phenomenon called T2-FLAIR mismatch. This method leveraged relationships between the T2 modality and the FLAIR modality of MRI to successfully distinguish patients with T2-FLAIR mismatch vs those without.

In this study, we investigate the use of GWR to determine PsP presence in patients with increased tumor size after adjuvant therapy. We leverage this lower-complexity statistical approach followed by logistic regression and naive Bayes to explore multimodal relationships between pairs of conventional MRI sequences as a potential tool to discriminate PsP from TP. We also explore the potential role of an alternative advanced imaging technique to PWI called Diffusion Weighted Imaging (DWI) as a possible discriminator of PsP, given some preliminary interest in the modality [10, 71, 95]. Although conventional MRI is known to be of little help in PsP diagnosis [225, 130], we hypothesize that some modality pairs may show differential patterns which can discriminate PsP from TP better than random. Our use of GWR attempts to identify whether class differences exist in the relationships between multimodal pairs and additionally uses fewer features than is expected in deep learning strategies. This methodology is novel in the context of PsP and TP discrimination in treatment-related changes and provides an alternative image processing approach to deep learning methods for small sample sizes.

3.3 Methods

3.3.1 Data Acquisition

Data were collected in accordance with relevant guidelines and regulations and approved by the Institutional Review Board at the University of Michigan (IRBMED, HUM00145517).

Data were analyzed retrospectively and retrieved from the Electronic Medical Record Search Engine (EMERSE) and DataDirect databases [110]. In this study, five MRI sequences from fifty patients with high grade (World Health Organization (WHO) grade III or IV) diffuse infiltrating glioma (astrocytoma or oligodendroglioma) were recorded from 2009 to 2018. All patients received an adjuvant therapy followed by two to three recorded follow-up visits. At the last follow-up, histopathological analysis of the tumor site was conducted. MRI sequence data included in this study come from the follow-up directly prior to histological analysis of the tumor site. A label of "pseudoprogression" (PsP) or "true progression" (TP) was determined based on the results of the histopathological analysis.

Characteristic	**PsP**	**TP**	**Total**	**p-val**
Total	13	29	42	
Age (mean)	53.34	45.69		0.100
Sex				0.524
Male	9	17	26	
Female	4	12	16	
IDH status				0.627
wildtype	6	14	20	
mutant	5	8	13	
missing	2	7	9	
1p19q status				0.217
wildtype	0	3	3	
mutant	3	4	7	
missing	10	22	32	
EGFR status				N/A
wildtype	0	0	0	
mutant	4	5	9	
missing	9	24	33	
P53 status				0.590
wildtype	5	4	9	
mutant	12	6	18	
missing	12	3	15	

Table 3.1: Patient demographic table

3.3.2 Dataset and Preprocessing

A summary of Preprocessing to Prediction can be seen in Figure 3.1. The dataset contained 50 patient MRI scans with masks (PsP=13, TP=37) for the pre-contrast T1 (T1pre), post-contrast T1 (T1post), T2, and FLAIR modalities. However, after removing patients with zero-signal modalities, 42 patients (PsP=13, TP=29) were remaining. Zero-signal modality

primarily meant patients with zero FLAIR signal. An additional subtraction map modality was created by subtracting T1pre from T1post (T1postpre). Lastly, a DWI modality called apparent diffusion coefficient (ADC) was also utilized from each patient.

Patient MRI were registered and sized such that all modalities were the same size then normalized via whitestripe. No skull extraction was necessary for this procedure because only the area indicated by the tumor mask was analyzed.

In order to mitigate the dataset imbalance, PsP patients were oversampled. The most ideal slice for all patients was first chosen based on the largest mask slice using Matlab R2022b. Two additional slices above and below the largest mask slice were extracted from PsP patients, if they existed. Model input values were selected from each modality by overlaying the mask slice over the image and extracting values within the mask bounds. After oversampling PsP patients, 63 samples were available for all modalities (PsP = 34, TP = 29).

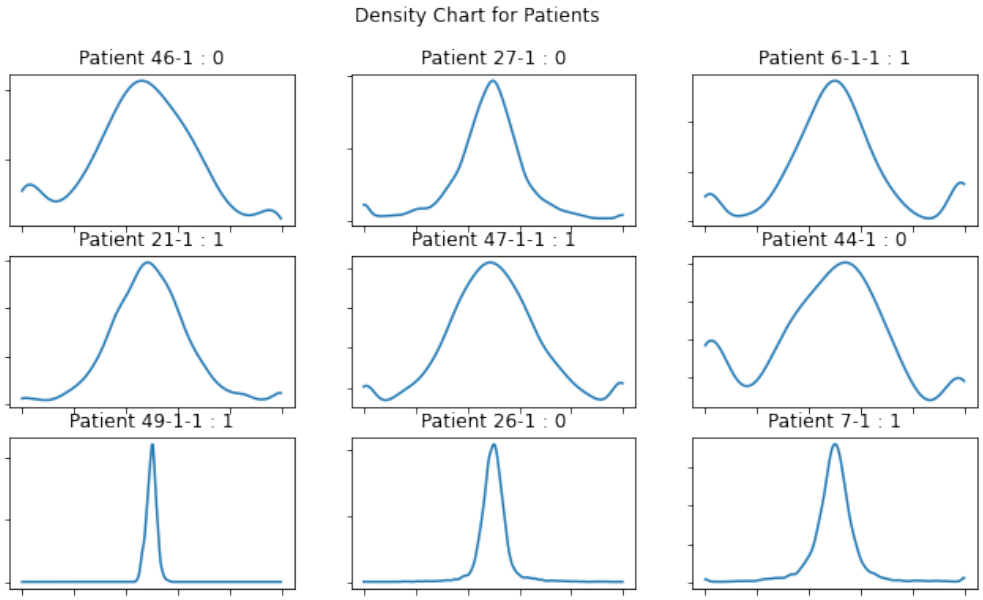

Figure 3.2: Residual density curves for individual patient MRI modalities of T2 regressed onto FLAIR. Next to the de-identified patient number at the top is the classification of the patient: 1 for pseudoprogression and 0 for true progression.

3.3.3 Model

We constructed a geographically-weighted regression (GWR) to analyze the relationship between pairs of MRI modalities. Based on the modalities T1pre, T1post, T2, FLAIR,

T1postpre and ADC, fifteen modality pairs were constructed. The predictor x of modality X of each pair was regressed onto y pixels from modality Y using the GWModel package in R(4.2.0) [59]. GWR is a form of linear regression that assumes regression weights will differ depending on the spatial locations of $x \in X$ and $y \in Y$. It therefore remains linear within specific localities i for all pixel locations x_i and y_i in a single MRI slice, such that

$$y_i = \beta_0 + \sum_M \beta_m x_{mi} + \epsilon_i \tag{3.1}$$

, where β_0 represents the bias at location i, M represents number of features, x_{mi} represents the m^{th} feature at i, and β_m represents the slope for feature m. In this study, the only feature analyzed was the pixel intensity at location i, so $M = 1$.

After fitting a model, residuals were calculated by subtracting the true y_i values from Y with the predicted $\hat{y}_i$ values from the model. Residuals were then flattened and densities were calculated into a density vector of 1×1000. Density curves for all patients were standardized to the same maximum and minimum bounds. Then, features from the curves were extracted as input into a logistic regression model using python 3.8 with Anaconda. In order to minimize the number of inputs into the regression model, we first shrunk the density to 1×500 by sampling every other value from the density curve.

There are two challenges with applying a residual density curve to logistic regression: 1) high dimensionality in a logistic regression model will encourage overfitting and will inappropriately overparameterize a model built on a small sample size, and 2) input expectations for logistic regression require independent and identically-distributed features. Features chosen which are too close together will be too colinear, perhaps biasing the model. In order to mitigate these two challenges, we address the first by sorting the regression vector by standard deviation and sorting in descending order, based on the assumption that areas of the curve with more overall variation will confer more information than those which fluctuate slightly. From these ordered areas of the curve, we select only the top k number of curve positions. In order to mitigate collinearity, we set a distance requirement between positions of the curve to discourage points from belonging to the same monotonic increase or decrease of a curve, and enforced an L2 regularization. Patient-wise LOOCV was utilized in the implementation and AUC was assessed for each test for each modality pair. Sensitivity and specificity metrics were also extracted for each test.

Lastly, we conduct a post-hoc test to assess the predictive power of combining multiple modalities, as [130] and [100] suggest that combining multiple predictive indicators may produce better predictability. Using a naive Bayes algorithm, we combined models for all predictive modality pairs (AUC$>$ 0.5) in a late fusion model. Modality pairs were added

into the model by their order of predictability in AUC from highest AUC to lowest. As in the above test, patient-wise LOOCV was utilized due to small sample size.

Pair	X	Y	AUC	Sens	Spec
1	T1postpre	FLAIR	0.5968	0.6923	0.4483
2	**T1postpre**	**T2**	**0.6180**	**0.6923**	**0.4828**
3	T1post	FLAIR	0.5889	0.6410	0.4483
4	T1post	T1postpre	0.5084	0.5641	0.4483
5	**T1post**	**T2**	**0.6145**	**0.6923**	**0.4828**
6	T1pre	FLAIR	0.5119	0.4615	0.6897
7	T1pre	T1post	0.4907	0.4615	0.5862
8	T1pre	T1postpre	0.4898	0.4615	0.5862
9	T1pre	T2	0.4598	0.4103	0.5862
10	T2	FLAIR	0.3917	0.1282	0.5862
11	**ADC**	**FLAIR**	**0.6684**	**0.6154**	**0.6552**
12	**ADC**	**T1post**	**0.6251**	**0.5385**	**0.5517**
13	**ADC**	**T1postpre**	**0.6260**	**0.5385**	**0.6207**
14	**ADC**	**T2**	**0.6525**	**0.6667**	**0.6207**
15	**ADC**	**T1pre**	**0.6242**	**0.5385**	**0.5862**

Table 3.2: AUC, Sensitivity and Specificity reported for detecting PsP from TP using Logistic Regression Analysis. Highlighted rows are for modality pairs where AUC$>$ 0.6

3.4 Results

Patient demographics can be viewed in Table 3.1. No significant differences in age, sex or any other clinical indicators were found. P-values for EGFR could not be calculated because no variation existed between PsP and TP sample populations.

Example images of residual density curves can be found in Figure 3.2. Density curves did not appear outwardly different based on class.

Results for the logistic regression assessment can be viewed in Table 3.2. We tested k values of 1,2,3,5,10, but found decreasing performance as the number of features increased past $k = 3$, signifying overfitting. All pairs involving ADC (pairs 11-15) demonstrated AUC values above 0.6 in our supervised model in the test set, with ADC regressed on FLAIR as the highest performing pair. Interestingly, two conventional MRI pairs also performed above 0.6 AUC: T1postpre regressed on T2 and T1post regressed on T2. Pairs 7-10 in Table 3.2 demonstrated values below 0.50, indicating that the signal for these were too weak given the sample size.

In order to understand why relationships between distributions resulted in relatively low

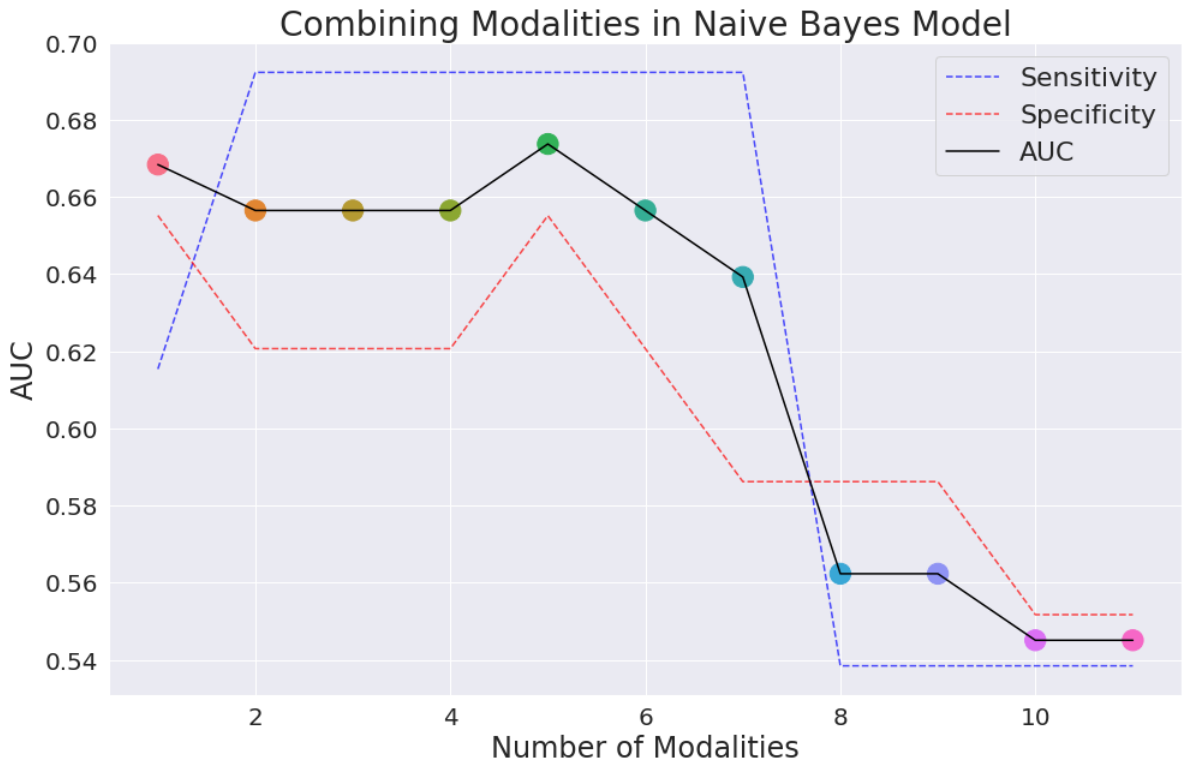

Figure 3.3: Results from naive Bayes late fusion model combining multiple modality pairs to distinguish PsP from TP. The best model performs at an AUC of 0.6737 and includes all ADC modality pairs.

AUCs, mean residual densities for the PsP and TP class for a single fail case modality pair were assessed in Figure 3.4. Marks in red signify the features chosen among the curve. PsP and TP residuals demonstrated significant overlap, suggesting that the X modality was well-conditioned to predict the Y modality in both PsP and TP cases, an indication that there is very little difference between the PsP and TP images in most fail cases.

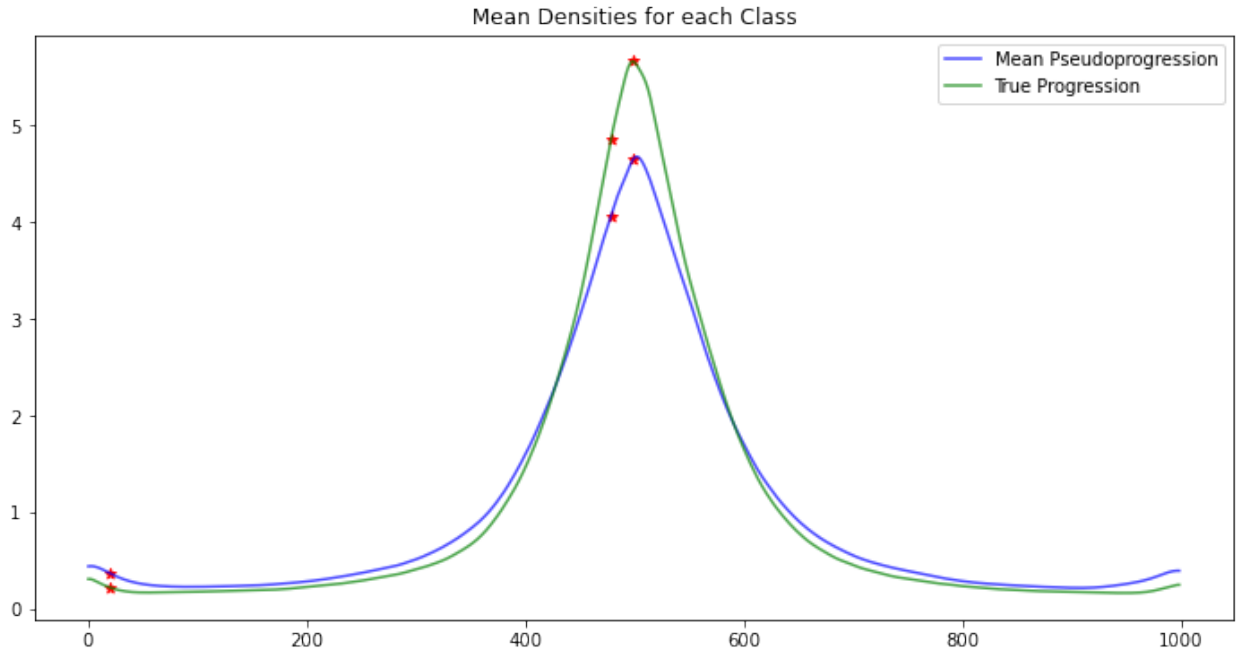

Figure 3.4: An example representation of mean densities for each class for T1 regressed onto T1postpre. The three red stars for each curve represent the locations of the top 3 PDF locations of the curve as features. This figure illustrates the difficulties of distinguishing PsP from TP, by lending evidence that even computationally, the PsP and TP images are nearly the same.

Results from the naive Bayes model are shown in Figure 3.3. In this analysis, the highest performing model was the result of a combination of the top 5 modality pairs, which included only those modality pairs with ADC as a predictor. Interestingly, after including the contribution of the conventional MRI sequences, the model performance decreases. Models including modality pairs beyond the 5 ADC modality pairs, T1post regressed on T2 and T1postpre regressed on T2, performed substantially worse.

3.5 Discussion and Conclusion

Distinguishing pseudoprogression, a result of adjuvant therapy, from true progression, has been a classically difficult challenge, particularly using conventional MRI sequences. How-

ever, there are benefits of being able to use conventional MRI to distinguish PsP in High-Grade Glioma (HGG) patients, because MRI is routinely used for analysis of glial tumors. In this study, we attempted to assess the predictive power of both conventional MRI sequences and one advanced imaging technique (DWI) at distinguishing PsP v TP using low-complexity models for smaller sample sizes and pathologist-confirmed data regarding PsP diagnosis. A set of 15 modality-pairs were investigated using GWR followed by logistic regression to analyze if correlations between imaging modalities could reap discriminative patterns. Most modality pairs did not reap AUCs above 0.6, indicating no unique patterns of pixel intensities in the images which could be discriminatory. This is corroborated by Figure 3.4, where distributions show an almost complete analysis. However, in the analysis, all modality pairs involving ADC demonstrated marked discriminative ability with AUCs above 0.6, and T1postpre regressed on T2 as well as T1post regressed on T2 also showed some promise in one analysis.

Our results on ADC's discriminative ability contribute to a small body of growing research which have also found a utility in diffusion weighted imaging. Two different groups found differences in ADC maps between recurrent and non-recurrent gliomas after radiotherapy using samples of 18 patients and 17 patients, respectively [10, 71, 95]. Thus, our study is the largest to assess ADC as a tool for recurrence (TP) v non-recurrence (PsP) in HGG patients to our knowledge. However, we acknowledge the limitations of our study due to small sample size and believe further studies with larger samples are needed to fully confirm the utility of ADC. Our study also stands as the first to find correlations involving T1post, T1pre and T2 as potentially helpful in distinguishing PsP v TP.

This study carries two main novelties. Firstly, although discrimination of PsP v TP is challenging with conventional MRI, our method found that relationships in some sequences, specifically DWI modalities such as ADC and possibly T1-post or T1postpre regressed on T2, carry some predictive power. Lastly, our method demonstrates the successful use of low-complexity models on conventional MRI sequences as an alternative to deep learning techniques in small sample sizes. The method leverages relationships between multiple modality pairs and can be used for supervised image analysis when deep learning applications are deemed inappropriate.

3.6 Publication and Acknowledgement

This chapter is a published work [211]: Elisa Warner, Joonsang Lee, Santhoshi Krishnan, Nicholas Wang, Shariq Mohammed, Ashok Srinivasan, Jayapalli Bapuraj, and Arvind Rao. Low-parameter supervised learning models can discriminate pseudoprogression and true pro-

gression in non-perfusion-based mri. In *2023 45th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*. IEEE, July 2023.

CHAPTER 4

Predicting Osteoarthritis of the Temporomandibular Joint using Random Forest with Privileged Information

4.1 Abstract

TMJ OA is the most common disorder of the Temporomandibular Joint (TMJ). A CDS system designed to detect TMJ OA could function as a useful screening tool as part of regular check-ups to detect early onset. This study implements a CDS privileged learning concept model based on Random Forest and dubbed RF$^+$ to predict TMJ OA. Our RF+ model was based on the hypothesis that a model which leverages high-resolution radiological and biomarker data as privileged features in training but not testing can improve upon a baseline model which only includes features from a questionnaire in training and testing. Under leave-one-out cross validation, RF+ predicted TMJ OA with an AUC of 0.6798 compared to the baseline model (AUC: 0.6518). Additionally, we introduce a novel method for post-hoc feature analysis, finding several features of the lateral condyles and joint distance to be the most important features from the privileged modalities for predicting TMJ OA.

4.2 Introduction

The TMJ plays an essential role in mouth movement and consists of a complex system of bone, cartilage and muscle. Osteoarthritis of the TMJ (TMJ OA), a degenerative disease which affects all structures therein, is the most common disorder of the TMJ [187]. Observations from radiological images show TMJ OA is associated with flattening or deformation of the lateral condyles, reduction of joint space, and possible alterations to the articular fossa region [171, 154]. Although prevalence of TMJ OA has been difficult to calculate, post-mortem analysis of modern bone collections have found a 30.2% prevalence among modern

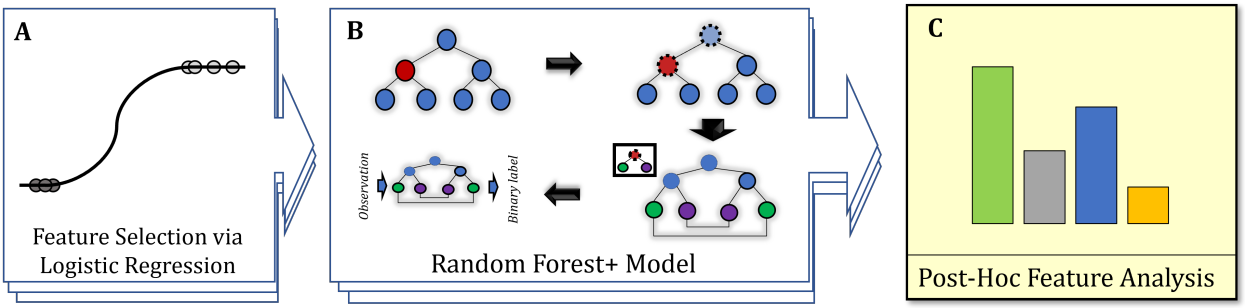

Figure 4.1: Workflow for the reported study. In this study, we utilize LOOCV on a sample of 97 patients. For each fold, a feature selection process consisting of Logistic Regression is computed (A), and then a Random Forest$^+$ model is constructed based on the selected features (B). After all folds have been calculated, a post-hoc analysis is conducted to determine the most important privileged and non-privileged features for tree-based transforms.

humans [154], with 40 to 75% of the population reporting at least one symptom of overall disorders of the TMJ (TMD) [171]. TMJ OA falls under the umbrella of osteoarthritis, which is the second most prevalent musculoskeletal disorder behind lower back pain, occurring with a global incidence of nearly 15,000 per year[205].

Recently, CDS models have made waves in the medical community, assisting in diagnosis of a wide range of conditions [193, 4, 155]. Although CDS models cannot replace the need for experienced dental experts, a CDS system designed to detect TMJ OA could function as a useful screening tool as part of regular check-ups, with the goal of detecting early TMJ OA and thus permitting dental experts to initiate treatment and preventive behavioral strategies to decelerate degradation of the TMJ at an early stage.

While clinical questionnaires designed to screen for TMD may help screen for TMJ OA, we hypothesize that including radiological imaging information from the TMJ site as well as protein biomarkers collected from serum/saliva could provide additional information which may be useful for discriminating TMJ OA patients from healthy patients. Studies analyzing protein biomarkers and radiological information in TMJ OA patients have already asserted the predictive utility of these features [229, 23].

However, although radiological imaging and protein biomarkers could be useful additions to a TMJ OA CDS model, it is not reasonable to expect that most clinics would be able to provide such data, as high-resolution Cone-beam Computed Tomography (CBCT) scans of the articular fossa and lateral condyle regions of the TMJ as well as protein microarrays of human serum and saliva samples are more common in research rather than clinical practice. Since typical predictive models require all modalities to be present with no missing data, multimodal co-learning strategies must be explored.

One such strategy incorporating privileged information was developed as a part of a concept called "knowledge transfer." [200]. In knowledge transfer models, a "privileged" modality of data exists in the model solely as a "teacher," providing information which assists the "student" model solely during the training phase, while disappearing in the test phase. With proper knowledge transfer, the final student model should perform more accurately with the assistance of the privileged information during training than without. In this study, we consider multimodal models which incorporate privileged information, where clinical features will be considered non-privileged information available in training and testing and radiological and biomarker features will be considered privileged information available in the training set only. This will allow the latter, rarer modalities to still assist the model while only requiring basic clinical questionnaire information at test-time, thus generalizing such a decision support model to a larger audience.

The most common privileged learning frameworks are based on Artificial Neural Network (ANN) or Support Vector Machine (SVM) frameworks [200, 32, 229, 78]. However, these models work best under very specific conditions. ANNs are primarily useful with large data samples and features, but considered largely inappropriate for smaller datasets due to the scale of trained parameters required. The well-known Support Vector Machine for Privileged Learning (SVM^+) model, a framework of SVM designed specifically to incorporate privileged information, can be problematic because the privileged modality functions as an error corrector in the model. This means that the privileged modality must provide discriminatory capabilities equivalent to a gold standard, or risk introduction of erroneous error corrections, thus reducing AUCs of the student model. Although some models such as [167] have attempted modifications of the SVM^+ algorithm to improve upon this shortcoming, such models are not widely available and come with large computational overhead. In another model, [128] developed a Random Forest model which incorporates privileged in-

Table 4.1: Patient Clinical and Demographic Data.

Feature	TMJ OA	Control	p-value
Total Sample (N=)	49	48	N/A
Gender	1.14	1.15	0.9672
Female (N=)	42	41	N/A
Male (N=)	7	7	N/A
Age	40.20	38.71	0.5730
Headaches [0,4]	1.59	0.60	0.0000
Muscle Soreness [0,4]	1.06	0.38	0.0004
Vertical Range Unassisted w/o Pain	36.08	44.94	0.0001
Restless Sleep [0,4]	1.29	0.58	0.0016

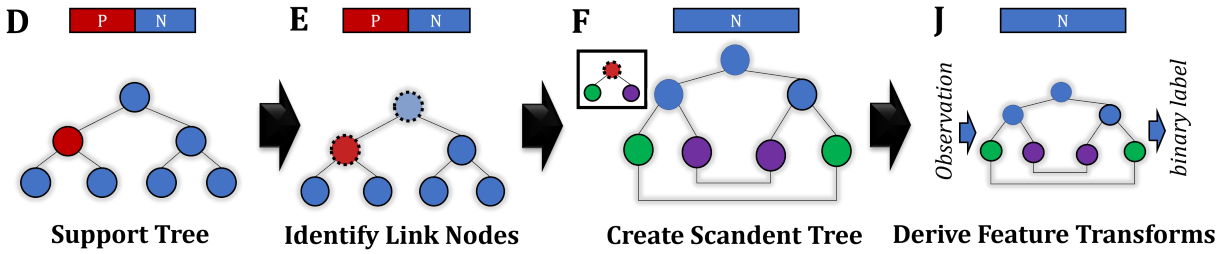

Figure 4.2: Workflow of the RF$^+$ framework using tree-based feature transforms. The top bar of the figure indicates the feature space used (N or ($N \bigcup P$)).

formation through the construction of "tree-based feature transforms." The authors claimed that their model can perform at least as well as a non-privileged model, even in the case of substandard privileged information, because of the Random Forest's unique ability to select best features from a given feature bag.

This study implements a CDS concept model based on the framework from [128] for predicting TMJ OA vs healthy controls, with the hypothesis that a model which leverages our available high-resolution radiological and biomarker data in training can improve predictions compared with a baseline model which requires only clinical features in testing. We further expand the work of [128] by introducing a novel method for post-hoc feature analysis, tracing back the most important features for prediction among both privileged and non-privileged feature sets.

4.3 Methods

4.3.1 Data Acquisition and Preparation

Our dataset consisted of 51 early-stage TMJ OA patients and 50 healthy controls recruited at the University of Michigan. All the diagnoses were confirmed by a Temporomandibular Disorders (TMD) and orofacial pain specialist following the Diagnostic Criteria for Temporomandibular Disorders (DC/TMD) [170]. The clinical, biological and radiographic data described below were collected from TMJ OA and control subjects with informed consent and following the guidelines of the Institutional Review Board HUM00113199.

Details on the dataset can be found in [23]. Briefly, the clinical dataset was collected following DC/TMD criteria. The biological data comprised of proteins that were previously correlated with arthritis initiation, progression and bone morphological alterations [30]. Using customized protein microarrays (RayBiotech, Inc. Norcross, GA), the expression level of 13 proteins was measured in the participants' saliva and serum samples, respectively.

The radiological data was collected from CBCT scans taken using 3D Accuitomo machine (J. Morita MFG. CORP Tokyo, Japan). It consisted of 3D superior condylar-to-fossa joint space measurements and radiomic features. Using BoneTexture module from 3D-Slicer software (www.3Dslicer.org), 43 radiomic features were attained following a standardized protocol reported by Bianchi et al [24].

Of the 101 patients obtained, four were removed due to missing data, resulting in a final sample size of 97 patients. Features were split into "privileged" and "non-privileged" information based on their probable availability in a real-world clinical setting. Due to the greater difficulty of obtaining high-resolution CBCT scans and microarray biological samples in a clinical setting, we classified these modalities as privileged information while the clinical data were marked as non-privileged features. In total, 68 privileged features and six non-privileged features were included in the dataset.

4.3.2 Model Construction

The primary model utilized in this study, which here is dubbed "RF$^+$", is based on the tree-based feature transforms framework from [128] and illustrated in Figure Fig 4.2. In our RF$^+$ model, a Random Forest model called the "support forest" consisting of K decision trees is first constructed based on both privileged features ($\{P\}$) and non-privileged features ($\{N\}$) in the training set only (Fig 4.2D). After the support forest is constructed, a simple algorithm searches through all nodes of each tree $t_k, k = 1 \dots K$ to identify nodes of interest called "link nodes" (Fig 4.2E). In order to qualify as a link node, any node n_i^k from tree k must satisfy at one of the following criteria:

1. Node n_i^k is a root node
2. Node n_i^k has a parent n_{i-1}^k with a node feature $f_{i-1}^k \in N$ and n_i^k has a node feature $f_i^k \in P$
3. Node n_i^k has a parent n_{i-1}^k with a node feature $f_{i-1}^k \in P$ and n_i^k has a node feature $f_i^k \in N$.

For each link node, the observations at the left and right children of the node are annotated as "0" and "1", respectively. Then, these labels are utilized to train a "scandent tree" for each link node, which attempts to replicate the discriminative power of the link node utilizing only non-privileged features (Fig 4.2F).

After all scandent trees are formulated, "tree-based feature transforms" are constructed for each data observation based on the label assigned by each scandent tree. Therefore, if

z link nodes are discovered, then z scandent trees are formulated, resulting in z number of binary-labeled tree-based feature transforms (Fig 4.2J). Then, a final model is formulated based on the non-privileged features and tree-based features only. Since the scandent trees are also based only on non-privileged features, no privileged features are required in testing.

4.3.3 Cross Validation and Evaluation

Two types of cross validation were utilized in this study. The first was LOOCV, due to its ability to demonstrate fullest use of the training data in a single run. In order to provide a more robust study, we also incorporated a second validation method consisting of 400 times random bootstrapping of 15% of the dataset. Because this method is essentially Out-of-Bag (OOB) sampling for Random Forest models, we denote this validation method with the acronym OOB from here onward.

For comparative analysis, four additional models were constructed: 1) one consisting of only privileged features, 2) one consisting of both non-privileged and privileged features, 3) one with only tree-based features, 4) the Baseline model, consisting of only non-privileged features. All models were evaluated for AUC for both LOOCV and OOB validation methods, and standard error was calculated. For OOB, mean AUC and mean standard error were calculated, respectively.

4.3.4 Post-Hoc Feature Analysis

Finally, after all models were run, a post-hoc feature analysis was performed on the tree-based feature transforms (example shown in Fig 4.6. For each tree-based feature transform, we traced back the link node from which it was based and analyzed the node feature at that link node. We then totaled up the the frequency with which each feature appeared as a node feature for a link node. Based on the definition of a link node, we decided to distinguish a feature at a link node as a "Root" feature if the node feature appeared at a link node defined by criteria 1 for identifying link nodes (See Section 4.3.2). This is because criteria 2 and 3 for defining link nodes are based on node features of a node given the node feature of a parent node. Thus, although our feature analysis identifies a specific feature at a link node, for non-"Root" features, the scandent tree formed for the link node listed will try to replicate the discriminatory ability of the node feature at the link node given settings of previous node features. Scandent trees from "Root" features, by contrast, will try to replicate the discriminatory ability of the node feature only.

4.3.5 Implementation

Due to the large number of privileged features, some of which may not be important, a univariate logistic regression to predict TMJ OA was run on the training set for each fold *before* initiating the RF$^+$ model workflow. Namely, only privileged features were analyzed by logistic regression, and privileged features with an AUC > 0.55 were included in the RF$^+$ model. Because there were only six variables included in the non-privileged feature set, all non-privileged variables were included for all folds.

The model implementation from [128] was preserved in our work. Namely, a feature bagging size of sqrt(num of features) was implemented, and the entire training set was utilized in the construction of the scandent trees. In order to reduce the number of unimportant tree-based features, we implemented a feature importance calculation on the training set immediately after construction of the features and features with an importance score of 0 were eliminated. Due to the large imbalance of non-privileged and tree-based feature transforms, we force equal sampling of each feature set to construct each tree in the final forest. We set max depth equal to 7 and number of trees equal to 100.

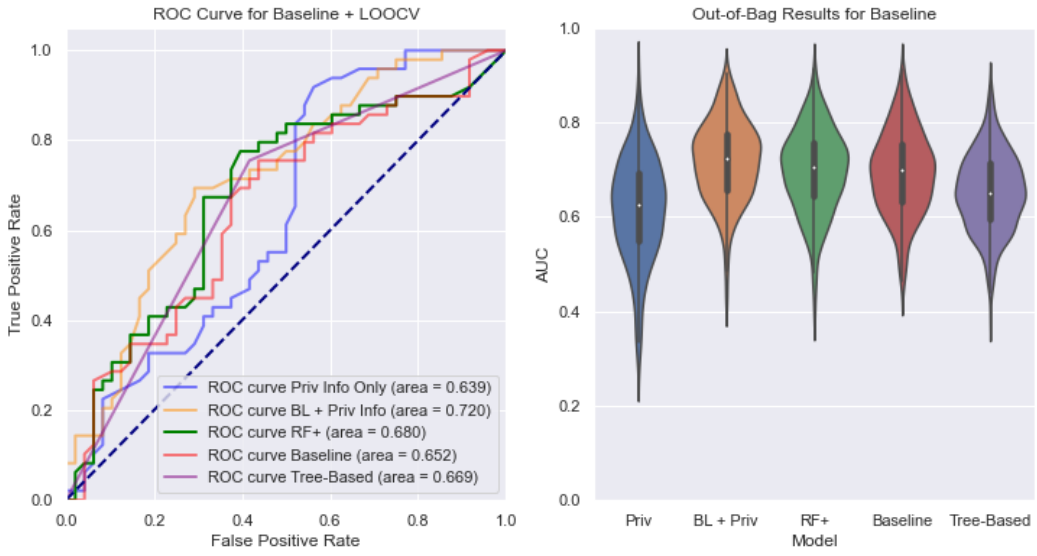

Figure 4.3: (*left*) ROC curves for RF$^+$ and comparative models using Leave-One-Out Cross Validation; (*right*) Violin plots illustrating the distribution of AUCs for Out-of-Bag validation tests

4.4 Results and Discussion

4.4.1 Dataset Analysis

A patient demographic data table with sample sizes and summaries of the baseline data can be shown in Table 4.1. As expected, clinical questions for TMD demonstrated discriminative ability to discern TMJ OA from healthy control patients. Averages for privileged information were omitted to save space.

Feature correlations are viewed in 4.5. It appears that clinical markers, lateral condyle imaging markers, lateral (articular) fossa imaging markers, and serum and saliva biomarkers tend to correlate more with themselves than with other groups of markers. However, articular fossa and lateral condyle demonstrate some overlapping correlations with each other.

4.4.2 Feature Selection Analysis

Results from the feature selection using univariate Logistic Regression are shown in Table 4.3 ranked by the percent of folds in which the features were included as well as the average AUC across all 97 folds. Included in the table was also the performance of non-privileged variables.

Table 4.2: Model Comparison Results

Model	LOO AUC	LOO stderr	OOB AUC	OOB stderr
Privileged Only	0.6390	0.0252	0.6163	0.0513
Baseline+Privileged	0.7198	0.0267	0.7184	0.0530
RF$^+$	**0.6798**	**0.0306**	**0.6974**	0.0602
Tree-Based Only	0.6692	0.0477	0.6535	0.0939
Baseline	0.6518	0.0309	0.6940	**0.0590**

4.4.3 Model Results

AUCs and their respective standard errors for each tested model are shown in Table 4.2 and Fig 4.3. The top box (top two models) consists of models in which privileged features are included in the test set, while the bottom box (bottom three models) consists of models in which only non-privileged features are included in the test set. The proposed RF$^+$ model outperformed both the Baseline model as well as the model based on tree-based feature transforms alone. Interestingly, while the Baseline+Privileged model (which incorporates privileged features during testing) outperforms all other models as expected, the Privileged

Only model performs lower than expected, even when a Logistic Regression is used for feature selection. This may indicate that although radiomic images are useful for detecting TMJ OA, the extracted features themselves may not be a better screen of TMJ OA compared to a simple clinical questionnaire, but when combined with the clinical questions, can provide some supplementary information.

The improved performance of the Tree-Based Only model over the Baseline model demonstrates the potential for tree-based feature transforms to mimic the predictive power of privileged features with only non-privileged features, and suggests that with only six non-privileged features, this model can still coax out interesting non-linear relationships between existing features that were not easily ascertained otherwise.

Lastly, the performance of the RF^+ model is interesting in that it can improve the baseline model, even where privileged features are not a "gold standard" source of information, confirming the advantages of this model stated in [128]. In privileged learning models where privileged information is utilized as an "error corrector" [200], privileged features must be close to gold standard quality in order to prevent introduction of erroneous error corrections to a non-privileged model. However, with tree-based transforms, when privileged information is poor, a decision tree can choose tree-based transforms which originate from a non-privileged root node if it outperforms those originating from privileged root nodes. Thus, the RF^+ can leverage the discriminative capabilities of privileged features, while downplaying weaknesses of the features.

4.4.4 Feature Importance Based on Tree-Based Feature Transforms

Feature importance for the top 20 features is shown in Fig 4.4. Frequencies were rescaled into a score in range $[0, 1]$ by dividing all feature frequencies by the total number of feature appearances. The top features were Vertical Range Unassisted w/o Pain, which is a clinical feature whereby a patient is asked to open their mouth to the fullest range before pain is felt. The most important privileged features were shortRunHighGreyLevelEmphasis of the lateral condyle and 3D_JS_SI (joint distances). Of the top 10 unique privileged features which ranked highest using this method, eight also appeared in the top 10 most predictive privileged features from the Logistic Regression rankings in Table 4.3.

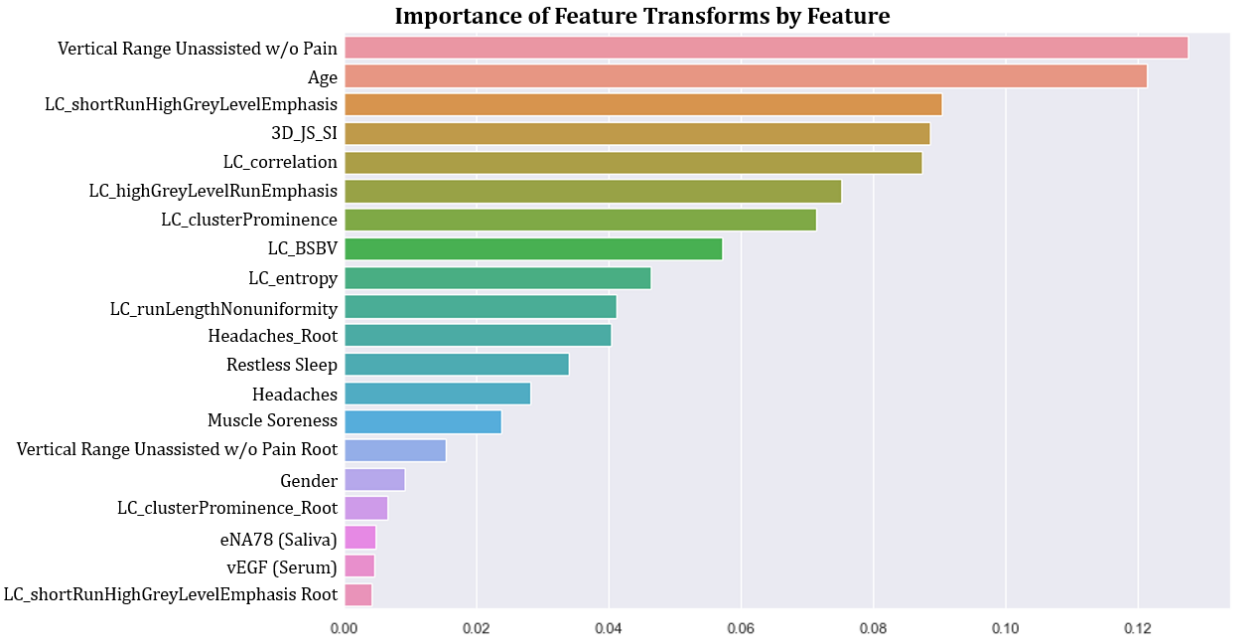

Figure 4.4: Top 20 features derived from tree-based feature transforms and their respective importance scores.

4.5 Conclusion

In this study we implemented an RF$^+$ CDS concept model based on tree-based feature transforms to detect TMJ OA in 97 patients. We incorporate two modalities of privileged information, namely radiological imaging features and biomarker protein data, and one set of non-privileged information consisting of clinical questionnaire data. We demonstrated that our proposed RF$^+$ model outperforms the baseline model, even though both models use only non-privileged information at test time. Furthermore, we expand upon the RF$^+$ model framework to incorporate our own feature importance scores based on appearance of link node features among the most popular tree-based features in the RF$^+$ framework. We show that tree-based feature transforms identify some of the most discriminative features of the dataset and sufficiently replicate their discriminatory capabilities with non-privileged clinical features alone. This work demonstrates both the usefulness of RF$^+$ in predicting TMJ OA and elucidates benefits of incorporating research-obtained information that is not normally obtained clinically as a means to improve upon CDS models.

Table 4.3: Top Features Selected Using Logistic Regression

Rank	Priv/Non-Priv	Feature	% Folds	Avg AUC
1	P	3D_JS_SI	37.11	0.6148
2	P	LC_correlation	32.99	0.6000
3	P	LC_entropy	30.93	0.5833
4	P	LC_shortRunHighGreyLevelEmphasis	30.93	0.6242
5	P	LC_longRunEmphasis	29.90	0.5285
6	P	LC_highGreyLevelRunEmphasis	29.90	0.5906
7	P	LC_clusterProminence	28.87	0.6378
8	P	LC_runLengthNonuniformity	28.87	0.5485
9	P	LF_correlation	22.68	0.4643
10	P	LC_BSBV	21.65	0.5132
1	N	Vertical Range Unassisted w/o Pain	100*	0.6867
2	N	Headaches	100*	0.6709
3	N	Restless Sleep	100*	0.5331
4	N	Muscle Soreness	100*	0.5136
5	N	Age	100*	0.3665
6	N	Gender	100*	0.000

4.5.1 Publication and Acknowledgements

This chapter is a published work [209]: Elisa Warner, Najla Al-Turkestani, Jonas Bianchi, Marcela Lima Gurgel, Lucia Cevidanes, and Arvind Rao. Predicting osteoarthritis of the temporomandibular joint using random forest with privileged information. In *Ethical and Philosophical Issues in Medical Imaging, Multimodal Learning and Fusion Across Scales for Clinical Decision Support, and Topological Data Analysis for Biomedical Imaging: 1st International Workshop, EPIMI 2022, 12th International Workshop, ML-CDS 2022, 2nd International Workshop, TDA4BiomedicalImaging, Held in Conjunction with MICCAI 2022, Singapore, September 18–22, 2022, Proceedings, pages 77–86.* Springer Nature Switzerland, Basel, 2022.

4.6 Supplementary Material

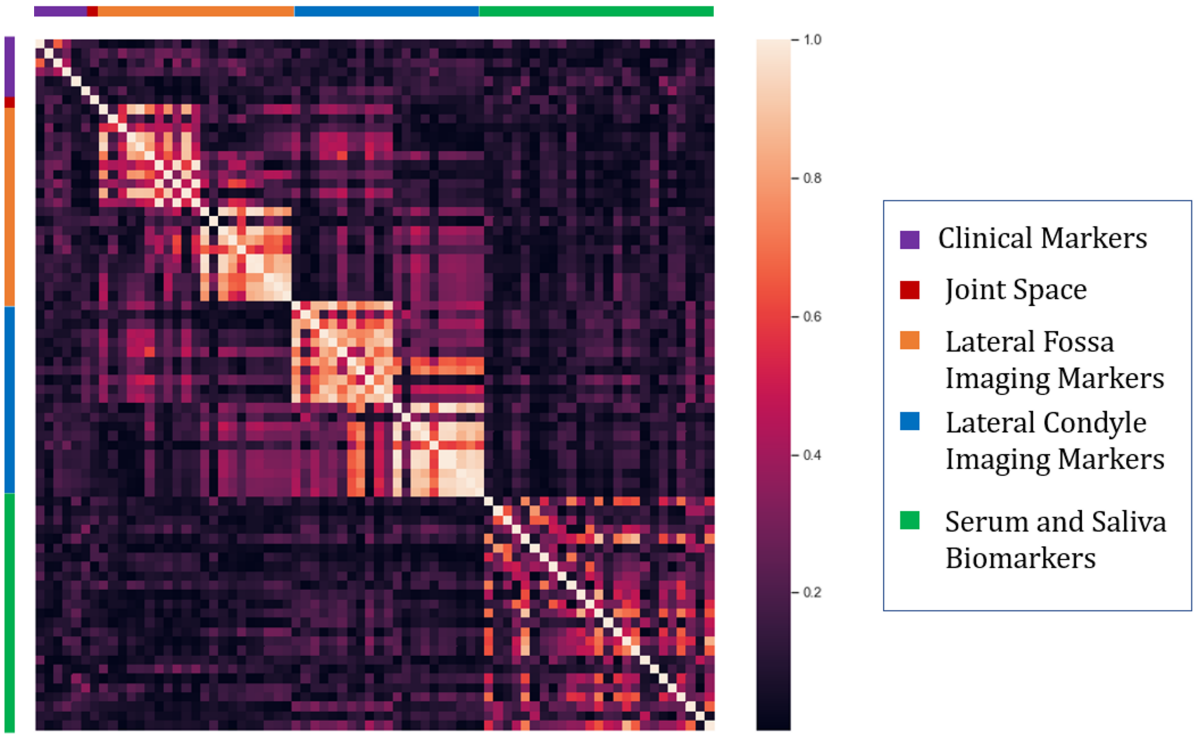

Figure 4.5: Correlation of Non-Privileged (Clinical Markers) and Privileged (all else) Features

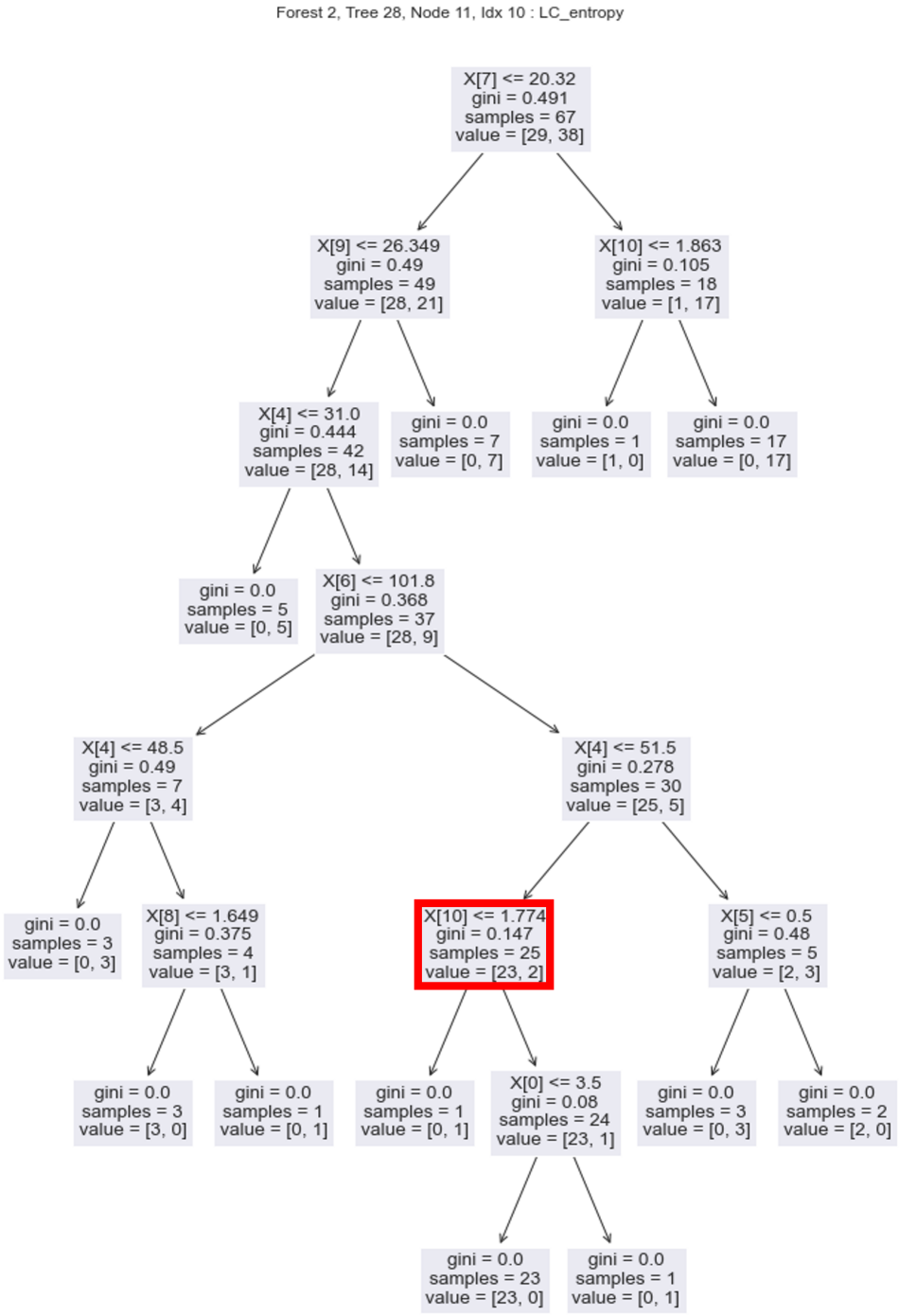

Figure 4.6: Example of following a link node back to the support tree to identify the feature at the link node. In this example, scandent tree features were built from a link node from the 28th tree of the 2nd support forest at the 11th node in the tree. The node feature at the node was the 10th index of the feature bag list for that tree, which was LC_Entropy, which is entropy of the lateral condyles in the CBCT image.

CHAPTER 5

IOL Prediction Models, Part I: Comparison of Cataract Surgery Patients in South Indian and Midwestern United States Populations

5.1 Abstract

Cataracts cause visual impairment or blindness in over 94 million people worldwide and are the most common cause of blindness in the world. In this study, we compared preoperative clinical and biometric measurements of patients undergoing cataract surgery across two populations and evaluate differences in refraction prediction accuracy in these populations. We obtained perioperative cataract surgery data from both Aravind Eye Hospital in Chennai, Tamil Nadu (Aravind) and the University of Michigan in Ann Arbor, Michigan (UMich). The study comprised 2729 eyes from Aravind (mean age at surgery 60.2 years ± 9.5 [SD]) and 1003 eyes from UMich (mean age at surgery 70.7 years ± 9.5 [SD]). The Aravind group demonstrated significantly lower ages at surgery, Axial Length (AL), Lens Thickness (LT), and Central Corneal Thickness (CCT), while the UMich group demonstrated lower K measurements, IOL power, and post-operative refraction compared to the Aravind group. In IOL formula assessment of the SN60WF Aravind group, Haigis, Hoffer Q, and Barrett had the worst performance in terms of error within 0.25D (46.90%, 47.79% and 51.33%, respectively), while the Nallasamy formula, SRK/T and Holladay1 performed the best (61.50%, 56.19%, 54.52%, respectively). This study illustrates significant differences in cataract patients from South Indian and Midwestern US populations. Differences also emerge in the distribution of errors in IOL formula predictions. Understanding population-level differences and designing better methods for integration of these factors into IOL formulas may help improve refractive surgery outcomes.

5.2 Introduction

Cataracts cause visual impairment or blindness in over 94 million people worldwide and are the most common cause of blindness in the world [97]. It is characterized by a gradual loss of transparency in the crystalline lens of the eye which results in loss of visual acuity.

Treatment of cataract is typically conducted through surgical replacement of the intraocular lens. This replacement requires implantation of a synthetic lens, usually constructed of collamer, acrylic, silicone, or polymethyl-methacrylate (PMMA)[195]. In order to determine the appropriate power of the lens, IOL diopter is calculated through a formula which takes into account various measures of a patient's eye, such as axial length (AL), anterior chamber depth (ACD) and keratometry (K). Since no model of the eye confers perfect diopter predictions, multiple IOL power calculation formulas exist, among them Barrett Universal II, SRK/T, HofferQ, Haigis and Halloday 1 being some of the most well-known [17, 161, 74, 67].

The overwhelming majority of both cataract and synthetic intraocular lens research is conducted on populations in the United States, Europe and Australia, where the majority of patients consume western diets and cultural influences. However, differences in diet, sun exposure (specifically UVB exposure), disease prevalence and socioeconomic status have been indicated as possible determinants of cataract onset and differs by population and regional demographic [188, 143, 84]. Some of these differences may be evident in the standard eye measurements obtained preoperatively in intraocular lens surgery, meaning that these cultural and genetic factors could alter the physiological characteristics of the eye. These differing measurements can further lead to inaccuracies in IOL formulas. For example, in one study of a Japanese population, it was discovered that ALs and AL/Corneal Radius (CR) ratios were longer in Japan than other countries compared. It was simultaneously discovered that longer ALs and AL/CR ratios resulted in higher error rates using Holladay 1 and Hoffer Q formulas compared with average measures [138]. Understanding if there are differences in measurements between different regional population demographics would be useful in assessing and customizing the need for public health preventive measures in different international demographics. With respect to IOL power calculations, many of whose constants were based on datasets from western populations, it may help elucidate if the model parameters for different IOL formulas hold true across populations.

In this study, we explore two obtained datasets from different regional populations, one from the Midwestern United States, and another from South India, representing differences in race, diet, sun exposure, and quality of life, among others, analyzing both the distributions of eye measurements as well as IOL predictions from a variety of formulas.

	Aravind Dataset				UMich Dataset			
	Female	Male	Total	p	Female	Male	Total	p
Count (N)	1427	1360	2787	N/A	570	433	1003	N/A
Laterality (R)	775	742	1517	0.895	283	229	512	0.310
Age at Surgery	59.15	61.30	60.20	< 0.001	70.92	70.48	70.73	0.467
AL (mm)	22.82	23.33	23.07	< 0.001	23.93	24.44	24.15	< 0.001
CCT (μm)	520.38	525.86	523.05	< 0.001	550.17	555.74	552.58	0.015
ACD (mm)	3.23	3.34	3.29	< 0.001	3.19	3.32	3.25	< 0.001
AD (mm)	2.71	2.82	2.76	< 0.001	2.64	2.77	2.69	< 0.001
LT (mm)	4.20	4.23	4.22	0.090	4.52	4.53	4.53	0.947
K1 (D)	44.60	43.98	44.30	< 0.001	43.74	43.06	43.45	< 0.001
K2 (D)	45.32	44.69	45.01	< 0.001	44.59	43.97	44.33	< 0.001
Km (D)	44.96	44.34	44.66	< 0.001	44.16	43.51	43.88	< 0.001
Astigmatism	0.72	0.71	0.71	0.745	0.85	0.91	0.88	0.259
WTW (mm)	11.65	11.85	11.75	< 0.001	12.01	12.20	12.09	< 0.001
IOL power (D)	21.31	20.47	20.90	< 0.001	20.27	19.37	19.89	< 0.001
refraction (D)	-0.05	0.04	-0.01	< 0.001	-0.64	-0.51	-0.594	0.021

Table 5.1: Demographic Table

5.3 Methods

5.3.1 Data Collection

The study was approved by the Indian Health Service Institutional Review Board (RET202100362) and the by the Institutional Review Board at the University of Michigan (HUM00160950). Due to the retrospective and de-identified nature of the data utilized, it was determined by the institutional review boards that informed consent was not required. All research was carried out in accordance with the Declaration of Helsinki.

Data from the South Indian population ("Aravind") were collected from patients undergoing cataract surgery at Aravind Eye Hospital in Chennai, Tamil Nadu, India. Preoperative biometry was obtained using IOLMaster 700 optical biometers (Zeiss, Oberkochen, BW, Germany). Demographics (patient age, gender and ethnicity), cataract surgery data, and postoperative refractions were obtained from the Aravind Eye Hospital electronic medical record. Patients included received one of the following lens models: Acrysof SN60WF lens (Alcon, Fort Worth, TX, USA), Auroflex FH5600AS (Aurolab, Madurai, TN, India), Aurovue HP760APY or HP760AP (Aurolab), or Toric FH560T* (Aurolab).

Data from the Michigan population (named "UMich") were obtained from patients undergoing cataract surgery at University of Michigan's Kellogg Eye Center and described

previously [106]. Preoperative biometry was obtained using Lenstar LS 900 optical biometers (Haag-Streit USA, EyeSuite software V.i9.1.0.0). Demographics (patient age, gender and ethnicity), cataract surgery data, and postoperative refractions were obtained via the Sight Outcomes Research Collaborative (SOURCE) Ophthalmology Data Repository.

Manifest refractions at both institutions were performed at the end of the first postoperative month by trained technicians. The inclusion criteria for the cases at both institutions were as follows: (1) Cataract surgery was performed; (2) No refractive surgery was performed before the cataract surgery; (3) No additional surgery was performed at the time of cataract surgery; (4) Visual acuity was 20/40 or better and (5) Data were complete and was not out of bounds for any of the formulas with which performance was compared.

Methods for collection of this data were conducted within regulations for patient privacy and can be viewed in [106]. The data collected consisted of primarily pre-surgical information such as axial distance, K1 and K2 keratometry measurements, age at surgery, anterior chamber depth, axial length, lens thickness, white-to-white distance, and central corneal thickness. Surgical information included the model and power of the implanted IOL.

5.3.2 IOL Power Prediction

IOL power prediction was performed using a collection of geometrical optics-based, regression-based, and machine learning-based formulas. These formulas included Barrett Universal II, Haigis, Holladay 1, HofferQ, Nallasamy, PearlDGS, and SRK/T. For regional comparisons of IOL formula performance, we selected both the Acrysof SN60WF lens which was implanted at both Aravind and University of Michigan. The Nallasamy formula was developed using perioperative cataract surgery data from the University of Michigan and has been described previously [106]. In order to ensure no information leakage, no patients included in the model development process for the Nallasamy formula were included in the UMich dataset considered here. Tables and figures of results from the UMich analysis are also provided in the previous work.

5.3.3 A-Constant Optimization

Lens constants optimization for all IOL formulas in this work are based on the optimized constants from 4390 patients implanted with the SN60WF lens in our previous work [106]. The The formulas for Haigis, Hoffer Q, Holladay 1, and SRK/T were implemented in Python based on their published equations and updates . The calculations confirmed with those obtained from Haag-Streit USA EyeSuite software V.i9.1.0.0. Prediction results for Barrett Universal II and PearlDGS were obtained through their online calculators. Briefly, the

optimal lens constant for each formula was determined through an empirical optimization process to zero out the mean prediction error. The optimized lens constants were: Barrett: 1.94, PearlDGS: 119.1, Haigis: -0.739, HofferQ: 5.727, Holladay: 1.860, SRK/T: 119.082.

	FH5600AS	SA60AT	SN60WF	TORIC	UMich (SN60WF)
Count	1130	672	985	136	1103
Age at Surgery	60.37	59.46	60.52	64.90	70.73
Laterality (R, %)	0.53	0.54	0.56	0.58	0.51
AL (mm)	22.99	23.03	23.19	23.01	24.15
CCT (μm)	522.07	520.21	526.13	517.88	552.58
ACD (mm)	3.30	3.28	3.28	3.25	3.25
AD (mm)	2.77	2.76	2.76	2.74	2.69
LT (mm)	4.19	4.23	4.24	4.30	4.53
K1 (D)	44.35	44.36	44.20	43.93	43.44
K2 (D)	45.14	45.03	44.86	45.77	44.32
Km (D)	44.74	44.69	44.53	44.85	43.88
Astigmatism	0.79	0.66	0.66	1.84	0.88
WTW (mm)	11.77	11.49	11.89	11.71	12.09
IOL Power (D)	20.51	21.37	21.02	21.52	19.89
Refraction (postop)	-0.15	0.10	0.010	0.12	-0.59

Table 5.2: Means by Lens Type

5.3.4 Statistical Analysis

Statistical tests and figures were conducted in Python 3.12 using scipy and matplotlib, respectively. Demographics for the Aravind and UMICH datasets were measured in pandas and assessed for p-value using the scipy stats package. All lens types were analyzed via their distributions and ANOVA with Tukeys analysis with Bonferroni correction and an alpha p-value of 0.01. For the predictive models, MAE and ME were compared across IOL formulas and assessed for statistical significance with the Friedman and Wilcoxon signed-rank tests with Bonferroni correction. Errors for each formula were also assessed based on different axial lengths. Because the Toric lens represents a group of patients not well-represented in the UMich dataset, Toric lens patients data was omitted in the Aravind-to-UMich comparisons but preserved in by-lens type comparisons.

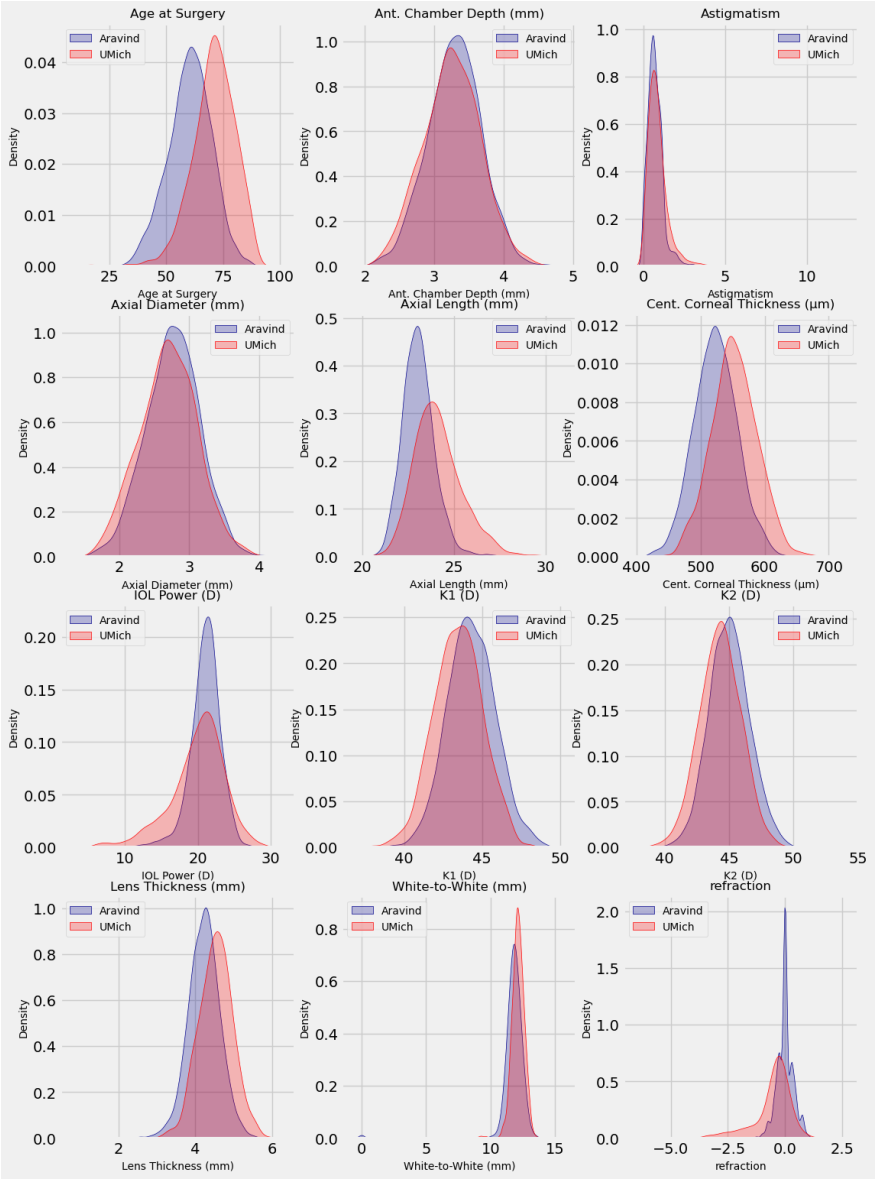

Figure 5.1: Distribution of Patient Measurements and Demographics. The Toric lens was removed so that both populations did not contain patients with astigmatism.

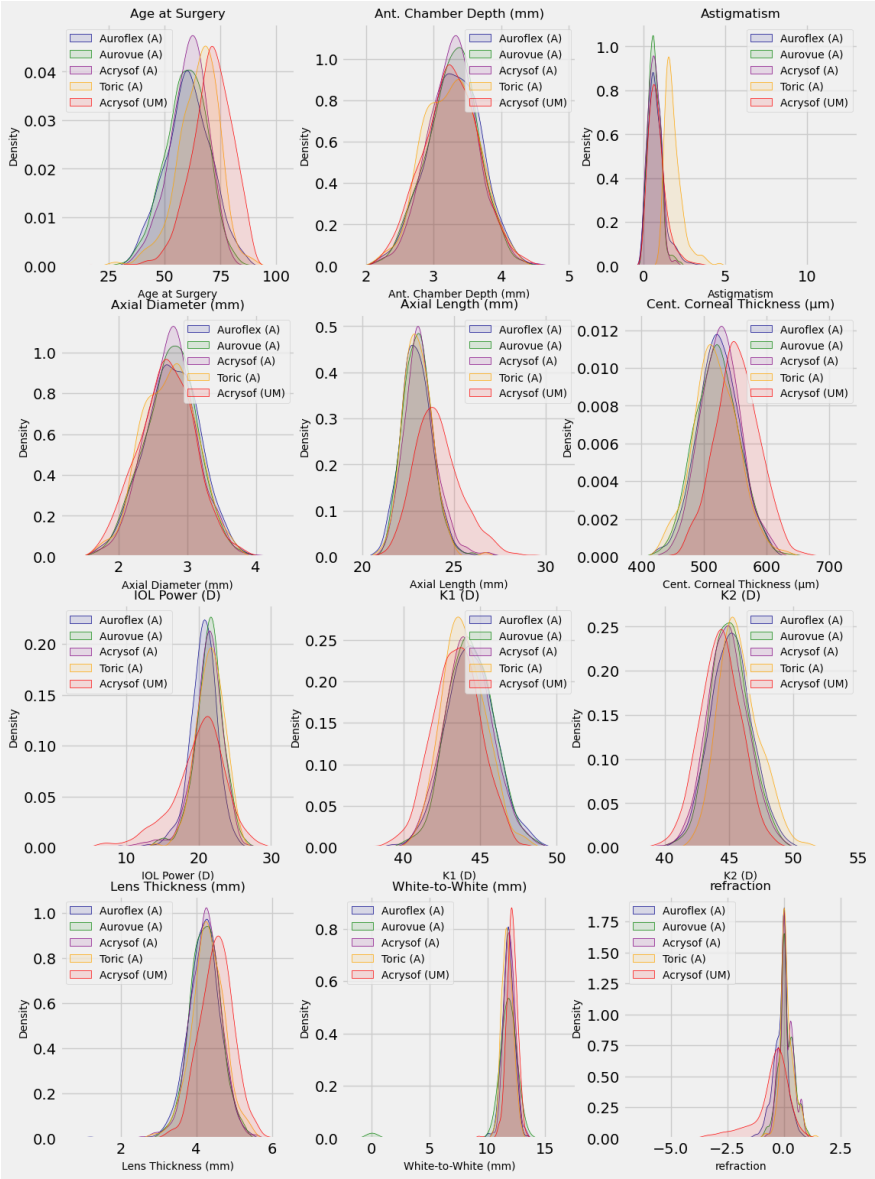

Figure 5.2: Distribution of Patient Measurements and Demographics by Lens Type

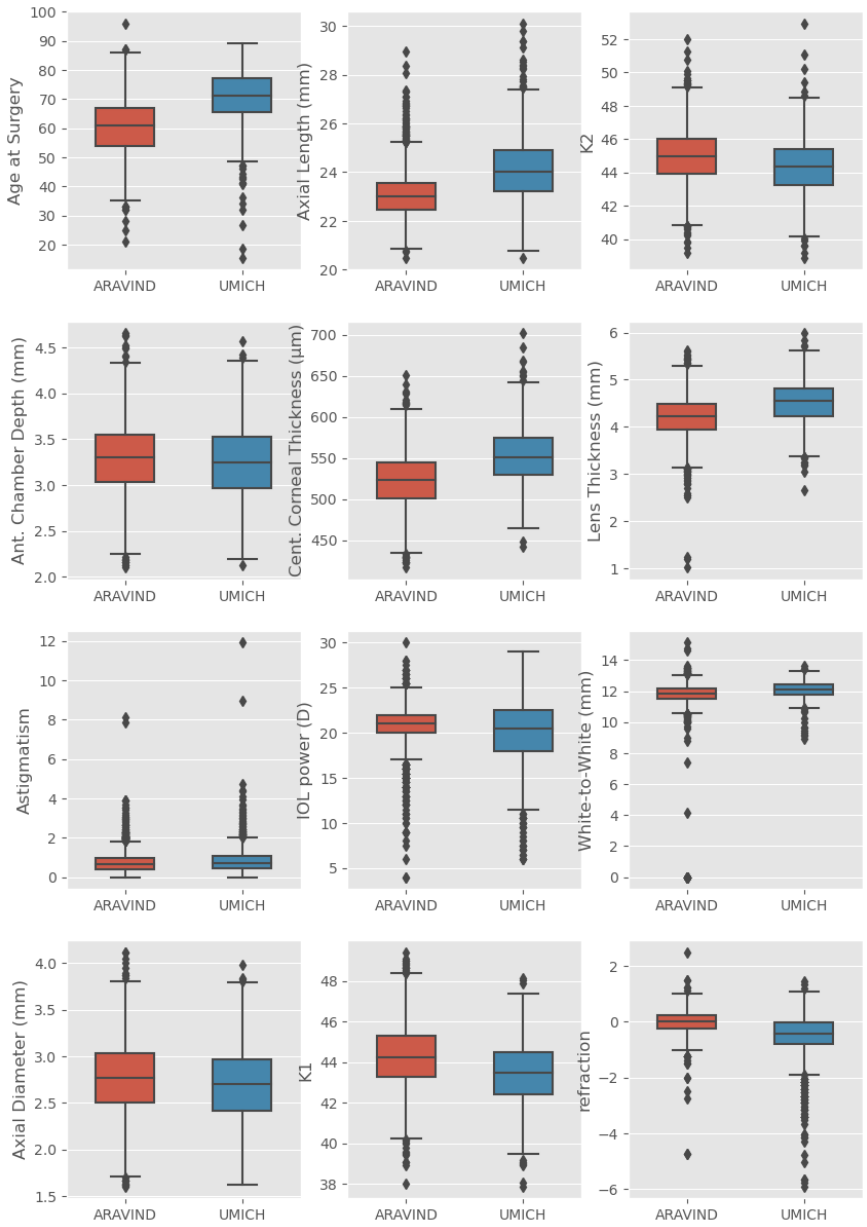

Figure 5.3: Boxplots of Patient Measurements and Demographics. The Toric lens was removed so that both populations did not contain paitents with astigmatism.

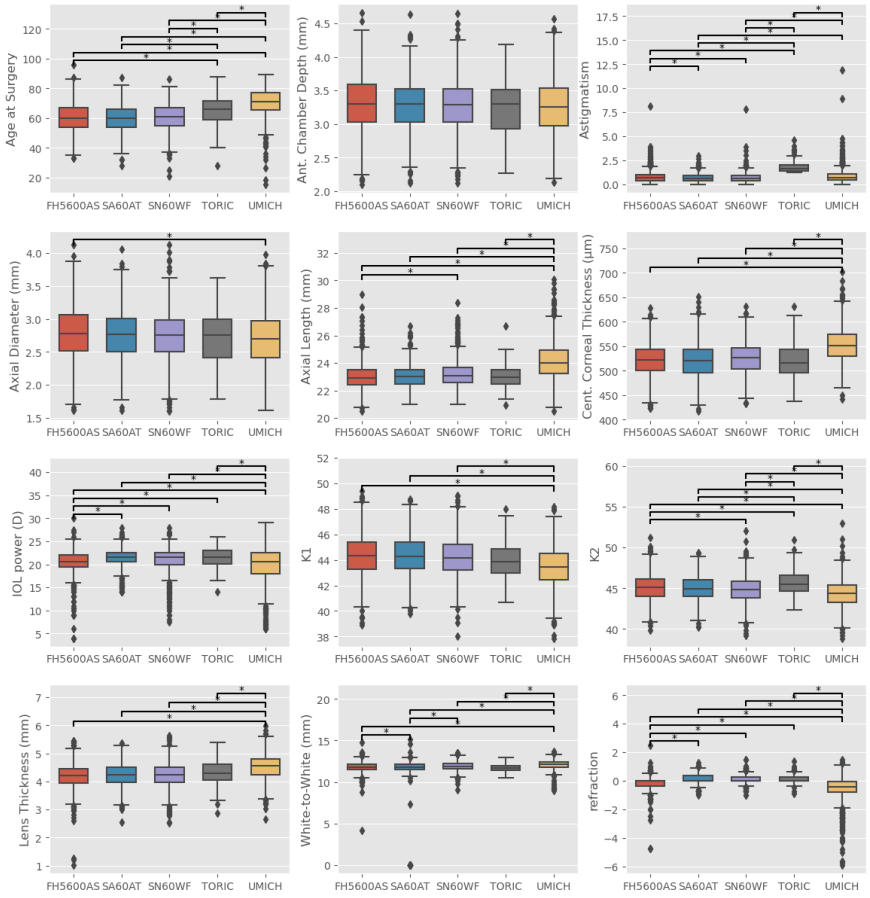

Figure 5.4: Boxplots of Patient Measurements and Demographics by Lens Type

5.4 Results

The Aravind dataset included 2929 patients who were administered either the Acrysof SN60WF lens (Alcon, M=568, F=417), or an Aurolab lens: Auroflex FH5600AS (M=646, F=484), Aurovue HP760APY or HP760AP (M=364, F=308), or Toric FH560T* (M=73, F=69) lens (Aurolab, Tamil Nadu, India). The UMich dataset included 1003 eyes of 1003 patients (M=433, F=570) who received the Alcon SN60WF (Acrysof IQ, Alcon, Fort Worth, TX) lens.

The dataset characteristics and biometry measures are summarized for all Aravind and UMich patients in Table 5.1. Data distributions across populations are depicted in Figure 5.1 and by lens type in Figure 5.2. Boxplots by population are depicted in Figure 5.3 and by lens type in Figure 5.4, with asterisks to represent Bonferroni corrected p-values < 0.01. Means by lens type are presented in Figure Table 5.2. The UMich population demonstrated significantly lower average IOL power (19.89 D , $p < 0.01$), post-operative refraction (-0.59, $p < 0.01$), and K1 (43.44 D, $p < 0.01$) and K2 (44.32 D, $p < 0.01$) measurements compared to the same measures in the Aravind population group, which were 20.90 D, -0.005, 44.30 D, and 45.01 D, respectively. The Aravind population demonstrated significantly lower mean age at surgery (60.20 y), lens thickness (4.22 mm), axial length (23.07 mm) and central corneal thickness (523.05 um) compared to the UMich group, whose measures were 70.73 y, 4.53 mm, 24.15 mm and 552.58 μm, respectively.

Distribution curves from the Aravind population showed lower variation standard deviations in IOL power (2.30), refraction (0.39) and axial lengths (0.91) compared to the UMich population, whose ranges standard deviations for IOL power, refraction and AL were 3.78, 0.93, and 1.35, respectively. When separated by lens type, it is notable that the Aurolab Toric lens appears to demonstrate significantly later age at surgery (64.9 y, $p < 0.01$) compared to the rest of their South Indian cohort (SN60WF: 60.52 y, HP760AP*: 59.46 y, FH5600AS: 60.37 y), although still earlier than the age at surgery of the UMich cohort (70.73 y). Aravind patients who were administered the Toric lens also demonstrated significantly higher measures ($p<0.01$) for astigmatism than any other group (1.84 D compared with 0.79 D for FH5600AS, 0.66 D for HP760AP* and SN60WF Aravind, and 0.88 D for SN60WF UMich).

The results from IOL power prediction can be seen in Tables 5.3 and by axial lengths for short and medium lengths in 5.4. Prediction error of each IOL formula by Axial Length can be viewed in a chart in Figure 5.6, while a breakdown of performance based on diopter of error is available in Figure 5.5. HofferQ and Holladay 1 tests appear to overestimate IOL power after 24 mm, while Barrett Universal II tends to underestimate IOL power with Axial

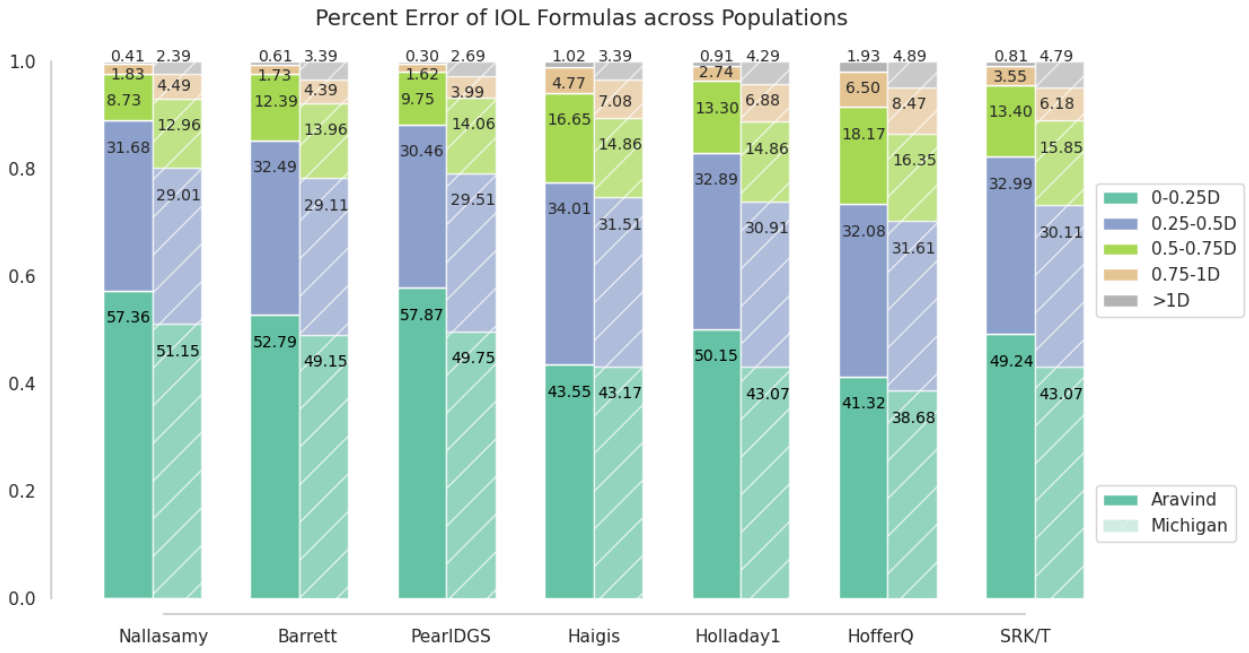

Figure 5.5: IOL formula performance on Aravind SN60WF and UMich SN60WF data

lengths < 25 mm and > 26.5 mm. PearlDGS appears to demonstrate the best performance at 24 mm and tends to increasingly underestimate at ALs > 26 mm. The Nallasamy formula appears to demonstrate the consistent performance with short and medium axial lengths, similar to PearlDGS, but appears to be less accurate with axial lengths > 26 mm compared to its performance on other axial lengths. However, it demonstrates the lowest MAE, lowest Median Absolute Error (MedAE), and the highest percentage of patients within 0.5 D of prediction for both the Aravind and UMich populations. Conversely, Haigis and Hoffer Q appear to demonstrate the highest percentage of errors over 0.25 D.

All formulas demonstrated better accuracy in the Aravind sample compared with the UMich sample. Among these, the Nallasamy formula appears to be the best-adapted, with a marked improvement in errors within 0-0.25 D of 6.4% compared to the UMich population and a 2.56% improvement in errors within 0.25-0.5 D. The Holladay 1 formula demonstrates the second best improvement, with a 7.08% improvement in 0-0.25 D errors and 2.50% improvement in 0.25-0.5D errors. Haigis and HofferQ formulas reflect the smallest improvement, with only 0.38% gains and 2.64% gains in the 0-0.25 D range of error, respectively.

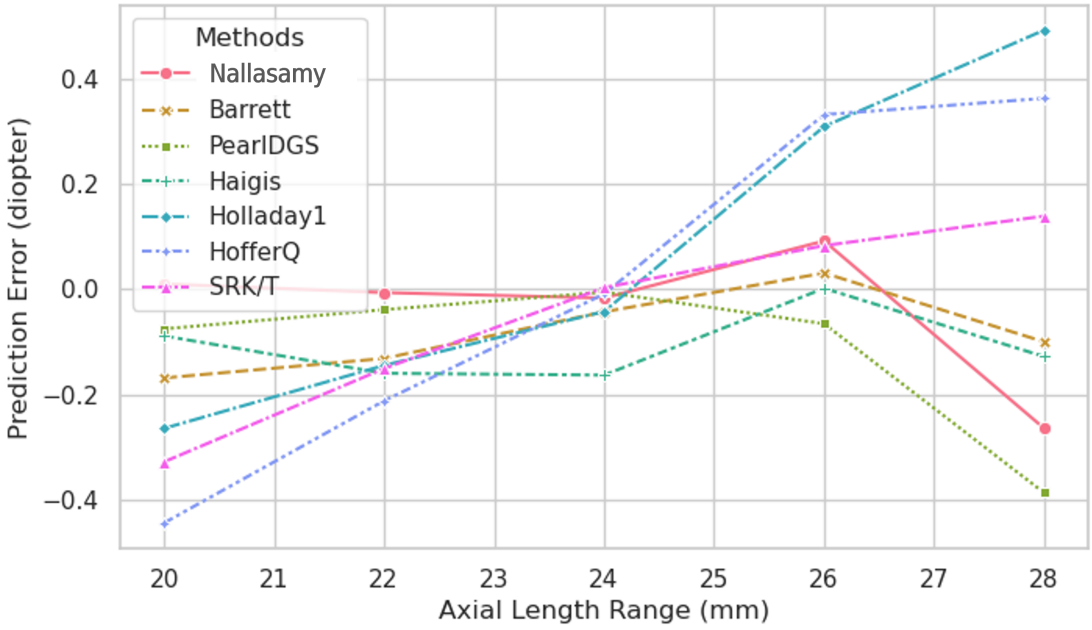

Figure 5.6: IOL formula performance on Aravind SN60WF data by Axial Length

5.5 Discussion

In this study, eye measurements from two populations undergoing cataract surgery were compared: 1) A Midwest population in the United States ("UMich"), and 2) A South Indian population ("Aravind"). This study has refrained from classifying populations as "Caucasian/American" and "Indian" because it cannot be ruled out that other factors or differences may be at play that do not generalize across ethnicity or national identity. In fact, sun exposure and clinical practice may explain a large proportion of differences in these populations.

Method	**MAE**	**ME**	**SD**	**MedAE**	**m**	**AE**< 0.5	**FPI**	**p-value**
Barrett	0.2798	0.1195	0.3309	0.2400	-0.3675	0.8528	0.1846	$<$0.01
Haigis	0.3296	0.1526	0.3815	0.2849	0.0062	0.7756	0.4957	$<$0.01
HofferQ	0.3583	0.1936	0.4018	0.3128	-0.8098	0.7340	0.0983	$<$0.01
Holladay1	0.2961	0.1331	0.3459	0.2478	-0.6638	0.8305	0.1185	$<$0.01
PearlDGS	0.2516	0.0375	0.3194	0.2110	-0.2411	0.8832	0.2455	$<$0.01
SRK/T	0.2963	0.1392	0.3434	0.2537	-0.7629	0.8223	0.1059	$<$0.01
Nallasamy	0.2497	0.0057	0.3202	0.2057	-0.0610	0.8904	0.4427	/

Table 5.3: Performance table of various formulas on SN60WF at Aravind dataset

The UMich dataset demonstrated a wider range of postoperative refraction, but also lower lens diopters of implanted lenses, implying that operations included patients who asked for reading vision instead of emmetropia. The lower range of implanted diopters and an average refraction around 0 for Aravind patients is likely a reflection of Aravind Eye Center practice, which is to correct all patients to emmetropia. One reason for this could be the need for speedy and consistent surgeries due to the high patient volume seen at the eye center. By contrast, University of Michigan practice is to ask patients whether they prefer to optimize long-distance vision or reading vision.

When broken down by lens type, Aravind patients who were administered the Toric lens also had significantly higher levels of astigmatism, marked by lower K1 and significantly higher K2 values compared to patients with other Aurolab lenses. This matches expectations that patients diagnosed with astigmatism are also prescribed the Toric lens as treatment. Interestingly, UMich patients overall demonstrated significantly lower K1 and K2 values than any other group. Since spherical error has been shown to link to lower Keratometry values [7], we hypothesize that preoperative refractive errors in the UMich patient group were higher and more frequent than the Aravind group. Although this could not be confirmed with the measurements available in our dataset, the prevalence of 55-60 year olds in a comparable Midwest state demonstrated a 50.1% prevalence of myopia [185] and the same age group at Aravind Eye Center in Madurai, India showed a 43-44% prevalence of myopia [90], supporting our hypothesis.

Interestingly, all South Indian patients recorded, regardless of the lens, demonstrated consistently lower axial lengths compared to the Michigan cohort. This may be explained in part by higher levels of ultraviolet (UV) exposure, which has been shown to be negatively correlated with Axial Lengths [156]. As South India is close to the equator, the yearly level of UV exposure for local residents is shown to be higher than for local residents in the midwestern United States, where winters are often overcast and sunlight hours shorter [194, 1]. The seriousness of this issue is also reflected in lower levels of vitamin D in the Michigan population [73]. Conversely, for South Indians, higher UV sun exposure could also explain earlier onset of cataracts and thus a younger patient age at time of surgery. This would be important to know for public health efforts targeting eye health.

Interestingly, central corneal thickness and lens thickness were also significantly lower in the South Indian population compared to their Michigan cohort. These cannot be attributed to UV exposure, as [141] suggests UV exposure should actually increase CCT. One study of a Chinese cohort found an association with age and lens thickness [120]. Since the UMich cohort also presented older ages at surgery, this could be a possible explanatory factor. In another study, [6] found ethnic differences between Caucasian and Japanese populations

	Short (n=78)				Medium (n=895)			
Formula	**MAE**	**MedAE**	**ME**	**STD**	**MAE**	**MedAE**	**ME**	**STD**
Barrett	0.33	0.30	0.17	0.38	0.28	0.24	0.12	0.33
PearlDGS	0.29	0.23	0.08	0.35	0.25	0.21	0.03	0.32
Haigis	0.34	0.30	0.09	0.40	0.33	0.29	0.16	0.38
Holladay1	0.37	0.34	0.26	0.36	0.29	0.24	0.13	0.34
HofferQ	0.49	0.45	0.44	0.38	0.35	0.30	0.18	0.39
Nallasamy	0.27	0.24	-0.01	0.33	0.25	0.20	0.01	0.32
SRK/T	0.39	0.34	0.33	0.35	0.29	0.25	0.13	0.34

Table 5.4: Performance of IOL formulas on SN60WF Aravind population

based on CCT. The South Indian cohort in this study demonstrated even lower CCTs than the Japanese group, and measurements of the Michigan group and the South Indian group in our study are consistent with findings in other studies [203]. Therefore, besides age, another explanation for these measurement differences could be ethnic differences.

In the IOL formula assessment, all formulas performed better with the South Indian Aravind population compared with the midwestern US UMich population. The wider range of post-operative refractions in the UMich dataset combined with longer axial lengths is a likely cause for the poorer accuracies in the UMich dataset compared with the Aravind dataset, which exhibited low ranges and refraction centered at 0, and lower non-skewed axial lengths.

In [106], IOL formulas were assessed on our mostly Caucasian UMich dataset, and Barrett was shown to outperform Haigis, Hoffer Q, Holladay 1 and SRK/T, both within 0.25 D and 0.5 D. However, Barrett's performance gains with the related Aravind data were considerably lower compared with SRK/T and Holladay1 showing markedly better performance with the Aravind dataset. As Barrett performs better than every other IOL formula listed except for Our Method (Nallasamy formula) in both samples, it is possible that the results are a reflection of Barrett's steady performance across a wide range of axial lengths, evidenced in Figure 5.6. This would corroborate claims in [138], which stated that Barrett performs more stable than Holladay 1 and Hoffer Q formulas when given longer ALs.

One limitation of our work with IOL formulas was a lack of large representation from longer axial lengths. Larger sample sizes of long axial lengths could affect the overall prediction error of the IOL formulas due to lower variation. Due to the link between sun exposure and axial lengths [156], it is reasonable to assume that longer axial lengths are rarer in South India. However, in the future this may change. Prior studies have found that younger college-educated South Indians had a nearly twofold increased odds of myopia compared to no education [90], while another asserted the association between near-work/outdoor time

ratios and myopia [61]. Indoor time spent on near-work such as homework or screen time may lead to future generations of Indian students with longer ALs than our sample group. Due to this, future studies may want to analyze relationships between AL and myopia in cohorts with Indian students, as well as how this may affect cataract populations in the future.

The benefits of understanding populations and trends in eye health are clear. For example, in the United States, younger generations experience lower exposure to UV in the US now compared to years prior because of public health efforts to prevent cataract. The South Indian population provides a unique patient pool whose needs must be accounted for carefully based on their own environmental factors. However, many studies relating to cataract surgery tend to focus on western populations. In this study, we have shown a comparative analysis of a South Indian group compared with a Midwestern United States group to demonstrate how location can affect the distribution of patient populations and thus why it is important to research a variety of different patient populations in cataract research. Additionally, we have shown that accuracy and precision of some IOL formulas may also differ across populations.

5.6 Publication and Acknowledgements

This chapter is a submitted work: Elisa Warner, Miles Greenwald, Tingyang Li, Prashanth Gupta, Jyothi Vempati, Karthik Srinivasan, Haripriya Aravind, Nambi Nallasamy. Comparison of Cataract Surgery Patients in South Indian and Midwestern United States Populations.

CHAPTER 6

IOL Prediction Models, Part II: Prediction of Postoperative Intraocular Lens Position in Cataract Surgery using Domain Generalization

6.1 Abstract

Cataract is a serious condition characterized by a protein buildup in the intraocular lens (IOL) leading to opacity and eventually loss of vision. Surgical replacement of the IOL necessitates precise prediction via IOL formulas that understand relationships between IOL power, patient biometry, and post-operative refractions. The Nallasamy Formula is an ML-based IOL formula previously introduced to predict post-operative refraction for patients implanted with Alcon SN60WF lens. This chapter presents the development of a domain-generalized approach to expand the Nallasamy Formula's prediction of post-operative refraction to additional lenses with different user populations. Since limited data was available to train our generalized model (which we dub "Nallasamy-G") on a multitude of various lenses, a domain-generalization approach via understanding of lens properties was designed to assist the model in the case of unreliable A-constants based on the supposition that target users may not have access to optimized A-constants. Using this approach, we leveraged patient biometric measurements such as axial length, keratometry, and central corneal thickness across four different datasets to demonstrate promising results in post-operative refraction prediction compared to classical IOL formulas such as SRK/T, Holladay 1, Hoffer Q, and Haigis. Our Nallasamy-G model provides a framework for understanding why the Nallasamy Formula is limited to the SN60WF lens and how to approach domain-generalization of the model to other lenses for future studies that intend to extend the model to generalized IOL power prediction of multiple lens models.

6.2 Introduction

Cataracts are a serious condition characterized by a protein buildup in the IOL leading to opacity and eventually loss of vision. It is the leading cause of blindness in older adults and requires surgical replacement of the IOL to restore vision. Surgical replacements require the use of IOL formulas, which understand the relationship between eye measurements, IOL power and post-operative refraction to predict the right power of lens for a given desired refraction. Failure to implant the correct lens can lead to patient dissatisfaction, the use of corrective lenses post-surgery, or an additional surgery to correct errors [129].

IOL replacements vary widely based on the manufacturer of the lens and the needs of the patient. Common IOL replacements are constructed from copolymers such as Polymethyl methacrylate (PMMA) and each material can hold unique refractive indices. Other properties can include haptics and convexity of the lens at different diopters. For example, most replacement IOLs tend to be biconvex, but some may tend to allow more curvature on the anterior or posterior sides of the lens as the diopter increases, whereas equiconvex lenses would require equal levels of curvature on each side as the diopter increases [3].

IOL prediction formulas based on models of the eye and empirical data attempt to assess the best fit diopter of lens for each patient based on a handful of critical measurements such as axial length (AL), the curvature of the cornea (keratometry, K), and central corneal thickness (CCT). However, most formulas depend strongly on a user-defined input that is known as the "A-constant," a catch-all for various unaccounted determinants of IOL diopter, including manufacturing differences and population-based differences. This somewhat simplistic error correction is a critical part of every IOL formula and each formula has a unique constant that must be tailored to their specific formula. The dependency on such constants harbors a liability of oversimplifying a complex system within the eye and this oversimplification becomes more evident as the patient's axial length diverges from the norm. This was demonstrated in a previous paper [106], which found HofferQ and Holladay 1 to be

Sym	Code	Lens	Mfr	Location	City	Size
A	FH5600AS	FH5600AS	Aurolab	Aravind Eye Hospital	Chennai	1130
B	HP760AP∗	HP760AP/ HP760APY	Aurolab	Aravind Eye Hospital	Chennai	673
C	SN60WF	SN60WF	Alcon	Aravind Eye Hospital	Chennai	985
D	UMich	SN60WF	Alcon	Kellogg Eye Center	Ann Arbor	5016

Table 6.1: Description of all datasets with the code used in the paper. Our generalized model (Nallasamy-G) was trained on the UMich dataset and evaluated on the UMich, SN60WF, FH5600AS and HP760AP* datasets. Note that

Input	Mean	Median	Min	Max	STD
IOL power	19.79	20.5	6	30	3.71
Sex	0.43	0	0	1	0.49
Age at Surgery	71.01	71.58	12	89.45	9.4
AL	24.16	23.98	20.44	31.57	1.34
CCT	551.41	551	418	702	35.73
AD	2.7	2.71	1.52	4.21	0.41
ACD	3.25	3.26	2.03	4.79	0.41
LT	4.53	4.52	2.66	6.08	0.44
K1	43.43	43.41	37.41	50.01	1.54
K2	44.33	44.27	38.69	52.94	1.63
AST	0.9	0.73	0	12.85	0.72
WTW	12.13	12.14	7.07	14.68	0.52
Laterality	0.51	1	0	1	0.5
Rrefraction (postop)	-0.55	-0.41	-6.16	1.84	0.86

Table 6.2: Demographic Table of the UMich dataset (Patients implanted with the SN60WF lens at University of Michigan)

some of the worst offenders of this. This is also visible in formula adjustments. In SRK II and SRK/T formulas, for example, corrective constants have to be added to the formula as Axial Lengths deviate from the norm, because the original SRK model is too simplistic to extrapolate far beyond average measurements. Other pitfalls of using A-constants as a catch-all is the assumption that the A-constant can represent both configuration of the lens as well as differences in patient populations, and that this relationship would somehow be linear.

Our previously published AI-based Nallasamy formula [106] was demonstrated to be a competitive alternative to other formulas such as Barrett Universal II, SRK/T, Holladay 1, Hoffer Q and Haigis formulas for the Alcon SN60WF lens. Because the model was designed specifically for this one lens, no A-constants were needed for the formula and the model weights were trained specifically for SN60WF. However, model weights and trained parameters may differ between lens model domains, meaning the Nallasamy formula trained for SN60WF may not extrapolate well to different lens types. Although training on data from a multitude of different manufacturer lenses is desirable, not enough data from additional lens types was available to train the model on diverse inputs. Therefore, in this study we sought to increase the generalizability of the Nallasamy formula trained on patient data from a single lens implant so that it can produce accurate predictions of other lens domains.

Domain generalization has been conducted in other studies with regards to the eye, typically in fundus imaging. In [217], mentioned in Chapter 2 augmentation of fundus images

was used to generalize the model's representation space to allow a greater classification net that encompassed input images from different cameras. In another study [57], domain generalization was conducted through specialized visual transformers that randomly add more variation and also classify using "soft" predictions. Lastly, [116] uses a deep-learning model for automated augmentations, where the model learns augmentation strategies that lead to different domains. The construction of a "domain decoder" in the model finally disentangles domain from segmentation of the fundus image. These approaches, however, are only applied to fundus images and cannot extrapolate to our IOL formula prediction problem. Additionally, complex deep-learning based methodologies are inappropriate here, where our sample sizes are relatively small for deep learning.

In this chapter, we attempt to robustify the Nallasamy formula through informed model architectures and feature engineering. We begin with an assessment of the Nallasamy formula performance. Then, we attempt a robustification of the Nallasamy formula's features with a focus on keeping predictions the same or better than the original model. This is done through training the model on patients implanted with the SN60WF lens only. Second, we evaluate our generalized model (which we dub "Nallasamy-G") using data from the Aravind Eye Hospital in India where patients were implanted with one of three lens models.

6.3 Materials and Methods

6.3.1 Data Collection

All data in the study were obtained in accordance with the statutes of the Declaration of Helsinki. The University of Michigan ("UMich") data were collected from Kellogg Eye Center in a retrospective cohort study approved by the Institutional Review Board of the the University of Michigan (HUM00160950). Data collected from 6138 patients (9452 eyes) who were implanted with the Alcon SN60WF (Acysoft IQ, Alcon, Fort Worth, Texas) lens were obtained from the SOURCE database in accordance with the methods described in [106]. After calculating valid IOL formula predictions for each patient, a full dataset of 6031 patients (9244 eyes) was collected.

A full view of patient inclusion, exclusion, and partitioning into training and testing during the model-building phase can be viewed in Figure 6.1. Patients with prior RK or refractive surgery were removed, as well as those exhibiting invalid measurement values. Patients exhibiting missing data or vision worse than 20/40 were also excluded from the study.

Testing on third-party data was performed using data obtained from the Aravind Eye

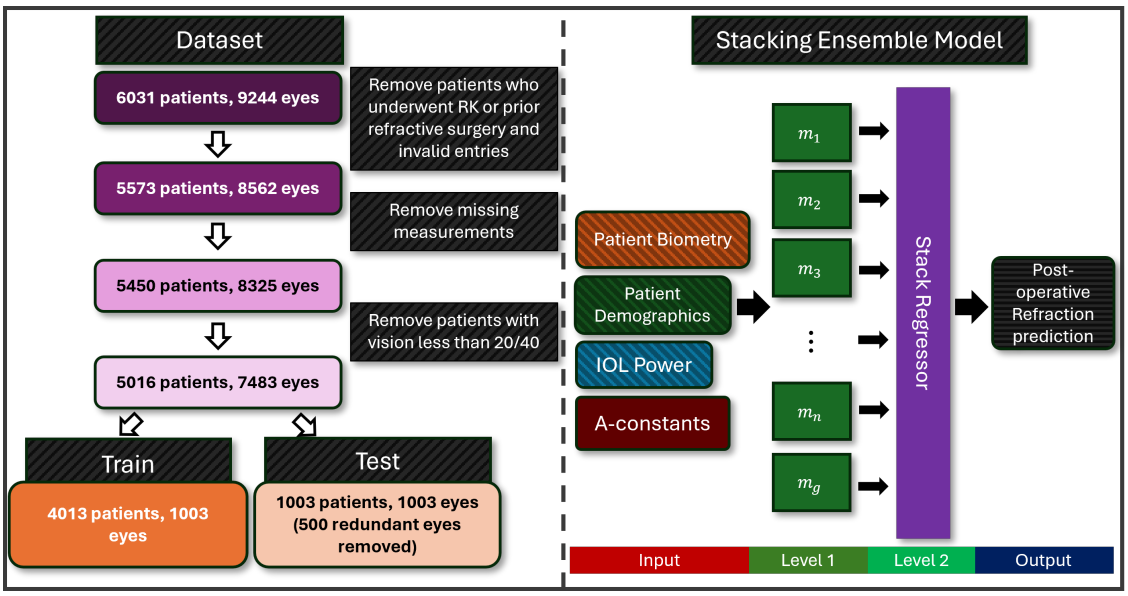

Figure 6.1: (*left*) A diagram describing inclusion/exclusion from the training set and dataset allocation into train and test for our generalized model (Nallasamy-G). (*right*) Model construction of Nallasamy-G. The model, based on the Nallasamy Formula model, consists of a 2-level ensemble network structure with multiple models in the level 1 ensemble which provide outputs to a level 2 Linear Regression Stack Regressor model. In our Nall-G model, we add a generalizing module in level 1 (m_g). The output of the level 2 model is the single post-operative refraction prediction used for analysis.

Hospital in Tamil Nadu, India. Data were collected in accordance with Indian Heath Service Institutional Review Board criteria (RET202100362). The data consisted of patients who were administered either the FH5600AS lens, HP760AP/HP760APY (dubbed "HP760AP*") lens, and SN60WF lens (Aurolab, Tamil Nadu, India). The data collected for both groups consisted of primarily pre-surgical information such as axial distance (AD), K1 and K2 keratometry measurements, age at surgery, anterior chamber depth (ACD), axial length (AL), lens thickness (LT), white-to-white distance (WTW), and central corneal thickness (CCT). Surgical information included the IOL power of the implanted lens and the postoperative refraction of each patient. Patients with BCVA worse than 6/12 (or 20/40 vision) were excluded from the study. Patients with prior refractive surgery were also excluded and missing data removed.

6.3.2 Evaluation of the Nallasamy Formula

The Nallasamy Formula metrics are included in this paper. Since the Nallasamy Formula is constructed with features that depend on A-constants, these features were generated for each dataset based on an A-constant set which matched the lens of the dataset. Since the

Nallasamy Formula is based on an ensemble framework [105] (illustrated in Figure 6.1, a feature importance analysis was conducted on two of the level 1 models.

6.3.3 Obtaining A-Constants

This study assumes that an IOL formula user would not have access to their own empirically-optimized A-constants. Therefore, two primary scenarios were considered: 1) The presence of no A-constants except the manufacturer A-constant. The manufacturer A-constants were obtained from manufacturer websites and product guides (Aurolab, Alcon). Because multiple A-constants are needed for our model, an A-constant conversion was performed using the ULIB A-constant converter [2]. 2) The presence of empirical A-constants given by another party. In this paradigm, we assume the user has access to only publicly-available studies where third party datasets were tested for optimal A-constants. In this case, A-constants for each of the lenses in FH5600AS, SN60WF were extracted from the ULIB optimized A-constants table [2]. Note that the Alcon SN60WF lens contains multiple entries, including a general set of A-constants, one from Japan, and one from India. Because the Aravind data comes from India, we believed the A-constants listed for SN60WF (India) were a better representation of a use case of our data. For the UMich dataset, we utilized the general SN60WF A-constants, which were not labeled with any country. A-constants were obtained in this study for SRK/T, Haigis, Holladay and Hoffer Q IOL formulas.

For a final comparison, optimized A-constants were also obtained in a post-hoc analysis. These A-constants are obtained via a grid search optimization whereby the selected constant minimizes the absolute mean error (ME) of the test dataset. Note that this last paradigm was conducted as a post-hoc test, meaning that all IOL formula results shown for SRK/T, Haigis, Holladay1 and HofferQ, are at their best predictive values fit to the test datasets after all other analyses were finished.

6.3.4 Data Interpolation

To improve model performance, data was interpolated at each fold to increase training set size by 10000 samples. Each interpolated data point was created by averaging the measurements and refraction for four randomly chosen patients. Interpolated data were then concatenated with the fold's training set before model training.

Type	Dataset	Haigis	HofferQ	Holladay1	SRK/T
ULIB	FH5600AS	0.68	4.92	1.120	117.8
	SN60WF	1.35	5.53	1.76	118.9
	UMich	-0.769	5.64	1.84	119.9
Manufacturer	FH5600AS	1.52	4.85	1.11	117.8
	HP760AP*	1.714	5.37	1.62	118.7
	SN60WF	1.904	5.56	1.80	119.02
	UMich	1.904	5.56	1.80	119.02

Table 6.3: A-constants used for this analysis. ULIB did not contain A-constants for HP760AP*. Haigis a1 and a2 constants were assigned as the default (a1=0.4,a2=0.1) for every dataset except for the SN60WF University of Michigan (UMich) dataset, which was assigned as a1=0.234, a2=0.217 according to the ULIB A-constant table. It was assigned with the default constants, however, in the Manufacturer set, as assigned by the A-constant converter.

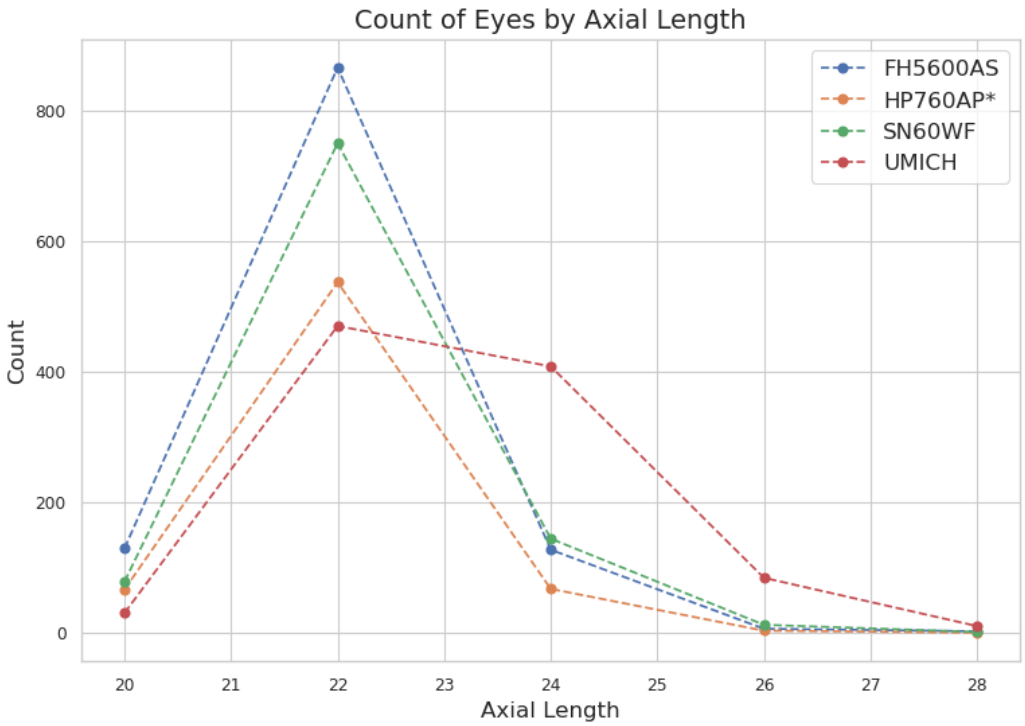

Figure 6.2: The count of available eyes by Axial Length for each dataset.

6.3.5 Model Development

Due to our limited dataset size and number of observable features, we selected a classical machine learning approach similar to the Nallasamy Formula [106]. Since a single model is prone to heavier bias, we selected a late ensemble approach. The model consists of two

levels of computation : 1) an initial predictor level consisting of individual predictions from multiple ML models (Level 1), and 2) a Stacking Regressor (Level 2), designed to take as input the output of each of the level 1 models and compute a final prediction.

Because many of the initial inputs are measurements that can benefit from additional mathematical operations that connect and represent interactions between values, we believed ANNs to carry some of the best potential for accurate prediction. Therefore, we added a "generalizing model" ANN consisting of 3 hidden layers of 700 nodes each. Models were optimized via a grid search approach in the training set.

6.3.6 Feature Engineering

As the feature analysis of the previous Nallasamy Formula indicated the Nallasamy Formula model to be highly dependent on the correct choice of A-constants, a focus of our new generalized model was for less reliance on A-constants and as a result focus on various eye calculations that do not include A-constants. Consequently, 33 features were added to the model and 9 features were removed which were believed to not contribute additional information towards the predicted output.
New features include the following A-constant based measurements:

1. `A1A2` : $a1 \times ACD + a2 \times AL$ (as calculated by Haigis)
2. `SRKII` : SRK II predicted IOL power given a refraction,

New features also include the following non-A-constant based features:

1. `defaultBarrettModIOL` - the predicted IOL by our modified Barrett for 0 refraction (see 6.3.6.1)
2. `barrettModRefraction` - the predicted refraction by Barrett Mod for given IOL (see 6.3.6.1)
3. `defaultBarrettOrigIOL` - Original IOL predicted by Barrett for 0 refraction [16]
4. `ThinLens` : Thin Lens predicted IOL power given a refraction,

   and the following non A-constant-based theoretical eye equations:
5. `defaultThickness` - lens thickness calculated by Barrett [17]
6. `Preop_RCP` - preoperative peripheral radius of the cornea calculated by Barrett [17]
7. `Preop_RG` - preoperative radius of posterior segment of globe calculated by Barrett [17]

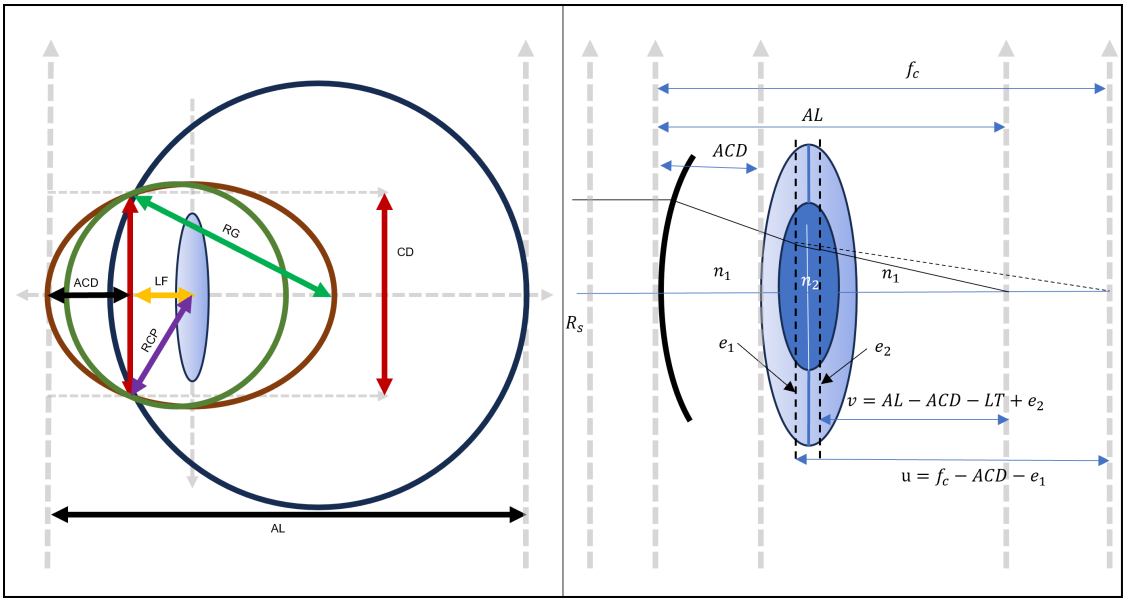

Figure 6.3: (*left*) Diagram of the eye model we propose for calculating the `pred_RCP` feature, which estimates the radius of the peripheral cornea. Our diagram is similar to that shown in [17], but is a simplification which estimates the RCP as slightly longer to make the length calculable. (*right*) Diagram of the universal eye model proposed in [16], used as the basis for our Modified Barrett I formula. The pre-surgical lens thickness was estimated as a proxy for the lens capsule size and used to make estimates for v, which measures the distance to the posterior focal point of the eye. Note that e_1 and e_2 represent first and second principal planes of the lens, respectively. R_s represents a corrected post-operative refraction. Variables n_1 and n_2 are described in Appendix D

8. `Preop_RC` - preoperative central anterior radius calculated by Barrett [17]

9. `Preop_P` - Preoperative Q factor (called P factor in Barrett's paper), as calculated by Barrett [17]

10. `Orig_Curv` : predicted curvature of lens by original Barrett formula [16]

11. `Orig_CornealPower` : predicted corneal power by original Barrett formula [16]

12. `F2_Orig` : predicted power of posterior surface of implant by original Barrett formula [16]

13. `F1_Orig` : predicted power of anterior surface of implant by original Barrett formula [16]

14. `E2_Orig` : projected distance from posterior side of lens to second principal plane by original Barrett formula [16]

15. `E1_Orig` : projected distance from anterior side of lens to first principal plane by original Barrett formula [16]
16. `U_Orig` : image distance by original Barrett formula [16]
17. `V_Orig` : object distance by original Barrett formula [16]
18. `defaultSRK` : SRK assessment of IOL given refraction of 0 [161]
19. `SRKRefraction` : SRK assessment of refraction given IOL [161]
20. `HofferACD`: Hoffer's relationship between postoperative ACD and AL
21. `BinkhorstACD` : Binkhorst's modified ACD
22. `OlsenpACD` : Olsen's postoperative ACD
23. `NaeserpLoc` : Naeser's posterior lens capsule location
24. `SRK_C2` : corneal width as calculated by SRKT [161]
25. `SRK_H` : Corneal height as calculated by SRKT [161]
26. `SRK_LCOR`: Corrected axial length as calculated by SRKT [161]
27. `SRK_pACD`: post-operative ACD as predicted by SRKT [161]
28. `Preop_Foc` : $AL - ACD - LT$
29. `Haigis_ACD` : anterior chamber depth as calculated by Haigis
30. `pred_RCP` : A guess at Barrett's predicted radius of corneal power based on his 2nd paper
31. `Barrett_pACD` : A guess at Barrett's predicted postop ACD based on his 2nd paper

In particular, the barrettModRefraction and its inverse defaultBarrettModIOL, along with pred_RCP, were novel constructs developed for this formula. Selected equations for features 1-31 can be viewed in Appendix D. The predicted radius of peripheral cornea or `pred_RCP` algorithm pseudocode can be viewed in Appendix E, and our modified Barrett algorithm can be viewed in Appendix G. The original Barrett from which our modified Barrett was based can be found in Appendix F.

As stated above, nine features from the Nallasamy Formula were omitted from this study. Among those were removed in our model were features that demonstrated no variation within training data or were deemed poor predictors of the outcome.

NAME	MAE	ME	MedAE	0.5D%
STACKING	**0.3098**	**0.0331**	**0.2424**	**0.8016**
Generalized Module	0.3142	0.0487	0.2438	0.8016
Holladay1	0.3706	0.0207	0.2980	0.7398
SRK/T	0.3760	0.0144	0.2999	0.7318
HofferQ	0.4038	0.0091	0.3311	0.7029
Haigis	0.3632	0.0237	0.2894	0.7468
Barrett	0.3280	0.0376	0.2564	0.7827
Kane	0.3148	-0.0196	0.2436	0.7986

Table 6.4: Nallasamy-G's performance on the test set (UMich) by ensemble model. Comparisons with optimized constants for Holladay1, SRK/T, HofferQ, Haigis, Barrett and Kane are given at the bottom

6.3.6.1 Our Modified Barrett I

One new non-A-constant-based feature we included in the model to boost generalizability was a modified formulation of Barrett's first IOL prediction formula [16]. In this first rendition, Barrett makes several assumptions, including a fixed lens thickness of 1 mm and a posterior radius of curvature of the implanted lens of 25 mm. In our modified version, we made three express changes:

1. **IOL thickness:** IOL thickness was calculated within the formula using the following equation, based on Barrett's introduction to his Universal II formula [17]

$$T = (RA - \sqrt{RA^2 - ((OD/2)^2)}) + (RP - \sqrt{RP^2 - ((OD/2)^2)}) \tag{6.1}$$

, where RA is anterior radius of curvature and RP is posterior radius of curvature.

In the original formula, the posterior radius is given a fixed constant of 25 mm. Instead of this, we consider the lens convexity in our calculation. Based on optics of biconvex lenses, we know that

$$OD = P_1 + P_2 \tag{6.2}$$

, where OD is the overall power of the biconvex lens in diopters, P_1 is the power of the anterior lens and P_2 is the power of the posterior lens. Lens power is related to curvature with the following equation:

$$P = (N_2 - N_1) * 1000/R \tag{6.3}$$

, where R is radius of curvature of the lens and P is the power in diopters. N_2 and N_1 are equivalent to the refraction index of the lens material and the aqueous humor, respectively.

In our modified formula, anterior and posterior curvature is based on a lens power which is dependent on the manufacturer build of the model, allowing currently either equiconvex or anterior asymmetric designs. In the equiconvex design, the assumption of the relationship between P_1 and P_2 is such that

$$P_1 = P_2 \tag{6.4}$$

. However, in anterior asymmetric designs where we assume greater curvature is assigned to the anterior lens side, the assumption is a relationship where a greater fraction of the overall OD of the IOL is placed on the anterior surface:

$$P_1 = 2P_2 \tag{6.5}$$

Note that this change expressly affects IOL thickness and that this thickness differs based on the curvature of each lens, even if the total IOL power remains the same. This can be viewed in Figure 6.11, where an fixed IOL power of 21.0 D demonstrates a quadratic relationship with the anterior lens power, reaching it's lowest point when the lens is equiconvex (anterior power equals posterior power).

2. **Placement of principal planes:** Because biconvex lenses with stronger anterior power would contain a thinner posterior lens than an equiconvex lens, the placement of posterior principal plane (see measurement e_2 in Fig 6.1) is assumed to be positioned more anterior than its location in a biconvex lens. This adjustment is small.

3. **Change to Olsen's number:** The refractive index for the aqueous humor of the eye was updated from Binkhorst's number of 1.336 to Olsen's number of 1.3315 [173].

4. **Improved Axial Length assumptions:** Experimentally-derived axial lengths for all patients were increased by 0.13 for the formula as an estimation of the distance between the back of the retina and the fovea in accordance with Barrett's modifications in the Universal II formula [17].

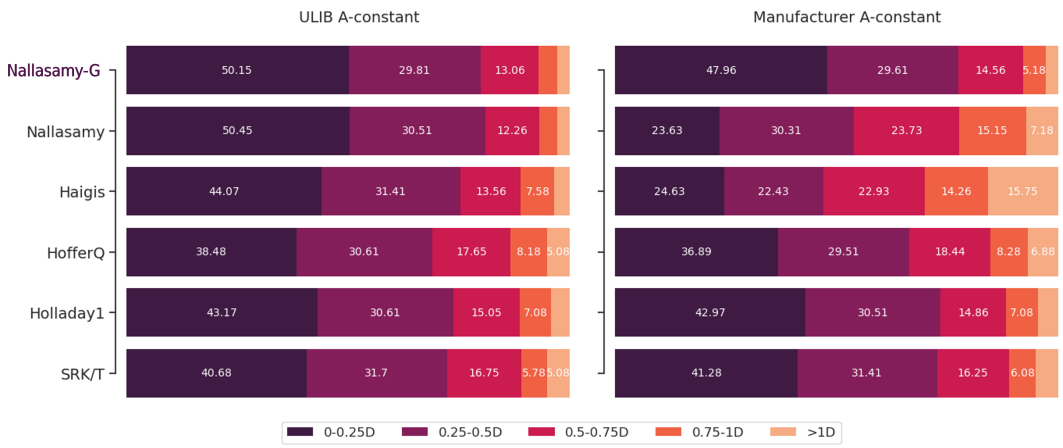

Figure 6.4: Predictive Error of the UMich dataset (Patients implanted with the Alcon SN60WF lens at the University of Michigan Kellogg Eye Center) broken down by diopter range of error.

6.3.7 Postoperative ACD and Radius of Corneal Power

Anterior Chamber Depth (ACD) is a measure from the anterior surface of the cornea to the iris (see Figure 6.1). While preoperative ACD measurements can be obtained, placement of the IOL during cataract surgery is known to change this number. While the calculation of postoperative ACD is widely disputed [173, 74, 136], this study follows the equations of Barrett in [17]:

$$pACD = AL - 0.593 + 0.13 - RG - \sqrt{RG^2 - RCP^2 + (RCP - ACD)^2} \tag{6.6}$$

, where AL is Axial Length, RG is radius of curvature of the globe's posterior segment, and RCP is peripheral radius of the cornea. While Barrett's process for pACD calculation appears to be iterative, we instead use preoperative ACD (ACD) on the right side of the equation with the assumption that preoperative ACD is a predictor of postoperative ACD. Radius of curvature (RG) is assessed with the following formula:

$$RG = 0.35066 \times AL - 0.06607 \times K + 5.70871 \tag{6.7}$$

,where K represents the mean of preoperative keratometry measurements for both eyes. RCP is given in [3] as the following formula:

$$RCP = (\sqrt{RC^2 + (1 - PZ) \times 5^2})^3 / RC^2 \tag{6.8}$$

, where PZ is the "p-factor" of the cornea, a measurement of asphericity. (P-factor is now

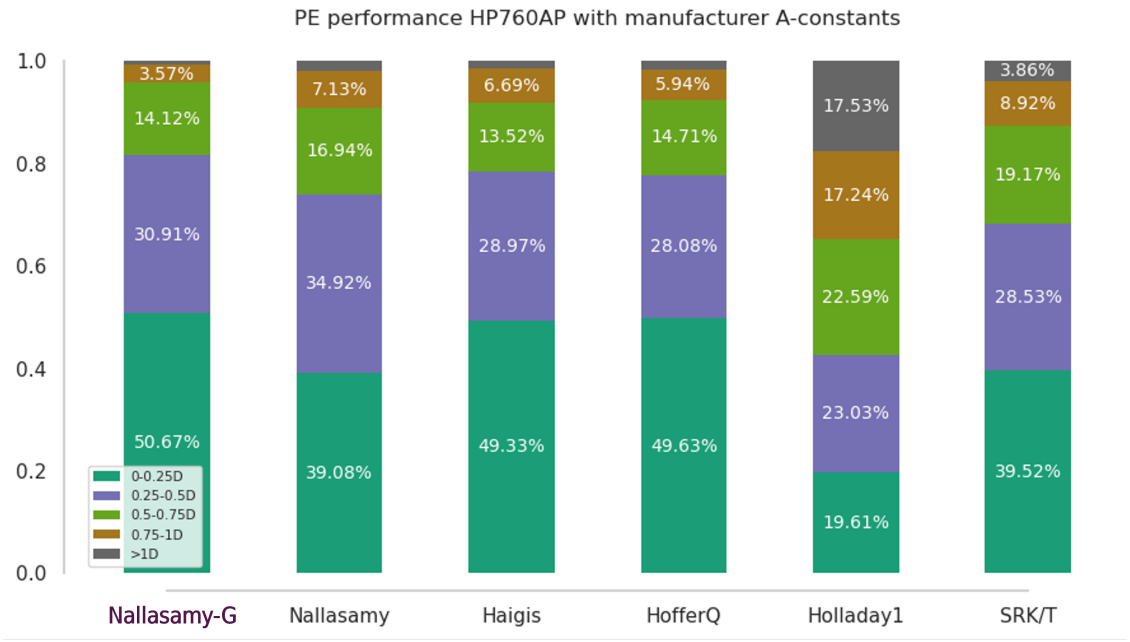

Figure 6.5: Prediction Error performance of HP760AP* models with Manufacturer A-constants. Our model demonstrated the highest percentages of errors below 0.25D, 0.75D and 1D compared to all other models presented. Given that the HP760AP* has no ULIB record for empirically derived constants, this dataset presents a perfect use case of generalized model and demonstrates that the model remains robust under Manufacturer A-constants.

called Q-value and has been shown in other studies to influence post-operative refraction outcomes [169]). RC represents the central radius of the cornea and is based on solving for RC in the following formula which defines the relationship between radii of the cornea and the power of the cornea, from [17]

$$KC = (376/RC) - (40/RCC) + (0.00052/1.376) \times (376/RC) \times (40/RCC) \tag{6.9}$$

. KC is the power of the cornea and was set to K, the mean of preoperative keratometry measurements for both eyes. RCC is the central posterior radius of the cornea and can be easily calculated as $RC \times 0.883$ based on empirically-observed similarities between the central radius (RC) and central posterior radius (RCP) (see Figure 6.1).

Note that while RC can be calculated with our empirically-obtained measurements, PZ cannot. Therefore, we conduct an iterative process for PZ which includes the SRK/T A-constant and white-to-white distance as a way to estimate best fit for predicted RCP.

In order to test whether or not the Q-value PZ is appropriate, a predicted RCP is calculated for some assigned PZ value based on the equation above. If the predicted RCP

with this assigned PZ is higher than RG, a new PZ is assigned and RCP is recalculated. Conversely, if the RCP is greater than RG, we next try to predict ciliary distance using RCP. Note that RCP and RG are both radii and thus form two intersecting circles (Figure 6.1). The distance between the two radii can be estimated to be roughly equivalent to $RG - LF$, where LF is the lens factor. The lens factor is defined as the distance between the ciliary plane or iris and the second principal plane. We strove to calculate a rough estimate of corneal distance to estimate the appropriateness of the predicted RCP.

Corneal distance (CD) is estimated to be equivalent to the length of the chord of the intersecting circles. Once CD is calculated based on the predicted RCP and RG, it is assessed for feasibility. CD must be greater than 0. Additionally, we use white-to-white (WTW), a measurable variable, to assess good fit. If the calculated CD from the predicted RCP is within a tolerable distance from the WTW calculation, the iterative process is stopped and both the predicted RCP and the resulting postoperative ACD are obtained from this fit. If not, the model iterates to a new assigned PZ until an appropriate fit is found. Note that some patients do not have a WTW measurement. If WTW is not available, an estimate of 12 mm is imputed here, as most human corneal diameters range from 11-13 mm [17].

6.3.8 Model improvement evaluation and validation

A five-fold cross validation was utilized for training and validation across the level-1 models. A stacking model consisting of the results of all level 1 models and based on linear regression was utilized in the training set on all five folds, then applied to the test set. All models were trained with the "UMich" Alcon SN60WF model but tested on all other models and populations.

Models were evaluated using refraction MAE, ME, MedAE, and Formula Performance Index (FPI). We define mean absolute error as the following:

$$\text{refraction MAE} = \frac{\sum_{i=1}^{n} |y_i - \hat{y_i}|}{n} \tag{6.10}$$

Mean error is defined simply as

$$ME = \frac{1}{n} \sum_{i=1}^{n} |y_i - \hat{y_i}| \tag{6.11}$$

. Lastly, we define FPI as the following, first defined by Hoffer et al [75]:

$$FPI = \frac{1}{SD + MedAE + 10 * abs(m) + 10 \times (n10)^{-1}} \tag{6.12}$$

, where m is defined as the slope coefficient of an ordinary least squares estimator where prediction error is regressed onto axial length, SD is sample standard deviation, and n represents the proportion of the dataset with an absolute predictive error < 0.5.

6.3.9 Statistical analysis

All analyses were performed using Python 3.8 with the scipy package.

To analyze the importance of each feature, we conducted multiple feature analyses of

Dataset	Formula	MAE	ME	MedAE	STD	AE< 0.5	m	FPI
A	**Nall-G**	**0.2936**	**0.1378**	**0.2339**	**0.4194**	**0.8434**	**0.0866**	**0.3697**
	Nall.	0.6471	0.6291	0.6071	0.435	0.3584	-0.1531	0.1864
	Haigis	1.2309	1.2239	1.215	0.5283	0.0575	-0.5198	0.0411
	HofferQ	0.3322	-0.1065	0.2611	0.4641	0.7947	-0.2135	0.2428
	Holl.	0.2848	0.0054	0.2041	0.4395	0.8522	-0.0483	0.4348
	SRK/T	0.297	-0.055	0.233	0.4378	0.8575	-0.0634	0.4048
B	**Nall-G**	**0.2891**	**-0.1088**	**0.2469**	**0.3541**	**0.8158**	**0.2041**	**0.2586**
	Nall.	0.3626	0.2817	0.3144	0.3439	0.74	-0.0851	0.3496
	Haigis	0.6121	0.5839	0.566	0.4135	0.4264	-0.4568	0.1267
	HofferQ	0.3882	-0.2902	0.3291	0.3873	0.6805	-0.1953	0.2416
	Holl.	0.3204	-0.1922	0.2527	0.364	0.7771	-0.0049	0.5123
	SRK/T	0.322	-0.1814	0.2532	0.3751	0.7831	-0.1152	0.3271
C	**Nall-G**	**0.2525**	**0.0473**	**0.2126**	**0.3154**	**0.8883**	**-0.1339**	**0.3341**
	Nall.	0.547	0.5251	0.5417	0.3311	0.4447	-0.4941	0.124
	Haigis	0.8298	0.8104	0.8155	0.4406	0.2122	-0.8537	0.0689
	HofferQ	0.3051	-0.0337	0.2494	0.387	0.8234	-0.694	0.1138
	Holl.	0.2729	0.052	0.224	0.3426	0.8538	-0.5947	0.1301
	SRK/T	0.2782	0.0868	0.2342	0.3423	0.8416	-0.7156	0.1121
D	**Nall-G**	**0.3316**	**0.1355**	**0.2612**	**0.4155**	**0.7757**	**-0.2433**	**0.2273**
	Nall.	0.5155	0.4411	0.4702	0.4301	0.5394	-0.6601	0.1069
	Haigis	0.582	0.4608	0.5332	0.5424	0.4706	-1.2086	0.0654
	HofferQ	0.4233	-0.2212	0.3468	0.4984	0.664	-0.7949	0.0971
	Holl.	0.3744	-0.0963	0.2904	0.4826	0.7348	-0.6978	0.1097
	SRK/T	0.3775	-0.0633	0.3112	0.4843	0.7268	-0.4304	0.1544

Table 6.5: Performance comparison of our model against five formulas (the Nallasamy formula, Haigis, HofferQ, Holladay1, and SRK/T formula) for the following datasets: A) FH5600AS, B) HP760AP*, C) SN60WF (at Aravind), D) UMich (patients implanted with SN60WF). All models were constructed with Manufacturer A-Constants.

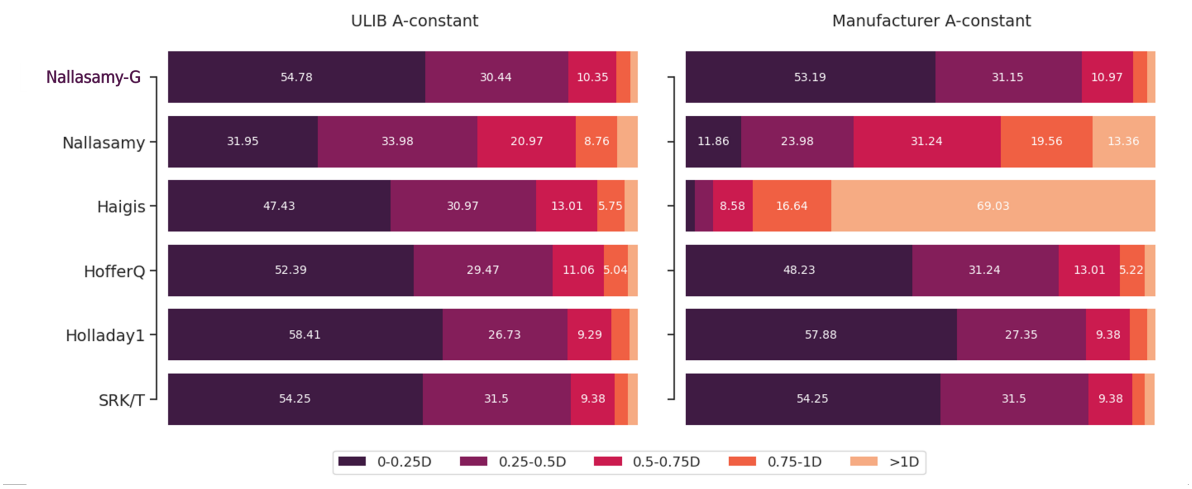

Figure 6.6: Prediction Error for Aurolab FH5600AS lens (Aravind). Bar labels show the percentage of patients within each error category. No labels are shown for percentages less than 5%. Errors less than 0.5 D are considered fair.

both the model and its performance under variations in each dataset's features. As tree-based models have an intrinsic ability to denote feature importance, two tree-based models inside of our ensemble model were analyzed on feature importance using scikit-learn and also compared to our predecessor, the Nallasamy formula. In addition, partial dependence plots were constructed for the Stacking Regressor to analyze the impact of incremental changes to specific features on the output error. ICE plots were also constructed to analyze the impact of IOL formulas as features in our model. Lastly, a univariate analysis consisting of each of the new features' abilities to predict post-operative refraction were deployed using statsmodels.api in Python. Correlation coefficients of variables versus post-operative refraction were calculated with the pearsonr() function in scipy.stats.

6.4 Results

6.4.1 Data characteristics

Count of Eyes by Axial Length can be seen in Fig 6.2. Among the datasets collected at Aravind (FH5600AS, HP760AP*, SN60WF), the most common ALs were between 21-23 mm, with a slight left skew to higher ALs. The Aravind datasets for the SN60WF and FH5600AS lenses contain higher proportions of short ALs ($<$ 22.0 mm) compared ot the UMich dataset. The UMich dataset, however, contained a comparatively larger proportion of long ALs ($<$ 26.0 mm). However, note that even for the UMich dataset, there is still only

a small proportion of long ALs available (N=9).

Demographic information for the UMich dataset, which made up training and testing, can be found in Table 6.2. Demographic information broken up by sex for the HP760AP*, FH5600AS, SN60WF and UMich datasets can be found in Chapter 5 table 5.2. The mean IOL power contained in the dataset was 19.79 D with a standard deviation 3.71 D. As sex was labeled "1" for a male and "0" for a female, the mean of 0.43 signifies a slightly larger proportion of female patients in the dataset. Both mean (-0.55 D) and median (-0.41 D) refraction for the UMich dataset was negative.

6.4.2 Model performance improvement

Performance breakdown of the Nallasamy-G generalized model on the UMich test set are given in Table 6.4. The level-2 Stacking Model performed with an MAE of 0.3098 and a ME of 0.0331. It managed to predict postoperative refraction with 0.5 D for 80.16% of the dataset. Although not all single models in the level-1 ensemble performed better than the best comparable IOL formula, Kane (MAE: 0.3148), the level-2 Stacking model (our final generalized model), performed better in MAE, MedAE (0.2424 vs Kane 0.2436) and percentage of errors within 0.5 D (0.8016 v Kane 0.7986).

The breakdown of feature importance for the two level 1 models can be viewed in figure 6.9. The breakdown of feature importance in the same models for the Nallasamy formula can be viewed in figure 6.12. In the first model, the most important features are the two A-constant based features and the modified Barrett predictions. For the second model, two A-constant based features and the modified Barrett are also the top three features. In the original Nallasamy formula, all top features were A-constant based. Note that our modified Barrett is among the top features in the Nallasamy-G model and does not require an A-constant. Also note that the importance of anatomy-based features are higher in Nallasamy-G than in the Nallasamy formula, where feature importance for anatomy-based features appears close to zero.

Partial Dependence plots demonstrating the effects of the input variables on refraction prediction error for FH5600AS are given in figure 6.13, for HP760AP* are given in figure 6.14, for SN60WF (Aravind) are given in figure 6.15, and for UMich are given in figure 6.16. Partial Dependence plots show how the average MAE is influenced by incremental changes in a single feature. The graphs for all datasets demonstrate a similar pattern among the input variables. CCT and age at surgery appear to have some of the largest impacts on change in refraction error.

Partial Dependendence graphs demonstrating the effect of predicted RCP, our modi-

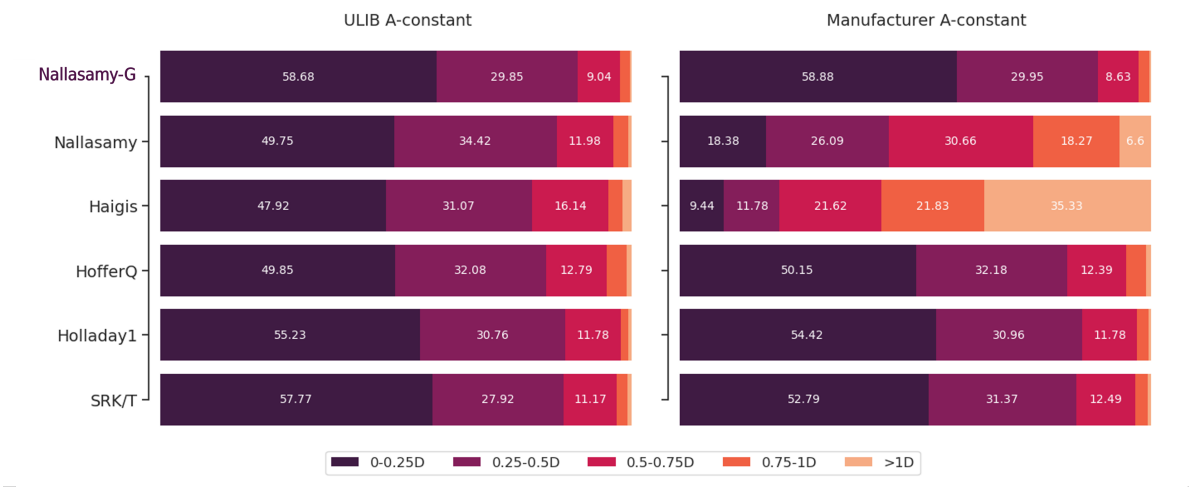

Figure 6.7: Patient predictive error for Alcon SN60WF at Aravind. Labels on the bars represent percentages of patients within the error range.

fied Barrett formula, and the ThinLens formula on refraction prediction error are given for FH5600AS in figure 6.17, for HP760AP* in figure 6.18, for SN60WF in figure 6.19 and for UMich in figure 6.20. For all datasets, the Barrett Modified refraction formula demonstrates the highest slope, or the strongest influence on refraction prediction error of the three features listed. This signifies that predictions made by the Barrett Modified formula feature are among the most powerful in swaying the prediction.

Finally, ICE graphs demonstrating the A-constant-based features from the Nallasamy Formula compared with ThinLens and Barrett Modified formula features are given for FH5600AS in figure 6.21, for HP760AP* in figure 6.22, for SN60WF in figure 6.8, and for UMich in figure 6.23. ICE graphs show partial dependence like a Partial Dependence plot, but also indicate how a single instance's MAE changes when the feature changes. For all datasets, the Barrett Modified formula demonstrates one of the highest slopes of the six IOL formula features, indicating that changes in this feature have among the highest influences towards refraction prediction and subsequently refraction prediction error. Note that in the UMich dataset, two of the original Nallasamy Formula A-constant based features also appear to have steeper slopes, but this effect is less apparent in the Aravind datasets, possibly signifying our modified Barrett as the primary determinant of postoperative refraction when A-constants are less reliable.

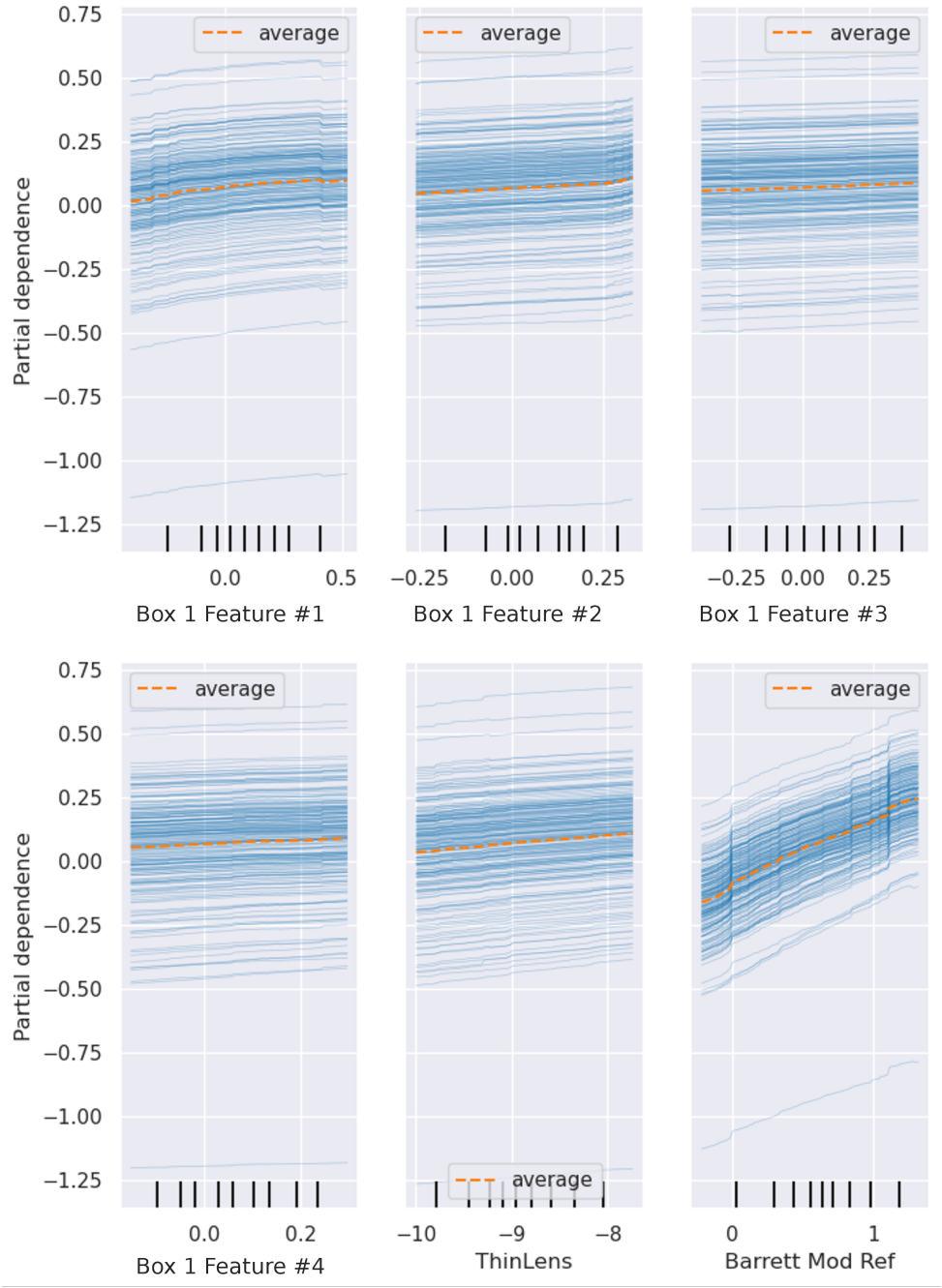

Figure 6.8: Partial dependence plots for IOL formula-based input features of the SN60WF (Aravind) lens dataset. Empirical constants are being used to demonstrate the model's performance in the first use case.

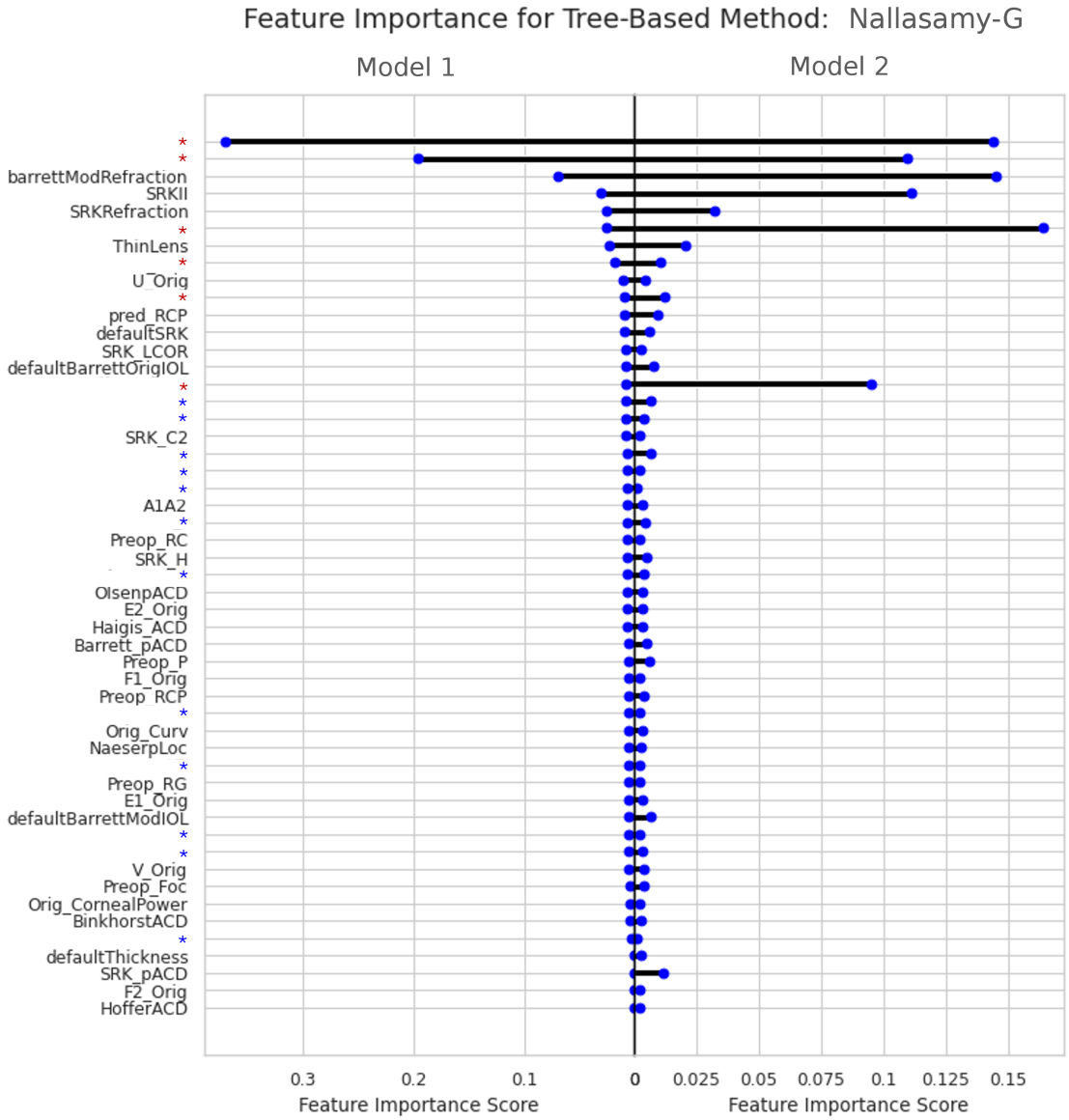

Figure 6.9: Feature Importance of Our Generalized Method (Nallasamy-G). The feature importances were extracted from two tree-based level 1 models. Asterisks denote features belonging to the original Nallasamy formula and the color of Asterisk indicates the category of each feature (Box 1 A-constant based, Box 2 Non A-constant anatomic, Box 3 constant)

6.4.3 Model generalization performance

Performance of Nallasamy-G on the UMich test set under different A-constants can be seen in Figure 6.4. Under the experimentally-derived A-constants from ULIB, our method slightly underperforms at 50.15% predictions within 0.25 D and 79.96% predictions within 0.5 D compared to the Nallasamy Formula which performs at 50.45% of predictions within 0.25 D and 80.96% predictions within 0.5 D, but our model remains among the top IOL formulas compared Haigis (44.07% at < 0.25 D and 75.48% at < 0.5 D), HofferQ (38.48% at < 0.25 D and 69.09% at < 0.5 D), Holladay 1(43.17% at < 0.25 D and 73.78% at < 0.5 D), and SRK/T formulas (40.68% at < 0.25 D and 72.38% at < 0.5 D). When comparing formulas under the Manufacturer A-constant, our generalized Nallasamy-G formula demonstrates not only the highest percentages of errors below 0.25 D (47.96%) and 0.5 D (77.57%), but also significant improvement over the Nallasamy Formula, which exhibits only 23.63% of predictions with errors within 0.25 D and 53.94% of predictions with errors within 0.5 D. The Haigis formula is also notably low, with only 24.63% of errors within 0.25 D and 47.06% of errors within 0.5 D.

Performance of our domain-generalized Nallasamy-G method on all datasets with Manufacturer A-constants compared with the Nallasamy formula, Haigis, HofferQ, Holladay 1, and SRK/T are given in Table 6.5, as measured by MAE, ME, MedAE, Standard Deviation (STD), Absolute Error (AE)< 0.5 and FPI. For all datasets, our Nallasamy-G model significantly outperforms the Nallasamy Formula under manufacturer A-constants, demonstrated lower MAE, ME and MedAE, with higher percentages of absolute error below 0.5 D. FPI is also higher with Nallasamy-G than for the Nallasamy Formula. When compared to other IOL formulas to which a manufacturer A-constant could be converted, our Nallasamy-G demonstrates competitive performance across all datasets. MAEs for UMich, SN60WF (Aravind) and HP760AP are the lowest with the Nallasamy-G model compared to all other formulas. For the equiconvex FH5600AS dataset, our model is only beat by Holladay 1 in MAE. We also note that MedAE remains the lowest among all other competing formulas for the UMich, SN60WF (Aravind), and HP760AP* datasets, but in the FH5600AS lens our Nallasamy-G generalized model produces the third lowest MedAE.

One critical observation from this chart is that Haigis is the worst performer of all formulas under manufacturer A-constants. Note that the $a1$ and $a2$ constants provided by the ULIB A-constant converter for manufacturer A-constants were Haigis' default values of $a1 = 0.4, a2 = 0.1$. This differs from the $a1$ and $a2$ constants trained in the model, which were specifically tailored for the SN60WF.

Performance of our method with ULIB A-constants are given in Table 6.9. Note that HP760AP does not have ULIB A-constants and therefore does not appear in the table. Under

these empirical constants, note that our model makes improvements upon the Nallasamy Formula as measured by MAE and FPI. Although ME is slightly worse with the UMich dataset, it is dramatically improved for the SN60WF at Aravind dataset and FH5600AS. Additionally, although MedAE reamins the same for the UMich dataset compared with the Nallasamy Formula, there is some improvement with the SN60WF at Aravind dataset and dramatic improvement with the FH5600AS dataset. When compared to the other IOL formulas, our Nallasamy-G model remains competitive. Across both SN60WF datasets, at Aravind and UMich, we see the best MAE, STD and FPI. With our equiconvex FH5600AS lens dataset at Aravind, we note MAE and MedAE of Nallasamy-G was second-best to the Holladay 1 Formula. Critically, we note for all lenses that the absolute ME of Nallasamy-G is less than 0.1, which signifies a relatively "well-centered" model under empirical A-constants, even for the FH5600AS lens dataset.

A stacked bar chart visualization of performance break down by absolute error can be seen in Figure 6.6 for FH5600AS lens dataset. Note that for the FH5600AS dataset, Haigis performs particularly poorly for under the manufacturer A-constant, with less than 5% of errors below 0.25 D and 0.5 D, respectively. Accordingly, the Nallasamy Formula appears to also produce very low predictions, with only 11.86% predictions below 0.25 D. This is in stark contrast to its performance with ULIB empirical A-constants, where 31.95% of predictions are below 0.25 D. Contrastly, Nallasamy-G remains robust across empirical and manufacturer A-constants, demonstrating 54.78% of errors below 0.25 D with ULIB empirical A-constants and 53.19% of errors below 0.25 D with the manufacturer A-constant. Our model's performance is similar to SRK/T.

Stacked bar charts for the SN60WF at Aravind dataset are given in Figure 6.7. Recall that this lens is the same as the training set, but population demographics differ. Note that once again, the Nallasamy Formula follows a similar pattern with Haigis, where performance dramatically drops in the presence of the manufacturer A-constant (Haigis: 9.44% errors within 0.25 D and Nallasamy 18.38% errors below 0.25 D). Nallasamy-G demonstrates the highest absolute error percentages below 0.25 D (56.68% under ULIB A-constants and 56.88% under manufacturer A-constant) and below 0.5 D (88.53% under ULIB A-constant and 88.63% under manufacturer A-constant) compared to all other formulas, including the next most competitive formula, Holladay 1 (55.23% and 54.42% under 0.25 D error and 85.99% and 86.38% under 0.5 D error under ULIB and manufacturer A-constants, respectively).

Since HP760AP* was not listed in the ULIB A-constant list, a performance analysis under manufacturer A-constants only is give in figure 6.5. Our model outperforms all other models in under 0.25 D error (50.67% compared with 39.08% Nallasamy, 49.33% Haigis, 49.63% Hoffer Q, 19.61% Holaday1, 39.52% SRK/T) and 0.5 D (81.51% compared to 74.0%

Nallasamy, 78.0% Haigis, 77.71% Hoffer Q, 42.64% Holladay1, 68.05% SRK/T).

All model performances using optimized A-constants is given and described in Appendix H.

6.4.4 Retraining performance

A line graph comparison of performance of retraining the level 2 Stacking Model can be seen in figure 6.10. Both the FH5600AS and HP760AP* lenses appear to consistently perform better than the pre-trained model after about 75 samples. For the SN60WF lens implanted at Aravind, the model improves over the pre-trained model after approximately 250 samples.

6.4.5 IOL Formula Correlations and Univariate Analysis

Correlations of all IOL formula post-operative refraction predictions against ground truth are given in Table 6.8. Both ML-based methods, including our generalized Nallasamy-G and the Nallasamy Formula, outperformed the other comparative methods.

Univariate Analysis results sorted by correlation metric ρ can be seen in Table 6.7. Our modified Barrett, SRKII and SRK refraction predictions appear to be among the top predictors of post-operative refraction with ρ correlation coefficients of 0.8435, 0.7566, and 0.6759, respectively. As individual predictors of post-operative refraction, they explain 71.15%, 57.24% and 45.68% of the variability observed in post-operative refraction, respectively. Our modified Barrett formula remains among the fourth best correlates overall when including the Nallasamy Formula features, explaining its presence among the top predictive features from the tree-based feature analyses.

6.5 Discussion

It is important for cataract surgeons to have a proper understanding of the relationship between an implanted lens and the post-operative refractions of patients after cataract surgery. Wrongful predictions can result in the need for corrective lenses or a second surgery to correct for errors. The previously published AI-based Nallasamy formula [106] was demonstrated to predict post-operative refraction outcomes with high accuracy, serving as a competitive alternative to other formulas such as Barrett Universal II, PearlDGS, SRK/T, Holladay 1, Hoffer Q and Haigis formulas for the Alcon SN60WF lens. However, the Nallasamy formula was never constructed for lens models other than the Alcon SN60WF and was shown here to generalize poorly to even populations with the same lens implant if A-constants no longer match the trained A-constant set. This is because model weights and trained parameters

differ between lens model domains, meaning the Nallasamy formula trained for SN60WF does not extrapolate to different lens types. Therefore, in this study we sought to increase the generalizability of the Nallasamy formula to other lens domains and lower the model's overall dependence on A-constants so as to robustify it.

In our revised version of the Nallasamy formula, we assessed generalization using one testing dataset from the University of Michigan in Ann Arbor and three datasets obtained Aravind Eye Hospital in Tamil Nadu. In the Aravind datasets, one dataset tested Nallasamy-G on the same lens implant in a different population, while the other two tested its performance on two different lens types. Given the difficult nature of obtaining optimized A-constants, we obtained two separate A-constant sets for our datasets under the assumption that a user would not have enough data from which to optimize an A-constant and therefore leverage an online public library (ULIB) for A-constants or would prefer to use the manufacturer's A-constant.

The testing set performance of the generalized Nallasamy-G demonstrated some small difference with the Nallasamy Formula under ULIB constants but a dramatic improvement over the Nallasamy Formula with only the manufacturer A-constant available. When compared to the Aravind datasets, Nallasamy-G was shown to demonstrate consistently low MAEs with high percentages of per-lens diopter error less than 0.25 D and 0.5 D in post-operative refraction prediction assessments across all three datasets under both A-constant scenarios, ranking either first or competitively among four non-proprietary IOL formulas

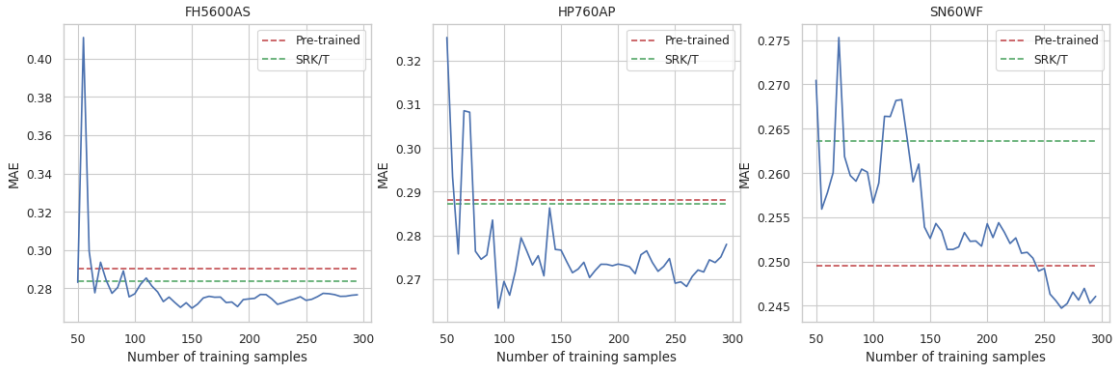

Figure 6.10: Retrain Performance of Different Lenses. For each of the lenses implanted at Aravind, a number of training samples were presented as retraining material for the pre-trained level 1 Stack Regressor. The x-axis illustrates the number of data points given to the level 1 Stack Regressor as training data and the y-axis gives the MAE of the model after retraining. The dashed lines indicate baseline comparators: in red, the pre-trained generalized Nallasamy-G model's performance without retraining, and in green, the SRK/T value. Note that the A-constants used for this analysis are optimized A-constants.

where A-constants could be obtained. Critically, Nallasamy-G showed in some cases dramatic improvement over the Nallasamy Formula for A-constant sets drastically different from the SN60WF (Manufacturer sets), or lenses different from the SN60WF. For the latter, this is most notable for the equiconvex FH5600AS dataset, because the lens thickness of an equiconvex lens is different from anterior- or posterior-dominant lenses, affecting the position of the lens planes.

The generalized Nallasamy-G model is the product of an expanded ensemble consisting of more opportunities for mathematical interactions between features to be found as well as the inclusion of an expanded set of features that do not require an A-constant. Notably, we have shown part of the model's strength to be related to the inclusion of our own modified Barrett formula, which we demonstrated through partial dependence plots and feature importance graphs to be a strong influence on the model and one of the model's most important features for two tree-based models in the ensemble. Independently, we verified that the Modified Barrett was shown to be the third most correlated feature with post-operative refraction prediction in a post-hoc univariate analysis.

We also found our algorithm for predicted radius of peripheral cornea (RCP) value to appear in the top 12 most important features in both level 1 models from our feature importance analysis. In partial dependence analyses, the predicted RCP variable demonstrated influence on Nallasamy-G that was similar to the performance of the ThinLens formula. The predicted RCP feature made use of white-to-white measurements and the SRK/T A-constant to predict a corneal radius measurement of the eye.

The use case of our model is specifically designed for scenarios where optimized A-constants cannot be derived due to a lack of data. In these cases, our method demonstrates promise as a robust method for lower mean absolute error compared to our more accurate and robust postoperative refraction predictions compared to our previous model, the Nallasamy formula, and is competitive with other models shown here. We also tested a paradigm where enough data was available to retrain the Stacking layer of Nallasamy-G and attempted to determine the amount of samples needed to improve the model under optimized constants. While each lens retrain resulted in a different number of training samples to improve the MAE of Nallasamy-G, we found that all tests demonstrated improvement at or before 250 samples.

However, this study still has many limitations that may require more work to ameliorate before the model is ready for real-world clinical use. This study was designed to improve upon the well-predictive Nallasamy Formula so that it could generalize to other lens types, with the significant constraint being that the training data comes from only a single lens model (SN60WF) in a single population (University of Michigan). This study finds that

generalization of the model to the lenses provided is possible under certain conditions such as understanding some information about the target lens geometry. However, more data from other lenses is recommended to understand the limitations of the generalized formula beyond the lenses provided. However, if enough data can be provided, it may be more ideal to simply train individual ML models for similar lens geometries for the most accurate results instead of using a generalized model.

Another limitations is that the generalized model is not as easy to use as the Nallasamy Formula. Firstly, the model must know some lens gemoetry to generalize well and this is not always easy to find, as some manufacturers may consider it proprietary. Second, since the model is built from the framework of the original Nallasamy Formula, the strongest features necessary for our generalized model required multiple A-constants. While we tried to ameliorate this inconvenience by demonstrating the model's efficacy with A-constants produced via a conversion of the manufacturer A-constant to the A-constants needed, this is far from ideal and could possibly lead to issues when tested on other lens geometries.

In conclusion, our generalized Nallasamy-G model provides a proof-of-concept alternative for the Nallasamy Formula to pursue generalization to additional lens constants under different A-constant scenarios. Not only did Nallasamy-G demonstrate robustness when assessed on a different patient population, but also generalized adequately to two additional lens models, even if only the manufacturer A-constant was known. The performance of our model was comparable to the performance of other known IOL formulas with experimental ULIB A-constants available. While we believe the model may need further testing on expanded, we have shown the informed construct of including features with lower dependence on A-constants does open the model to better post-operative refraction predictions for lens models not included in training.

6.6 Acknowledgement

This chapter is part of a written work (unpublished): Elisa Warner, Miles Greenwald, Dongxiao Yan, Tingyang Li, Prashanth Gupta, Jyothi Vempati, Karthik Srinivasan, Haripriya Aravind, Nambi Nallasamy. An Exploration of Generalizing ML-based IOL Formulas.

6.7 Supplementary Materials

6.7.1 Supplementary figures

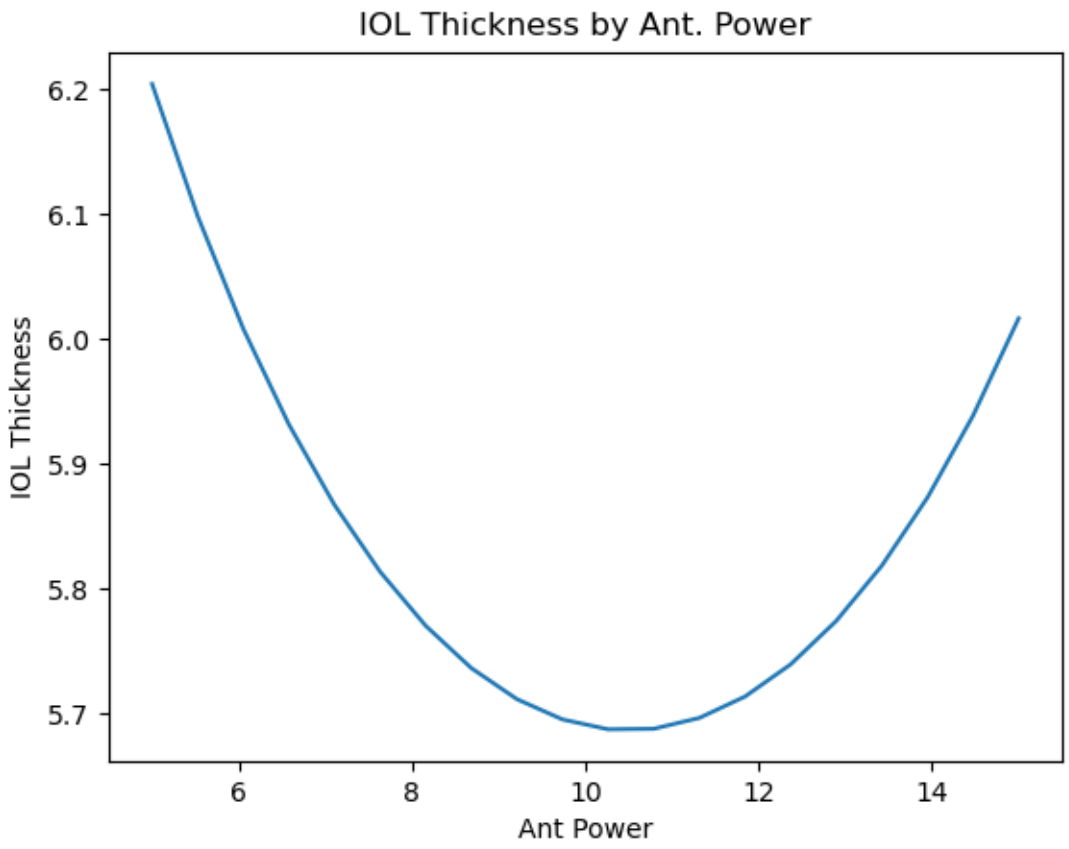

Figure 6.11: A view of how total IOL thickness depends on the thickness of each lens. In this figure, we show lens thickness for a total IOL power of 21.0 D. On the x-axis is anterior lens power and on the y-axis is total IOL thickness. Note that the total thickness of the lens is the least when both anterior and posterior powers are equal.

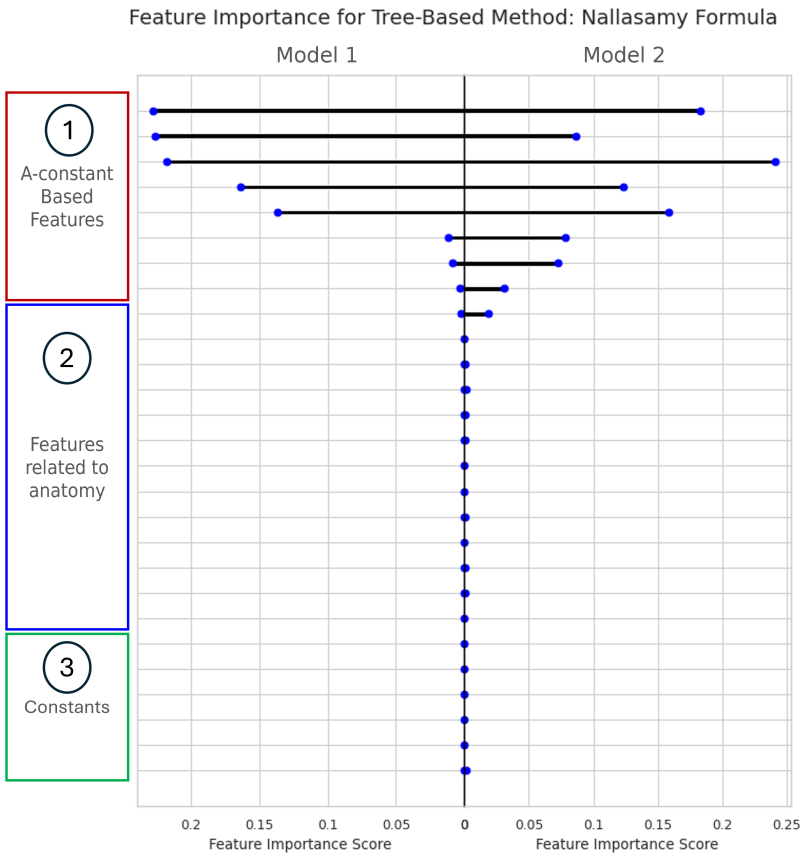

Figure 6.12: Feature Importance of the Nallasamy formula. The feature importances were extracted from two level 1 regressors. Features have been categorized into three groups: Box 1) A-constant based features, which were among the best-performing features, Box 2) Non A-constant-based predictions of anatomic measurements, Box 3) constants or near-constant values, which overall showed little relative importance in model predictions compared to the other feature categories.

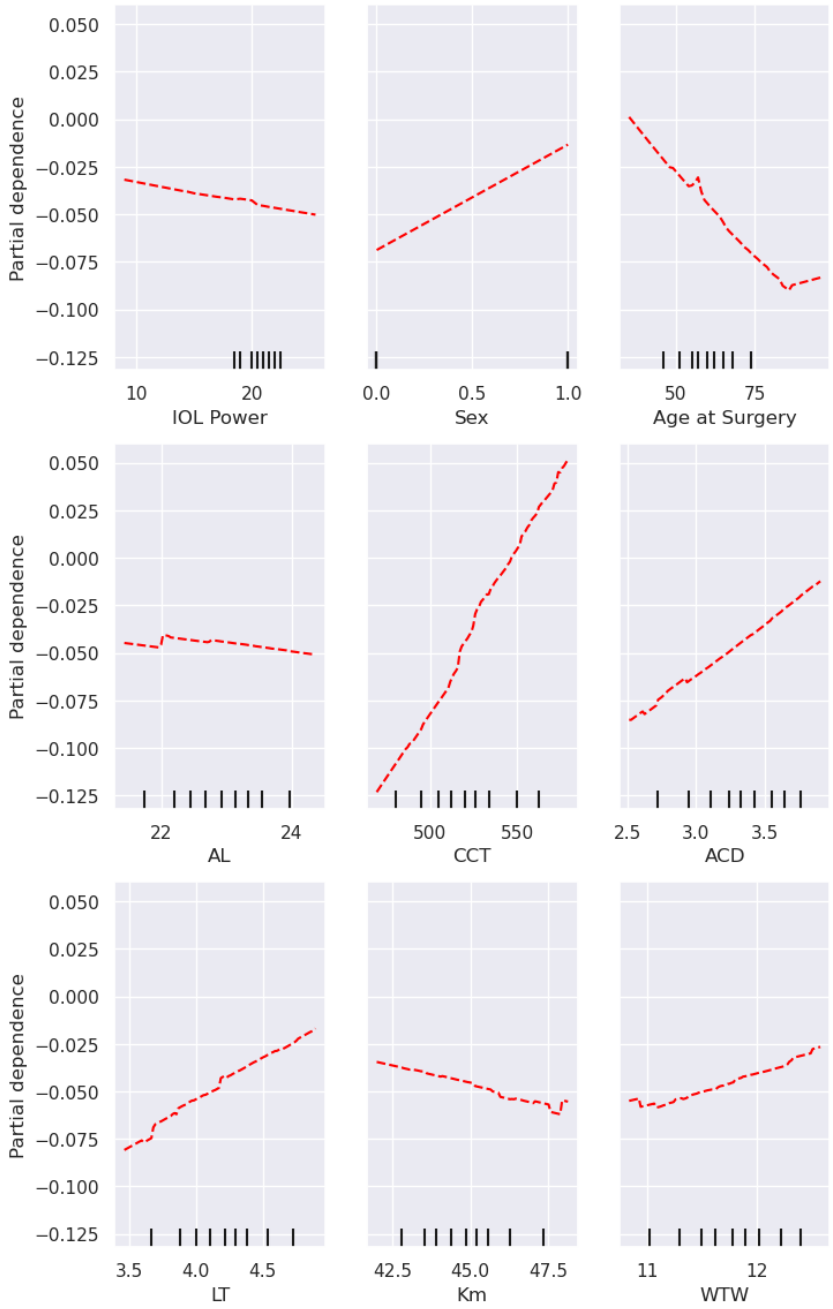

Figure 6.13: Partial dependence plots for patient biometric input features of the FH5600AS lens dataset. Empirical constants are being used to demonstrate the model's performance in the first use case.

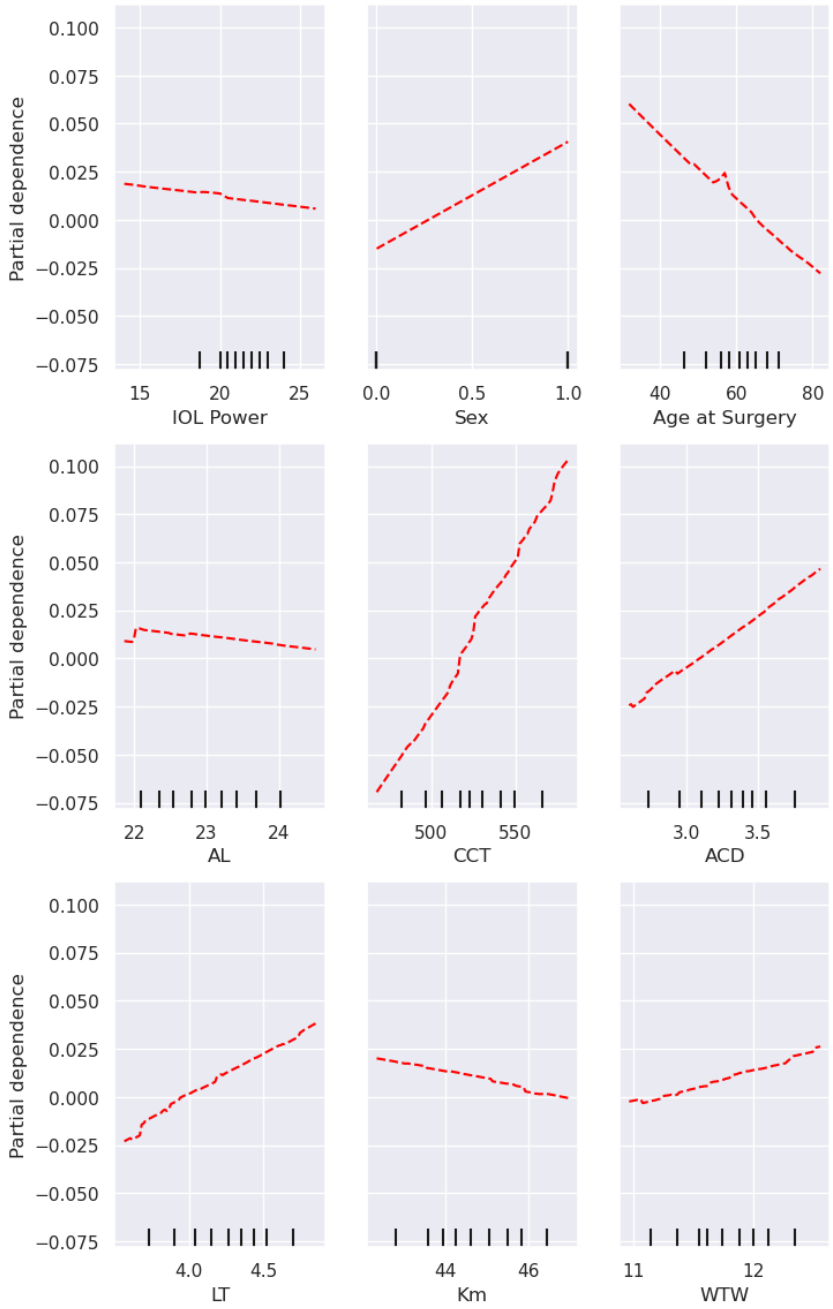

Figure 6.14: Partial dependence plots for patient biometric input features of the HP760AP* lens dataset. Empirical constants are being used to demonstrate the model's performance in the first use case.

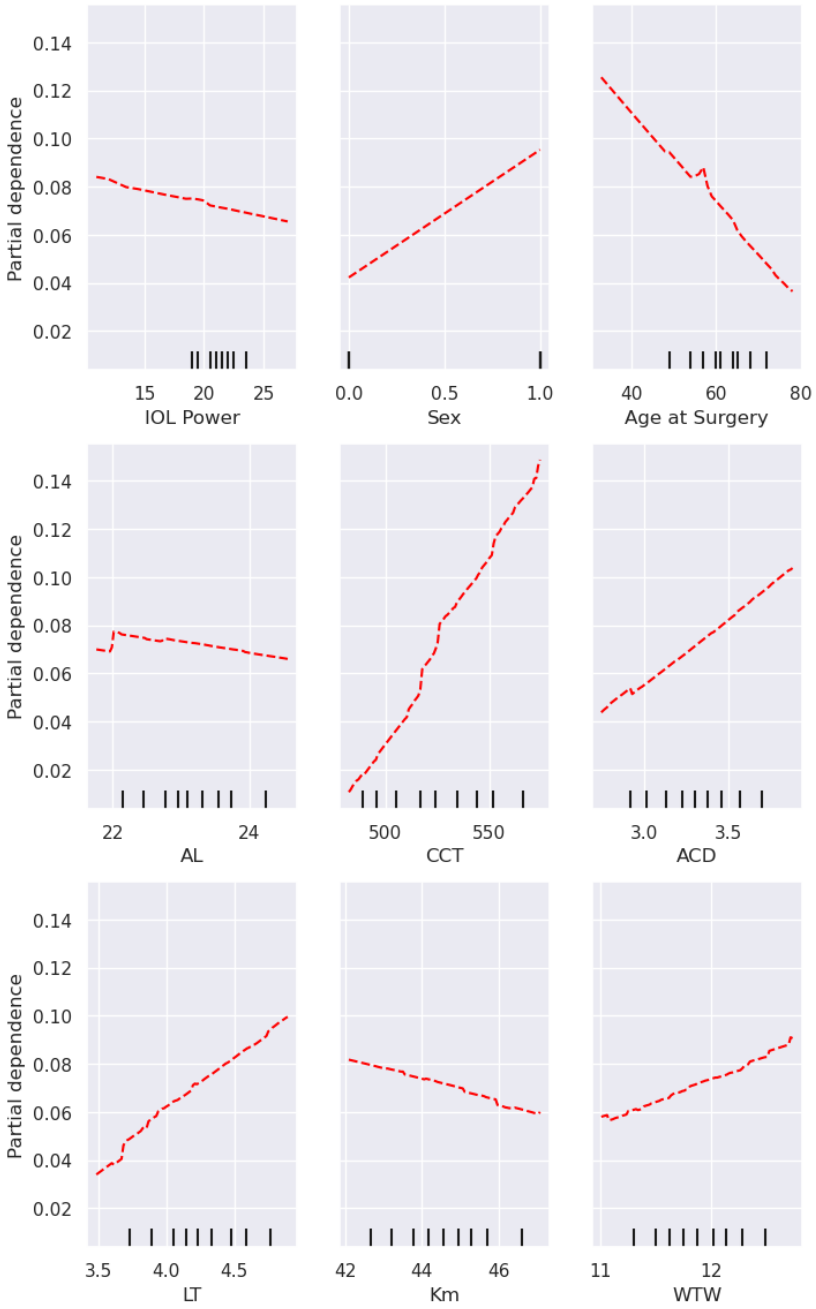

Figure 6.15: Partial dependence plots for patient biometric input features of the SN60WF (Aravind) lens dataset. Empirical constants are being used to demonstrate the model's performance in the first use case.

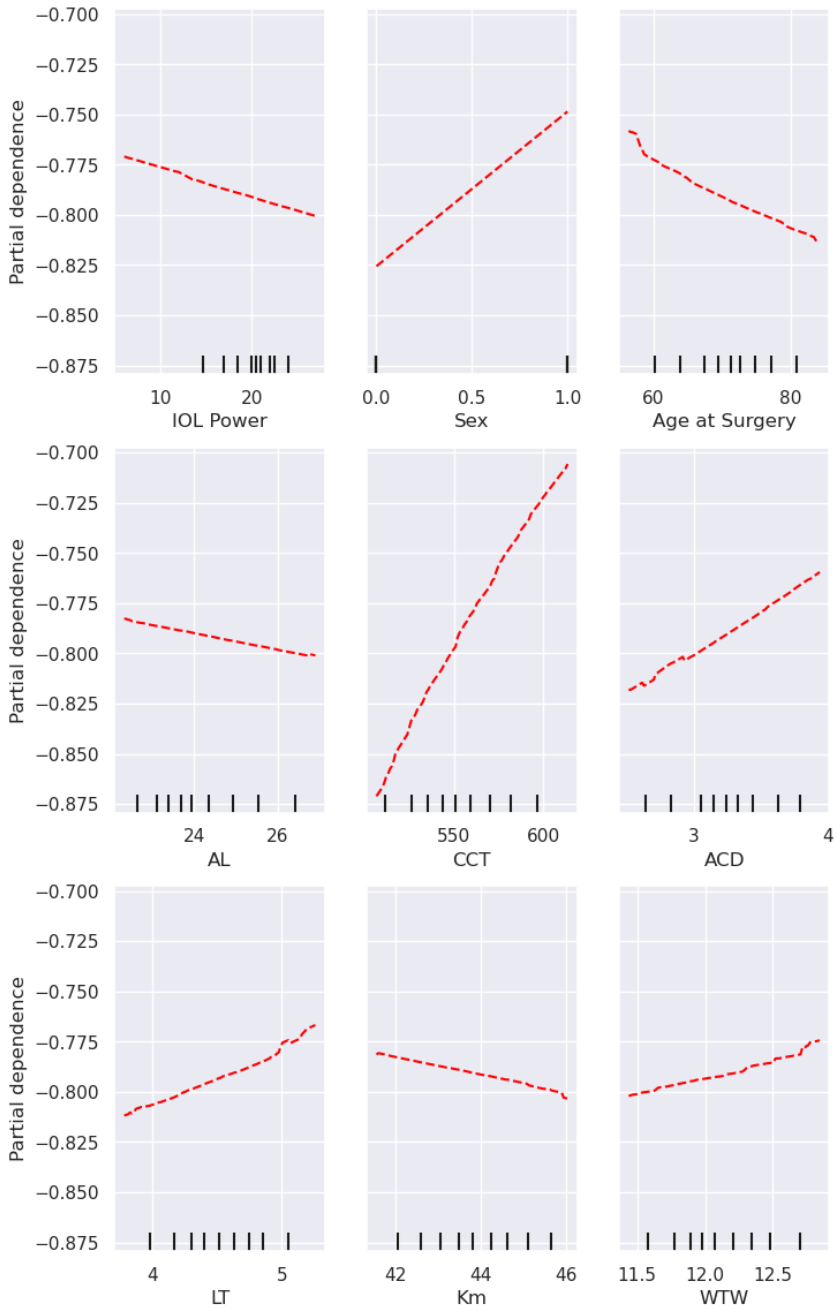

Figure 6.16: Partial dependence plots for patient biometric input features of the SN60WF (University of Michigan) lens dataset. Empirical constants are being used to demonstrate the model's performance in the first use case.

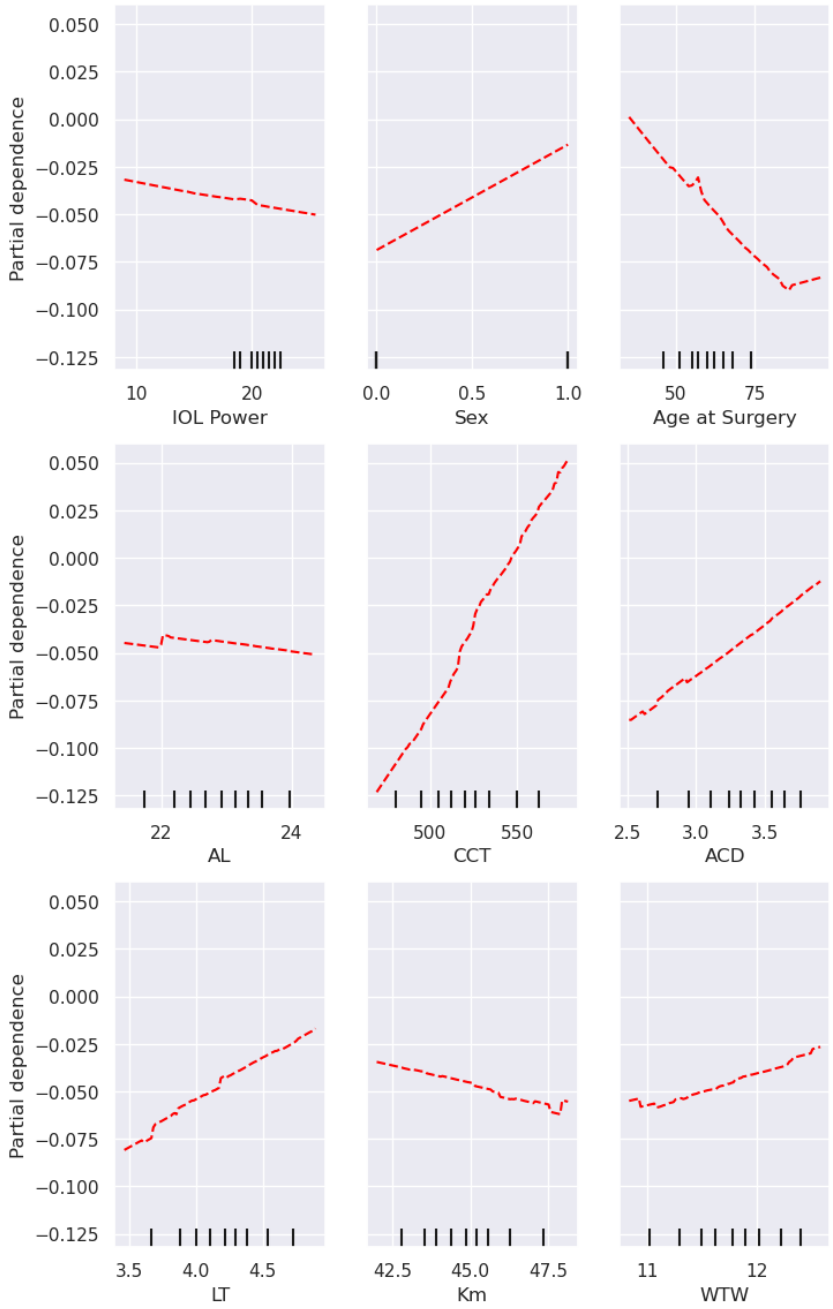

Figure 6.17: Partial dependence plots for patient biometric input features of the FH5600AS lens dataset. Empirical constants are being used to demonstrate the model's performance in the first use case.

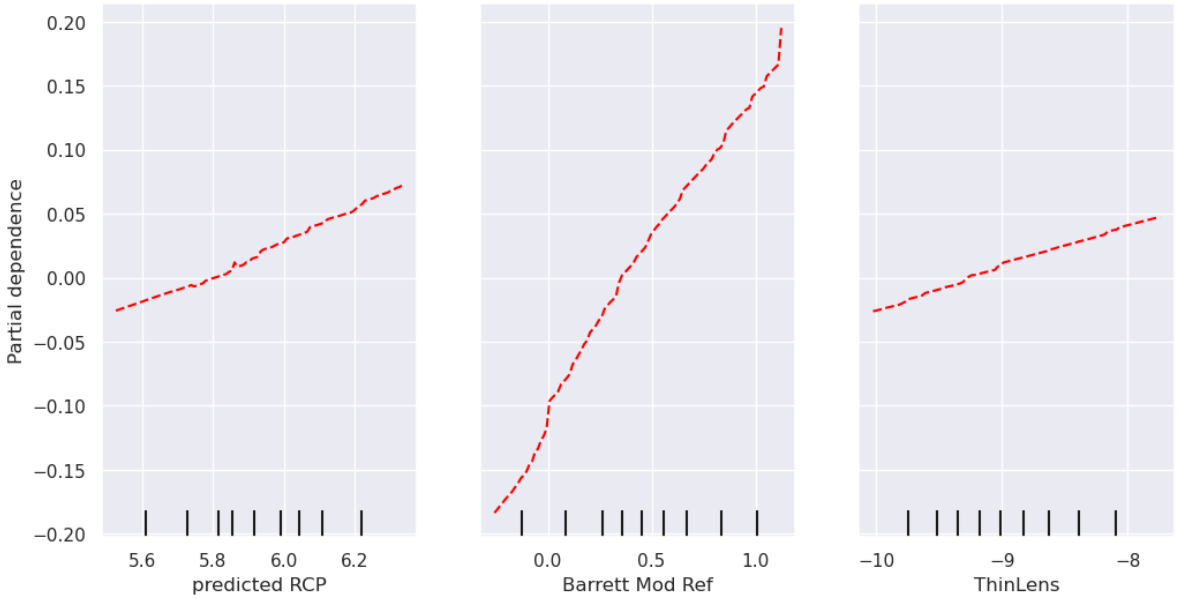

Figure 6.18: Partial dependence plots for three new input features in Nallasamy-G tested on the HP760AP* lens dataset. Empirical constants are being used to demonstrate the model's performance in the first use case.

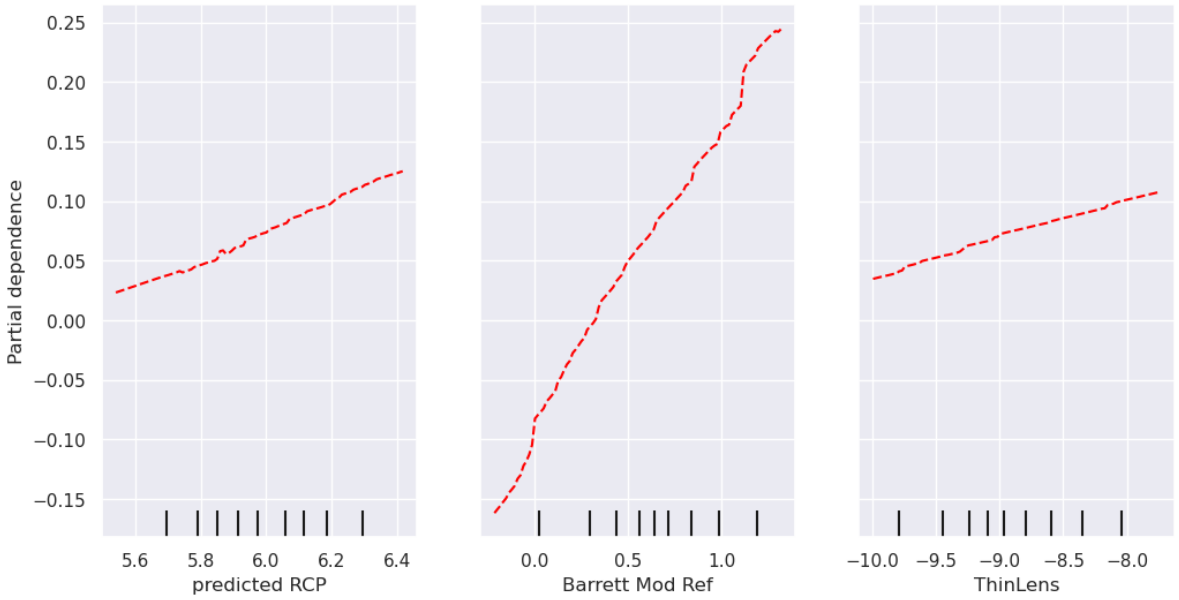

Figure 6.19: Partial dependence for three new input features in Nallasamy-G tested on the SN60WF (Aravind) lens dataset. Empirical constants are being used to demonstrate the model's performance in the first use case.

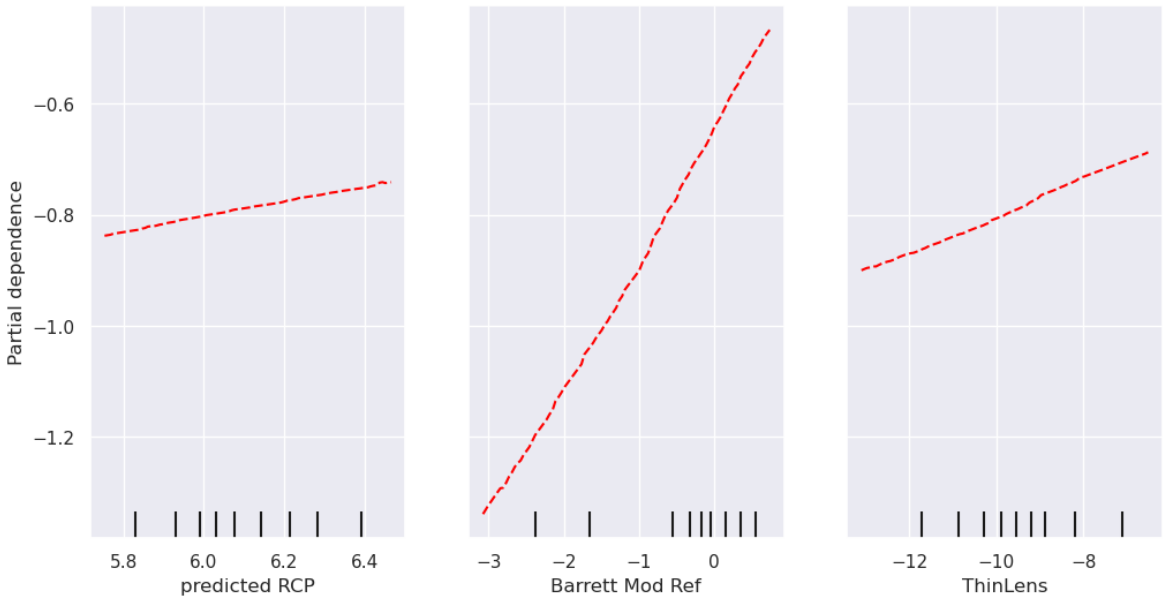

Figure 6.20: Partial dependence plots for three new input features in our generalized Nallasamy-G model tested on the SN60WF (University of Michigan) lens dataset. Empirical constants are being used to demonstrate the model's performance in the first use case.

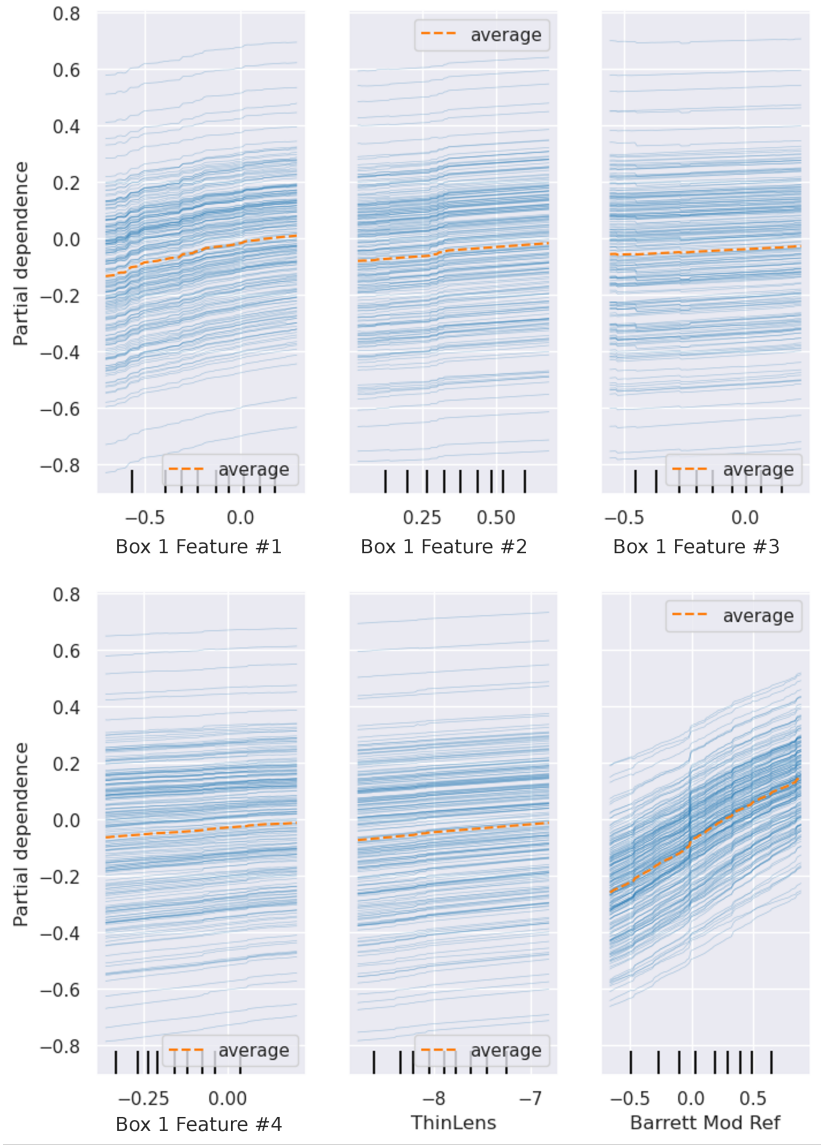

Figure 6.21: Partial dependence and ICE plots for IOL-formula based input features of the FH5600AS lens dataset. Empirical constants are being used here.

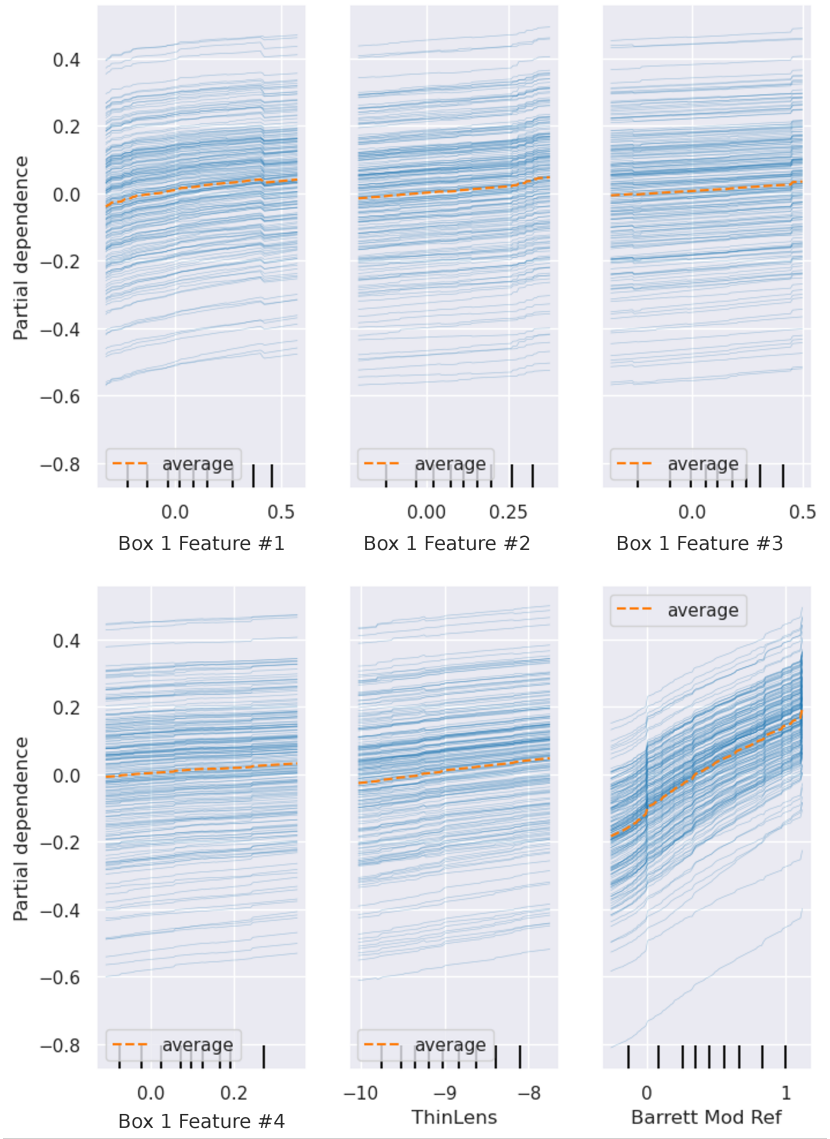

Figure 6.22: Partial dependence and ICE plots for IOL-formula based input features of the HP760AP* lens dataset. Empirical constants are being used here.

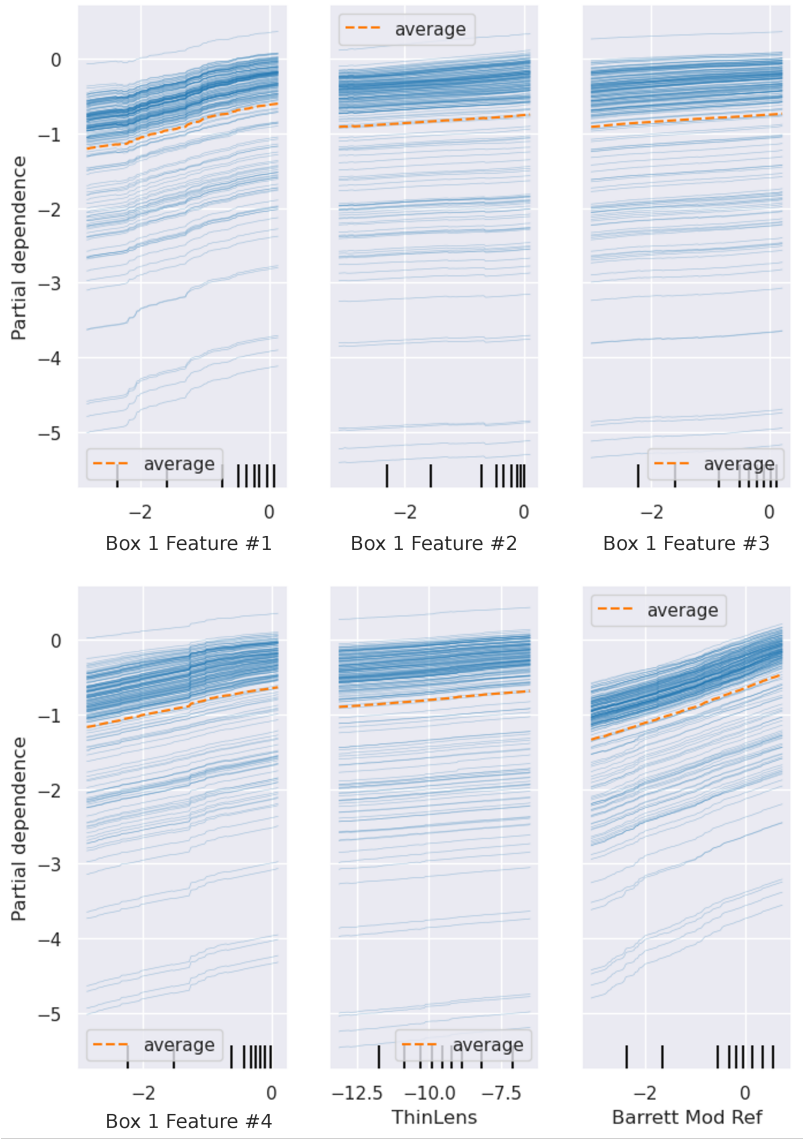

Figure 6.23: Partial dependence and ICE plots for IOL-formula based input features of the UMich (SN60WF at University of Michigan) lens dataset. Empirical constants are being used here.

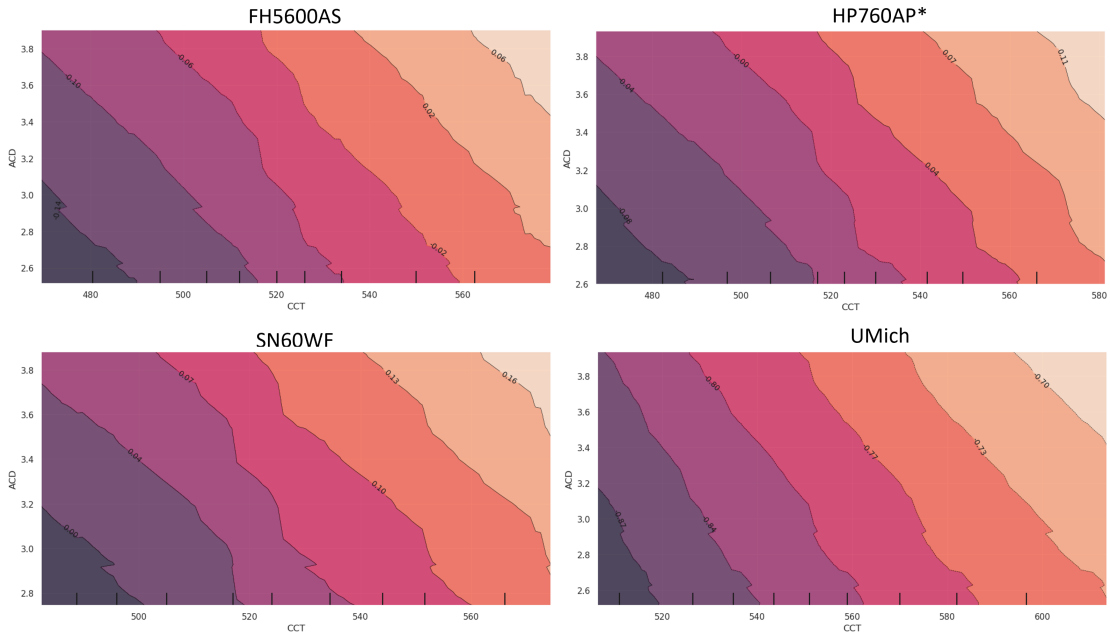

Figure 6.24: Partial dependence plots for interactions between ACD and CCT in each dataset. Empirical constants are being used here.

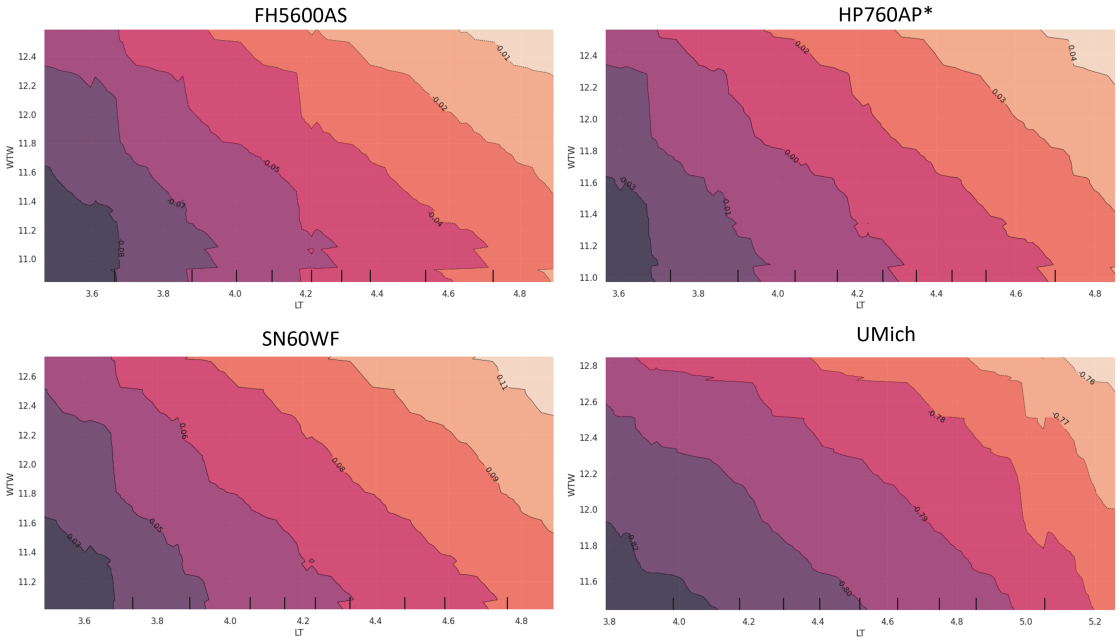

Figure 6.25: Partial dependence plots for interactions between LT and WTW in each dataset. Empirical constants are being used here.

6.7.2 Supplementary tables

Model	Equiconvex	RI	MAE	ME
Nallasamy-G	**True**	**1.46**	**0.2901**	**0.064**
Nallasamy-G	True	1.55	0.3023	0.1349
Nallasamy-G	False	1.46	0.3774	0.3002
Nallasamy-G	False	1.55	0.4143	0.3559
Nallasamy	N/A	N/A	0.3748	0.2873

Table 6.6: Performance of our method on the FH5600AS lens with different input parameters for Equiconvex lens and Refractive Index. Note that the bolded row is the true input for the lens, since FH5600AS is an equiconvex lens with a refractive index of 1.46. Note also the improvement that our generalized model makes in equiconvex lenses compared with the Nallasamy formula.

Feature Name	β	p-value	R^2	ρ
barrettModRefraction	0.727	*	0.7115	0.8435
SRKII	0.6741	*	0.5724	0.7566
SRKRefraction	0.9351	*	0.4568	0.6759
ThinLens	0.296	*	0.3221	0.5676
defaultBarrettOrigIOL	0.0656	*	0.0696	0.2638
defaultSRK	0.0729	*	0.0609	0.2467
A1A2	0.7259	*	0.0425	0.2061
Preop_RCP	0.2174	*	0.0395	0.1988
Orig_CornealPower	0.1552	*	0.0192	0.1386
NaeserpLoc	-0.5923	0.0057	0.0076	-0.0872
Orig_Curv	-0.1377	*	0.0191	-0.1383
OlsenpACD	-0.8704	*	0.032	-0.1788
Preop_RG	-0.3248	*	0.0324	-0.1801
SRK_LCOR	-0.1822	*	0.0365	-0.191
Preop_P	-1.3181	*	0.0379	-0.1948
Haigis_ACD	-0.5311	*	0.0385	-0.1961
Barrett_pACD	-0.4599	*	0.0429	-0.2071
HofferACD	-0.4965	*	0.0437	-0.2091
BinkhorstACD	-1.0145	*	0.0446	-0.2112
SRK_C2	-0.3894	*	0.049	-0.2213
SRK_pACD	-0.4379	*	0.0508	-0.2255
SRK_H	-0.4379	*	0.0508	-0.2255
Preop_Foc	-0.1697	*	0.0546	-0.2337

Table 6.7: Univariate linear regression analyses of each novel feature in our Nallasamy-G generalized formula as a predictor of post-operative refraction. The *beta* values refer to the slope of the predictor. Pearson correlation coefficients were also calculated and displayed as ρ. Values marked with * were less than 0.0001. Only features with a significant p-value were included in this table. Note that features not included in the table are not necessarily unimportant features, but rather do not demonstrate linear correlations with post-operative refraction.

A-Constant	Dataset	Nall-G	Nall.	Haigis	Hoff.	Holl.	SRK/T
Manufacturer	FH5600AS	**0.2649**	**0.258**	0.189	0.1865	0.1807	0.1564
	HP760AP	**0.2765**	**0.331**	0.2293	0.183	0.171	0.0837
	SN60WF	**0.355**	**0.3459**	0.1388	0.1657	0.2057	0.1788
	UMich	**0.8959**	**0.8882**	0.8377	0.8522	0.8582	0.8556
ULIB	FH5600AS	**0.2607**	**0.2695**	0.2153	0.1841	0.1807	0.1564
	SN60WF	**0.3573**	**0.3759**	0.1859	0.1689	0.2114	0.1849
	UMich	**0.8967**	**0.8933**	0.8703	0.8475	0.8568	0.8556

Table 6.8: Correlations of IOL formula predictions to ground truth post-operative refractions under different A-constants and different datasets. The column "A-C" refers to the A-constant used in the analysis: A) Manufacturer, B) ULIB. Note that in all scenarios, both our formula and the Nallasamy formula demonstrate the strongest correlations with post-operative refrection. However, as shown in Table UUU, our Formula also demonstrates lower MAEs and absolute MEs when only a manufacturer A-constant or ULIB constants are known.

Dataset	Formula	MAE	ME	MedAE	STD	AE< 0.5	m	FPI
A	**Nall-G**	**0.2891**	**0.0617**	**0.2255**	**0.4308**	**0.8522**	**0.1333**	**0.3162**
	Nall.	0.4319	0.3735	0.3722	0.436	0.6593	0.1051	0.2962
	Haigis	0.3427	0.0803	0.2655	0.49	0.7841	-0.0835	0.349
	HofferQ	0.3157	-0.0097	0.2355	0.4669	0.8186	-0.2601	0.221
	Holl.	0.284	0.0191	0.2037	0.4396	0.8513	-0.0562	0.4203
	SRK/T	0.297	-0.055	0.233	0.4378	0.8575	-0.0634	0.4048
C	**Nall-G**	**0.249**	**-0.0115**	**0.21**	**0.3177**	**0.8853**	**-0.0873**	**0.3952**
	Nall.	0.2933	0.1773	0.2521	0.3177	0.8416	-0.1687	0.2903
	Haigis	0.3133	0.0557	0.265	0.3946	0.7898	-0.4983	0.1447
	HofferQ	0.3069	-0.0748	0.2518	0.3846	0.8193	-0.67	0.1169
	Holl.	0.2655	-0.0022	0.2225	0.3406	0.8599	-0.5461	0.1391
	SRK/T	0.2637	-0.0148	0.2058	0.3407	0.8569	-0.62	0.1264
D	**Nall-G**	**0.3122**	**-0.0558**	**0.247**	**0.4135**	**0.7996**	**-0.0115**	**0.4935**
	Nall.	0.316	-0.0477	0.247	0.4199	0.8096	0.0666	0.3893
	Haigis	0.3637	-0.0624	0.2927	0.4673	0.7547	-0.179	0.2581
	HofferQ	0.4045	-0.1193	0.3357	0.507	0.6909	-0.875	0.0906
	Holl.	0.3706	-0.0459	0.2974	0.485	0.7378	-0.7487	0.1039
	SRK/T	0.3788	-0.0791	0.3078	0.4842	0.7238	-0.4123	0.1588

Table 6.9: Performance of IOL formulas under ULIB Constants for the following datasets: A) FH5600AS, C) SN60WF (at Aravind), D) UMich. The "DS" column refers to the dataset analyzed: A) FH5600AS, B) SN60WF (SN60WF at Aravind Eye Hospital), C) UMich (patients implanted with SN60WF). Note that no ULIB constants were present for the HP760AP* lens.

CHAPTER 7

Conclusion

7.1 Summary of Findings

Real-world clinical practice is intrinsically multimodal and multidomain. In this dissertation, we have explored multimodal and multidomain machine learning solutions for clinical decision support in the context of limited data, providing three case study examples. In this final chapter, we synthesize the key findings and contributions of this thesis, providing a comprehensive overview of the research journey undertaken. We reflect on the main objectives set forth at the outset and examine how they have been addressed through the course of this study.

In our first dive, we tackled the puzzle of differentiating between pseudoprogression and true progression in glioblastoma, leveraging a multimodal imaging approach with MRI. Uniquely, our study was challenged with the limitations of small sample size (less than 50 patients), resulting in our choice of a low-parameter statistical approach that looked for differences in residual densities based on one modality's prediction of another. The approach highlighted a strategy for multimodal image analysis in the case of limited sample size when no pretrained deep learning models are available for use. The study, although confirming that classical MRI sequences are generally not very helpful in distinguishing psuedoprogression and true progression, found potential discriminatory ability using the ADC modality and leveraging relationships between T1-post and T2.

Moving on to our second case, we explored how routine clinical data can gain a boost in disease prediction capabilities through the inclusion of less common tests from other modalities that can hold the key to better insights. Critically, our clinically-informed model accounted for limited availability of these less common tests in routine practice by opting to leverage them in the training stage only and not require them in testing. By embracing this concept, dubbed "privileged information," through our use of the Random Forest (RF)+ model, we built a model designed to detect temporomandibular joint osteoarthritis using a

routine questionnaire but leveraging CBCT scans and protein serum and saliva tests collected in research. In addition, the model was able to provide some clinical clarity as to what features were most important in determining TMJ OA. The RF+ model was our attempt at crafting a flexible framework that makes the most out of whatever information is available.

Finally, we tackled the challenge of predicting postoperative refraction after cataract surgery, despite the scarcity of data spanning different lens types. Due to a lack of diverse training data, we were constrained to an approach where our training set consisted of patients from one institution implanted with one lens model. Thus, we attempted an informed domain generalization approach that could leverage aspects of eye measurements and known information about lens build to overcome differences caused by different lens manufacturers. We started with a model trained on patients with one type of lens (the Nallasamy Formula) and fine-tuned it to work across various lens types. The overarching objective was to adapt this model to furnish precise and robust prognostications across various lens types. Innovative feature engineering techniques were introduced as a pivotal strategy aimed at mitigating error propagation across divergent domains. As a result of our changes, we were able to show consistently lower and more robust MAEs across different populations and lens types from different manufacturers under different A-constant scenarios through our model, and were able to attribute this to specific engineered features in the model. As a result, we concluded that adding information about lens geometry was a critical piece to overcoming the limitations of a non-diverse dataset when improving the Nallasamy Formula.

Thus, our work provides three examples of informed ML models built for proof-of-concept CDS which overcome the challenges of limited data. We believe approaches such as these are crucial to inclusivity among practices of all sizes, as these models are designed for limited data but can be scaled upwards to larger datasets.

The methods demonstrated in this work are assigned to three different biomedical problems, which provides a well-rounded scope of the work. Our work demonstrates approaches for imaging data as well as tabular data, situations where diverse data is available *in training* (privileged learning), or situations where diverse data is not available in training at all, leading to other approaches that need to be considered (domain generalization). We hope this strength of the work allows a reader to consider how any of these methods could be applied or adapted for new CDS applications.

7.2 Future Directions

7.2.1 Multimodal Fusion MRI for psuedoprogression detection

Our strategy for developing a multimodal low-parameter framework for image processing with small sample sizes uncovered further evidence that the ADC modality carries predictive value in discriminating true progression from pseudoprogression. However, it also un–discovered the classical MRI sequences are not good discriminators of these conditions, which has also been supported by the literature [225, 130]. Based on our results, there is a potential for the relationship between T1 post-contrast and T2 imaging to contain some discriminatory ability which is weaker than the ability of ADC but notable.

When applying a GWR-based framework to distinguish cases of 1p/19q co-deletion, [123] leverage unique and visibly-distinct visual patterns in MRI known as T2-FLAIR mismatch, where a hyperintense ring appears in the FLAIR modality. A limitation of the GWR-approach is that there must be some manifestation of unique pixel intensity patterns that are visible in one class but not another. Therefore, a realistic future approach using GWR should focus on automating arduous tasks for radiologists that are obvious but monotonous. This could, for example, be applied in class distinction scenarios beyond the MRI modalities, using histopathological tissue instead. One example would be an in-house discriminator of oligodendrogliomas vs astrocytomas in histopathological tissue. Since oligodendrogliomas contain a "fried-egg" appearance in tissue samples, when converted to grayscale, they should contain a higher proportion of white pixels compared to astrocytoma images [196]. However, preservation methods should be consistent and care should be given in images to prevent excessive tears in the tissue, which could confuse the model.

In the case of discrimination of pseudoprogression and true progression, our study demonstrated the largest study to assess the discriminative ability of ADC, but further interpretive work should be done to elucidate what aspects of the tumor image results in ADC demonstrating decent predictive ability. Understanding this may allow others in future work to add constrast-enhancements or other imaging procedures which enable these patterns to emerge more salient for machine learning models and thus improve the discriminative ability of ADC.

7.2.2 Privileged Learning for Temporomandibular Joint Osteoarthritis Prediction

TMJ OA appears to be best diagnosed by analyzing loss of bone density in the joint area and a shrinkage of joint distance in CBCT scans, but most patients suspected of having

TMJ OA will not undergo such scans. Additionally, protein serum and saliva tests may also be of potential help for diagnosis, but this is also not commonly conducted in the clinic. Therefore, our study attempted to leverage this information as "privileged" information in a model based primarily on clinical questionnaires asked of patients suspected of having TMJ OA. Our model demonstrated marginal gains over a baseline model containing only questionnaire items.

We suggest three future directions for this work. Firstly, we note that a limitation of this study was that the clinical questionnaire was relatively short, resulting in a small number of non-privileged features. However, other routinely-collected variables not included in the questionnaire could also be useful for the model. For example, patients exhibiting TMJ OA may also demonstrate other routinely-collected clinical indicators such as a family history of osteoarthritis or current comorbidities that would land them more adept at exhibiting TMJ OA. Additionally, information about number of teeth cleanings per year or oral exam results may give an overall indication of the health of gums and the underlying bone, providing potentially useful information towards onset of TMJ OA. Therefore, we suggest an inclusion of additional baseline features from data routinely collected which extend beyond the clinical questionnaire.

Second, we suggest experiments to better understand the limitations of the RF+ model. We hypothesize that when baseline features are relatively stronger than privileged features, privileged features will have little to no effect. Therefore, an understanding of how strong privileged information must be in relation to baseline non-privileged features to contribute to the model will be useful.

Another hypothesis is that frameworks where the non-privileged features and privileged features have stronger similarities, the privileged model can provide greater gains. In [128], the authors developed a model based on lower-quality CT as non-privileged features and higher-quality MRI as privileged features. In this framework, CT and MRI have strong structural similarities but differing levels of detail. Because of the strong similarities between modalities, CT can better build scandent trees that mimic splits at link nodes made with privileged features. Therefore, TMJ OA assessment could also focus on routinely-collected X-rays of the teeth or other similar imagery as input features, with the assumption that connections between bone density in the mandible may demonstrate similar patterns to that of CBCT scans in the jaw, enabling better learning of privileged features.

7.2.3 Post-operative Refraction Prediction

Our generalized model for post-operative refraction prediction demonstrated significant improvement over the Nallasamy formula to generalize to other population datasets under different A-constant sets for different lenses. This is in thanks to an expanded set of features that allow the model to learn factors affecting refraction prediction that extend beyond A-constant-based features.

In [107], Li et al propose two metrics called Mean Absolute Error in Prediction of Intraocular Lens (MAEPI) and Correct IOL Rate (CIR) to assess the ability of a model which predicts *IOL power*. While the Nallasamy formula is optimized to predict post-operative refraction, it is used in practice as a predictor of IOL power. However, we demonstrate in an analysis in Appendix I that our generalized formula, although an excellent predictor of post-operative refraction, is not well-fit for the reverse problem of predicting IOL power. In essence, it is likely the very presence of the features which allow generalizability of the model also cause the model to function poorly as a reverse predictor. From the analysis, it appears that the model cannot make proper assumptions about predicting data where post-operative refractions are greater than 0, likely because of a relative lack of data in the training set for low IOL powers, which are more likely to correspond to positive refractions. While the inclusion of data containing low IOL powers may seem of lower importance considering their rarity, the significance of the results in I may point more to a necessity for understanding relationships between refraction and lower IOL powers to build a truly robust model.

Since there is clinical value to developing a model which predicts IOL power, there are a few suggestions about how to proceed to build a model which better predicts IOL power which may still retain generalizability. Since we hypothesize that the *number* of features are causing the model to better understand the distribution of the University of Michigan data, we suggest the following adaptations.

The first suggestion is to add more data for less common cases, such as longer axial lengths and lower IOL powers. As alluded to in the analysis in Appendix I, the greatest contributor to a large MAEPI is errors in due to a poor understanding of relationships between refraction and lower IOL powers. This resulted in associations between lower IOL powers and more positive refractions that were not well-defined enough to produce a good reverse model. Providing more example of lower IOL powers and higher refractions will help the model better distinguish these. The addition of more axial length data will also provide a better understanding of how the model can adjust to especially long or short eyes, which was demonstrated in Figure 5.6 to increase prediction error for all IOL formulas.

Another suggestion is to retrain the model to weigh predictions for lower IOL powers higher. Predictions including lower IOL power inputs appear to be the most adversely

affected by the model. Alternatively, where post-operative refraction predictions are greater than zero, loss function weights can be higher, to encourage greater distinction between IOL power and refraction there.

Thirdly, we suggest reducing the number of features. Since both our generalized model and the Nallasamy formula consist of some level-1 tree-based models, the node features were able to better separate the impact of high refractions and lower IOL powers to minimize the effect of refraction predictions for lower IOLs and better focus on refraction prediction for higher IOLs. The additional variables allowed the tree-based methods to separate the values better in our generalized model than in the Nallasamy formula. Therefore, we hypothesize that reducing the number of variables can ameliorate this problem partially. However, a reduction of variables may also result in a loss of generalizability of the model. Therefore, future work must determine how to preserve as much generalizability as possible while also providing few enough features to prevent this form of learning to lower importance of lower IOL powers.

Lastly, we suggest training a model specifically optimized to IOL power prediction. If the target clinical goal is IOL power prediction, this is an important step to ensure that IOL power prediction is optimized over anything else. It cannot be guaranteed that a model accurately mapping $X \longrightarrow Y$ will also accurately map $Y \longleftarrow X$. Therefore, if bi-directionality of the model is the goal, or if predicting IOL power from target refraction is the goal, then the model should include provisions for optimizing IOL power.

Moving beyond our generalized model, the future of refraction prediction may also involve probabilistic modeling. This may result in an output of a distribution of possible refractions for a given patient, together with a probability for each refraction. Probabilistic approaches may be especially useful for surgeons juggling multiple procedures in the eye at once which may increase variability beyond standard single-procedure cases.

7.3 Conclusion

This dissertation focuses on developing proof-of-concept machine learning-based clinical decision support models tailored to the complexities of multimodal and multidomain clinical scenarios, driven by the recognition that real-world clinical work is diverse in nature. While these three works demonstrated application to specific medical case studies, these methods exhibit versatility and can be applied to various scenarios requiring privileged learning, multimodal image fusion, or domain adaptation. A key focus of this dissertation has been to address multimodal and multidomain challenges in cases of limited data, but we stress that the methods presented here can also be scaled up to any number of samples. This

wide-range of scalability for clinical decision models is crucial, mirroring real-world clinical and research data scenarios where massive datasets are often unattainable or require years to build up. Consequently, this work underscores the potential of machine learning-based clinical decision support models, showcasing their adaptability and laying the groundwork for future advancements in the field.

APPENDIX A

Brief History of Key Developments in AI Due to Image Processing

A.1 Background

Multimodal machine learning for medical applications has evolved primarily from two phenomena that occurred around the same time: 1) the growing presence of high-quality, digitally-available patient health data, and 2) rapid developments in machine learning and associated software. Prior to these events, multimodal models in healthcare were possible but uncommon, due to challenges in obtaining large amounts of patient data from local sources and finding models to process them. To address the fusion challenge, for example, data from multiple modalities were often converted to vector form via feature engineering and then simply concatenated into a single traditional ML model [226, 72, 219]. Other strategies could include usage of SVM kernels [213, 226, 49], or ensemble models [133, 135]. Other multimodal challenges such as co-learning were not often easy to address, although Vapnik et al's SVM+ method of information transfer via privileged information is a notable strategy [200]. Thus, an explosion of publicly-available multimodal datasets and the development of advanced ML models, particularly neural network-based, have rapidly developed the field. Here we discuss the history of big data and the modern architectures enabled by the former. Four models discussed in this section are illustrated in Figure A.1.

A.1.1 Growing Presence of Big Data

In 2011, the National Cancer Institute (NCI) paired with Washington University to initiate the The Cancer Imaging Archive (TCIA), collecting imaging and patient health data (PHI) related to cancer diagnosis from 31 separate collection sites [40]. The data was to be made publicly available for the purposes of cancer research. TCIA and its sister database The Cancer Genome Atlas (TCGA) are some of the most commonly used databases for

medical imaging and Patient Health Information (PHI) today, exemplifying the impact of such repositories for biomedical and translational research. Today, these large annotated databases, which include newer arrivals such as Stanford's CHeXpert [81], University of Pennsylvania's BraTs [121, 13, 14], and PhysioNet's MIMIC-III [87, 88], among others, can provide thousands of data samples of varying modalities for multimodal data models. Such repositories of medical information have also likely benefited from government interventions such as the Health Information Technology for Economic and Clinical Health Act (HITECH) act, designed to encourage clinics to move health records to a digital medium [69].

A.1.2 Significance of ImageNet

Rapid developments in machine learning and computer vision also occurred at this time. While the early-2000s saw a practical rise in applications of what we call here "traditional machine learning methods" such as SVM and RF, the advent of the ImageNet challenge and subsequent win of AlexNet in 2012 [94] spurred a new interest in neural networks, which gradually overtook the computer vision community. Prior to AlexNet, the state-of-the-art for image recognition consisted of extraction of specific engineered features, typically through some method such as SIFT or Bag-of-Words (BoW)s, which resulted in feature vectors, followed by processing through a traditional ML algorithm. By surprising contrast, AlexNet completely bypassed the feature extraction step, utilizing a simple 5-layer CNN to automate cross-correlation filters for feature extraction. Although a relatively simple architecture by today's standards, AlexNet was a revolutionary step for the time. The winning ImageNet model's top 5 error rate dropped from approximately 26% with feature engineering in 2011 to a little over 15% with CNNs in 2012 [165]. All following winning models in the ImageNet challenge were based on convolutional neural nets [165].

The ImageNet challenge and the successful implementation of CNNs spurred additional advances in computer vision in the next few years. VGG16 and VGG19, some of the first very deep ($>$ 10 layers) CNN-based architectures proposed from Simonyan and Zisserman [176], won top prizes in the 2014 ImageNet challenge and revolutionized CNN applications to data by presenting a form of standardized architecture that was organized in block-like structure of two to three convolutional layers followed by a max pooling layer. The ability to organize and justify the block-like structure of the network helped make CNNs more accessible to a broader audience, who were largely freed from the arduous task of layer construction and hyperparameter tuning for every kernel and channel size of each layer.

As the overall architectures for VGG16 and VGG19 became more widespread, this also enabled a broad new application of **transfer learning** using the ImageNet networks ap-

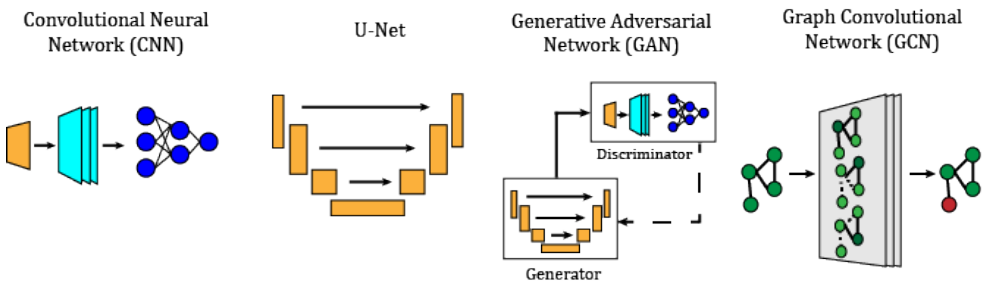

Figure A.1: Common deep learning models used for medical imaging and clinical decision support.

plied to in-house data [198]. Transfer learning allowed third parties to borrow a pre-trained VGG16/19 model with frozen weights in the convolutional layers, then retrain only the final layers of the network for their own classification tasks. Not only did transfer learning reduce compute time, but it also made deep learning accessible to third parties with only a limited number of data samples. Prior to this, deep learning, often containing thousands (and now millions) of trainable parameters, were only useful on large datasets such as ImageNet, which contained millions of images. Application of neural networks on most in-house datasets with less than a few thousand samples, such as those coming from a single hospital site, did not often confer better results than traditional ML methods and could be subject to extreme overfitting. Therefore, deep learning was at the time still constricted to use by computer scientists who specialized in architecture rather than the broader data science field and bioinformatics. However, transfer learning opened doors to these latter groups, allowing parties with comparatively smaller datasets to access deep learning and produce high-accuracy models. Conceptually, transfer learning worked well with ImageNet models because ImageNet contained 1000 categories of images [45], broad enough in theory for trained cross-correlation filters to selectively seek out generic visual characteristics from images, much like the human eye is believed to function. Such generic characteristics are theorized to be adaptable to any kind of image, where the model is trained to associate certain groups of these characteristics with new classes of objects. This conceptualization has resulted in years of successful medical models [181, 198, 177, 182, 58, 80, 37, 92], but has recently been challenged by proponents of self-supervised learning [221, 119, 8].

The ImageNet challenge engendered another radical shift in computer vision which is adopted in today's medical imaging models. At the time of introduction, the 16- and 19-layer VGG16/19 models were some of the deepest networks available. Deeper networks with more parameters imply more degrees of freedom and thus more complex models, which benefits modalities from medical images to natural language processing datasets. However,

networks deeper than the VGG architectures proved hard to conceive because training error begins to increase with more layers [70]. In 2015, Microsoft's 154-layer ResNet [70] model won first prize in the ImageNet competition, introducing **residual networks**. Residual networks, which consist of a simple architectural change whereby networks are organized into residual "blocks", were revolutionary in their own right because they reduced the computational burden required to process deep networks by allowing networks to easily learn identity parameters. What was previously mistaken as a "vanishing gradient problem" turned out to be easily-solved convergence issues. The authors even claimed to have built a 2000-layer network, the likes of which were previously unfathomable.

A.1.3 U-Nets and GANs

The ImageNet models had their limitations with medical imaging, however. While ImageNet models were specifically designed to solve classification problems, many classic clinical imaging challenges involved segmentation of specific organs or tumor tissue. For example, decision support models to detect tuberculosis or other lung-based conditions would require first a segmentation of the lung on a chest x-ray before extraction of features. Likewise, overall survival prediction of glioma patients using MRI would first require a segmentation of the tumor area before model development. Traditionally, computer vision algorithms would address such segmentation problems via arduously tailored trial-and-error computer vision toolbox methods such as identifying edges and convex hulls, opening and closing spaces, and applying active contours and fills. The introduction of new data, even of the same image type, may have meant additional tweaks to minimize obvious errors. While effective, the process was by no means automated, requiring time-intensive labor to customize to the available data. Therefore, the innovation of the CNN-based U-Net [163], specifically designed for medical imaging data, was a revolutionary advancement in image segmentation.

One of the critical advantages of the U-Net to this day is that the overall structure of the network is generally effective. The U-Net consists of a series of downsampled layers followed by a series of upsampled layers with concatenated skip connections, or copies of outputs from previous layers in the downsampled series. The authors of the U-Net argued that this structure is effective for image segmentation because it captures both pixel-level features and group-level context [163]. As of the date of publication, improvements have been suggested to improve U-Net architecture, but the overall downsampling, upsampling, and skip connection concepts which characterize U-Net remain state-of-the-art in image segmentation.

Interestingly, while the U-Net was originally intended for image segmentation, it turned out to also be useful for image generation. In 2014, Goodfellow et al [60] proposed another

major advancement critical to today's multimodal ML work called GANs. GANs function utilizing two networks that train against each other, each trying to outperform the other. One of these networks, called the generator, often consists of a general U-Net-style architecture, typically taking as input some kind of random noise or an image and outputting a fake generated image. The generator is challenged by the second module, called the discriminator, which is typically a CNN that takes as input the generated image and outputs a single binary classification that describes whether the generated image is real or fake. The discriminator is trained with real images as ground truth "real" images and previously generated output as ground truth "fake" images. The goal of the GAN is for the generator to understand the underlying distribution of the "real" images, and construct new "fake" images which appear to be real.

Modifications such as CycleGAN [234] and conditional GANs [82] have enabled improvements to computer vision tasks such as style transfer [103, 89] and super-resolution images [99, 64, 89]. Due to this, GANs have become essential to the growth of work on multimodal translation in recent years in fields as diverse as natural language processing [66, 109, 149], anomaly detection [104, 47] and biomedical applications [118, 191, 113, 151, 162, 54]. In the latter category, image-to-image translation, whereby both input and output are images from different medical imaging domains (e.g. input CT, output MRI), has been the most popular application of GANs, and the trend continues today.

A.1.4 On the Rise: Graph Convolutional Network (GCN)s

In 2016, Kipf and Welling [93] introduced the concept of GCNs. A simple graph structure requires a node and an edge, which indicates which nodes are connected with each other. Edges are often given weights to represent the strength of node connections, and nodes (edges) can also contain additional information as well, called node (edge) features. GCNs provide the opportunity to classify either nodes or entire networks as well as to predict edge weights, and have been utilized in fields such as social networks and recommendation systems [215, 126, 150, 224] as well as image captioning [223]. They are of particular interest in multimodal ML because multiple networks representing separate modalities can be combined together, as demonstrated in Zitnik et al's [237] work on polypharmacy side effects. Graph convolutional networks have recently been applied to other areas of multimodal biomedical research such as for survival outcome prediction in glioma [34] and modeling interactions for drug repurposing [208].

A.1.5 Summary

conclusion, the evolution of multimodal machine learning in medical applications has been significantly propelled by two key factors: the emergence of extensive, digitally-accessible patient health data and rapid advancements in machine learning techniques. Previously, challenges in acquiring and processing diverse data modalities hindered the widespread adoption of multimodal models in healthcare. However, the landscape changed with the availability of large, publicly-accessible datasets like TCIA, TCGA, and others, facilitated by initiatives like the HITECH act. Simultaneously, breakthroughs in machine learning, notably with the advent of deep neural networks spurred by the ImageNet challenge, revolutionized computer vision and paved the way for applications in medical imaging. Architectures like VGG16/19 and ResNet, along with innovations like U-Nets and GANs, introduced efficient methods for tasks like segmentation and image generation, fundamentally altering the approach to medical image analysis. Moreover, the introduction of GCNs extended the applicability of multimodal approaches by enabling effective fusion of data from disparate sources, showcasing their potential in diverse biomedical domains. Together, these advancements have propelled multimodal machine learning to the forefront of medical research, promising groundbreaking insights and transformative applications in healthcare.

APPENDIX B

Density-Based Classification in Diabetic Retinopathy through Thickness of Retinal Layers from Optical Coherence Tomography

B.1 Summary

Diabetic retinopathy (DR) is a severe retinal disorder that can lead to vision loss, however, its underlying mechanism has not been fully understood. Previous studies have taken advantage of Optical Coherence Tomography (OCT) and shown that the thickness of individual retinal layers are affected in patients with DR. However, most studies analyzed the thickness by calculating summary statistics from retinal thickness maps of the macula region. This study aims to apply a density function-based statistical framework to the thickness data obtained through OCT, and to compare the predictive power of various retinal layers to assess the severity of DR. We used a prototype data set of 107 subjects which are comprised of 38 non-proliferative DR (NPDR), 28 without DR (NoDR), and 41 controls. Based on the thickness profiles, we constructed novel features which capture the variation in the distribution of the pixel-wise retinal layer thicknesses from OCT. We quantified the predictive power of each of the retinal layers to distinguish between all three pairwise comparisons of the severity in DR (NoDR vs NPDR, controls vs NPDR, and controls vs NoDR). When applied to this preliminary DR data set, our density-based method demonstrated better predictive results compared with simple summary statistics. Furthermore, our results indicate considerable differences in retinal layer structuring based on the severity of DR. We found that: (a) the outer plexiform layer is the most discriminative layer for classifying NoDR vs NPDR; (b) the outer plexiform, inner nuclear and ganglion cell layers are the strongest biomarkers for discriminating controls from NPDR; and (c) the inner nuclear layer distinguishes best between controls and NoDR.

Although this case is not a multimodal or multidomain problem, we include this work in

the thesis as a reference to a unimodal/unidomain strategy for creating a clinical decision support concept model with limited data. The sample size of 107 subjects was too small to perform deep-learning methods, so a density-based strategy, similar to that demonstrated in Chapter 02 and Appendix B was applied.

B.2 Publication and Acknowledgment

This appendix is a published work [125]: S. Mohammed, T. Li, X. D. Chen, E. Warner, A. Shankar, M. F. Abalem, T. Jayasundera, T. W. Gardner, and A. Rao, "Density-based classification in diabetic retinopathy through thickness of retinal layers from optical coherence tomography.," Scientific reports, vol. 10, no. 1, p. 15937, 2020

APPENDIX C

Quantifying T2-FLAIR Mismatch Using Geographically Weighted Regression and Predicting Molecular Status in Lower-Grade Gliomas

C.1 Summary

The T2-FLAIR mismatch sign is a validated imaging sign of isocitrate dehydrogenase-mutant 1p/19q noncodeleted gliomas. It is identified by radiologists through visual inspection of preoperative MR imaging scans and has been shown to identify isocitrate dehydrogenase-mutant 1p/19q noncodeleted gliomas with a high positive predictive value. We have developed an approach to quantify the T2-FLAIR mismatch signature and use it to predict the molecular status of lower-grade gliomas. We used multiparametric MR imaging scans and segmentation labels of 108 preoperative lower-grade glioma tumors from The Cancer Imaging Archive. Clinical information and T2-FLAIR mismatch sign labels were obtained from supplementary material of relevant publications. We adopted an objective analytic approach to estimate this sign through a geographically weighted regression and used the residuals for each case to construct a probability density function (serving as a residual signature). These functions were then analyzed using an appropriate statistical framework. We observed statistically significant (P value = .05) differences between the averages of residual signatures for an isocitrate dehydrogenase-mutant 1p/19q noncodeleted class of tumors versus other categories. Our classifier predicts these cases with area under the curve of 0.98 and high specificity and sensitivity. It also predicts the T2-FLAIR mismatch sign within these cases with an under the curve of 0.93. On the basis of this retrospective study, we show that geographically weighted regression-based residual signatures are highly informative of the T2-FLAIR mismatch sign and can identify isocitrate dehydrogenase-mutation and 1p/19q codeletion status with high

predictive power. The utility of the proposed quantification of the T2-FLAIR mismatch sign can be potentially validated through a prospective multi-institutional study.

This work is the predecessor the Chapter 02 and lays the groundwork the interest in geographically-weighted regression. However,the application of this work involves assessment of images where a clear visual anomaly (a hyperspectral ring around the effective tumor area) occured in the FLAIR modality. This anomaly was apparent to the naked eye, and therefore the purpose of the study application was to save physicians time in identifying the T2-FLAIR anomaly. In the case of pseudoprogression, differences between pseudoprogression and true progression were not outwardly apparent to the naked eye and thus there was widespread interest in applying machine learning techniques to observe latent patterns in the images. Therefore, our work in Chapter 4 demonstrates an interesting and novel application of this method to discrimination of conditions in MRI that are not outwardly apparent to the naked eye.

C.2 Publication and Acknowledgment

This appendix is a published work [124]. S. Mohammed, V. Ravikumar, E. Warner, S. Patel, S. Bakas, A. Rao, and R. Jain, "Quantifying t2-flair mismatch using geographically weighted regression and predicting molecular status in lower-grade gliomas," American Journal of Neuroradiology, vol. 43, no. 1, pp. 33–39, 2022.

APPENDIX D

Feature Equations in Nallasamy Formula

The following section describes selected equations included in the Nallasamy formula. For information on the equations, refer to the reference given for each equation.

Abbreviations:

1. OD : predicted lens Dioptre (D)
2. n_1 : refractive index of acqueous
3. n_2 : refractive index of IOL
4. DL_d : a predicted lens Dioptre (D) in a single iteration
5. FL_1 : power of anterior surface lens implant (D)
6. FL_2 : power of posterior surface lens impalnt (D)
7. t : represents a thickness of both the lens and the capsular bag as a unit, usually represented as the lens thickness.
8. RA : anterior radius of IOL
9. RP : posterior radius of IOL
10. RG : posterior segment of the globe
11. RCP : peripheral radius of cornea
12. RC : central radius of cornea
13. RCC : central posterior radius of cornea
14. PZ : P-factor of cornea

15. Ac : SRK/T A-constant

16. $pACD$: posteroperative ACD

17. ELP : effective lens position

The features used in the generalized Nallasamy formula (Nallasamy II) are as follows:

1. `Km` : average keratometry
$$K_m = (K_1 + K_2)/2$$

2. `AD` : axial diameter
$$(ACD - CCT)/1000$$

3. `AST` : measure of astigmatism
$$\text{abs}(K_1 - K_2)$$

4. `defaultBarrettOrigIOL` was derived from the following equation at a target refraction of 0 [16]:
$$OD = \frac{n_1 \times 1000}{AL - d - t + \frac{n_1}{n_2} \times \frac{DL_d - FL_2}{(1 - fractn_2 \times FL_2)} \times \frac{t}{DL_d} - \frac{n_1 \times 1000}{\frac{n_1}{K_m} - d - \frac{n_1}{n_2} \times \frac{FL_2}{DL_d} \times t}}$$

5. `defaultThickness` : a measurement of the assumed thickness of the lens based on the following equation from [17]:
$$T = (RA - \sqrt{RA^2 - (OD/2)^2} + (RP - \sqrt{RP^2 - (OD/2)^2}$$

6. `Preop_RCP` : prediction of preoperative radius of peripheral cornea based on equation from [17]:
$$ACD = AL - 0.593_0.13 - RG - \sqrt{RG^2 - RCP^2 + (RCP - ACD)^2)}$$

7. `calcRG` : calculation of radius of globe from Barrett Universal II eye model [17]:
$$RG = 0.35066 \times AL - 0.06607 \times K - 5.70871$$

8. `RCC`: calculation of central posterior radius of cornea according to [17]:
$$RCC = RC \times 0.883$$

9. `calcRC`: calculation of central radius of cornea from [17] is based on a derivative of the following equation:

$$Km - (376/RC) - (40/RC) + (0.00052/1.376) \times (376/RC) \times (40/RCC)$$

10. `Preop_P`: Preoperative "P-value" as described in [17] was assessed based on a derivative of the following equation:

$$RCP = (RC^2 + (1 - P) \times 5^2)^{3/2}/(RC^2)$$

11. `defaultSRK` and `reverseSRK` are based on derivatives of the following SRK formula [160]:

$$OD = Ac - 2.5 \times AL - 0.9 \times K_m - (R/(1/0.0875 \times Ac - 8.55))$$

12. `HofferACD`: Hoffer's postoperative ACD prediction [74]

$$pACD = 0.292 \times AL - 2.93$$

13. `BinkhorstACD`: Binkhorst's prediction for postoperative ACD [17]

$$pACD = 0.17 \times AL + 0.017$$

14. `OlsenpACD`: Olsen's prediction for postoperative ACD [137]

$$pACD = 1.14 \times 0.22 \times ACD + 0.10 \times AL$$

15. `NaeserpLoc`: predicted position of posterior lens capsule [131]:

$$2.40 + 0.011 \times age + 0.171 \times ACD + 0.051 \times AL$$

16. `a1a2`: Developed from a derivative of the ELP equation from Haigis [173]:

$$ELP = a_0 + a_1 \times ACD + a_2 \times AL$$

APPENDIX E

RCP Algorithm

Algorithm 1: Calculate Radius of Peripheral Cornea and post-operative ACD

Input: A,WTW,AL,$K1$,$K2$
Output: RCP, $pACD$
$RG \leftarrow$calcRG(AL,K1,K2);
$RC \leftarrow$calcRC(K1,K2);
$LF \leftarrow A * 0.5825 - 67.6627$;
$\alpha \leftarrow RG - LF$;
for $PZ \in [-2, 2]$ **do**
    $pred\_RCP = (RC^2 + (1 - PZ) * 25)^{3/2}/RC^2$;
    if $pred\_RCP <= RG$ **then**
        $x = (pred\_RCP^2 - RG^2 + \alpha^2)/2\alpha$;
        if $pred\_RCP^2 - x^2 < 0$ **then**
            continue;
        else
            h = $\sqrt{pred\_RCP^2 - x^2}$;
            $pred\_CD = 2h$;
            if $abs(WTW - pred\_CD) < 1e^{-}2$ **then**
                break
            end
        end
    end
end
$RCP = pred\_RCP$;
$pACD = AL - 0.593 + 0.13 - RG - \sqrt{RG^2 - RCP^2 + (RCP - ACD)^2}$;

APPENDIX F

Barrett Model

Original Barrett I algorithm in pseudocode [16].

Algorithm 2: Barrett

Input: $K1, K2, S, AL, ACD = 4.8, N2 = 1.435, R = 25, T = 1, N1 = 1.336$
Output: $P2$
$C \leftarrow (K1 + K2)/2 + s/(1 - 0.012 * S)$;
$F2 \leftarrow 1000 * ((N2 - N1)/R)$;
$P1 \leftarrow 21.5$;
for $x \in [1, 10]$ **do**
    $F1 = (P1 - F2)/(1 - (T/(N2 * 1000)) * F2)$;
    $E2 = (N1/N2) * (F1/P1) * T$;
    $E1 = (N1/N2) * (F2/P1) * T$;
    $L = (N1/C) * 1000$;
    $U = L - ACD - E1$;
    $V = AL - ACD - T + E2$;
    $P2 = (N1/V) * 1000 - (N1/U) * 1000$;
    $P1 = P2$;
end

APPENDIX G

Modified Barrett Formula

Algorithm for Modified Barrett Formula. Input takes in $K1$, $K2$, AL, ACD, LT, IOL and outputs post-operative refraction.

Algorithm 3: Modified Barrett

Input: $IOL, ACD, AL, K1, K2, N2, T, eqc$
Output: R
$N1 \leftarrow 1.3315$;
$OD \leftarrow 6$;
$K \leftarrow (K1 + K2)/2$;

$RA = ((N2 - N1) * 1000)(IOL/(1.5 + (0.5 * \text{int}(eqc)))$;
$PA = ((N2 - N1) * 1000)/RA$;

`/* power of posterior surface */`
$R = ((N2 - N1) * 1000)/(IOL - PA)$;
$F2 = (1000 * ((N2 - N1)/R))$;

`/* projected thickness of IOL */`
if $RA * *2 - (OD/2) * *2 < 0 or R * *2 - (OD/2) * *2 < 0$ **then**
  $IOL\_T = T$;
else
  $IOL\_T = (RA - \text{sqrt}(RA * *2 - (OD/2) * *2)) + (R - \text{sqrt}(R * *2 - (OD/2) * *2))$;

`/* Corrected power of anterior surface w/ thickness */`
$F1 = (IOL - F2)/(1 - (IOL\_T/(N2 * 1000)) * F2)$;

`/* Distance to Principal Planes */`
$E2 = (N1/N2) * (F1/IOL) * IOL_T$;
$E2\_adj = (T/2) - (IOL\_T/(2 + \text{int}(eqc)))$;
$E2 = E2 + E2\_adj$;
$E1 = (N1/N2) * (F2/IOL) * T$;

$V = AL - ACD - T + E2$;
$a = (N1/V) * 1000$;
$b = ACD + E1 - (N1/(IOL - a) * 1000)$;
$c = (N1/b) * 1000$;
$d = c - K$;
$R = d/(1 + 0.012 * d)$;

APPENDIX H

Comparisons of Our Method Using Optimal A-Constants

Analyses conducted in chapter 4 were done using manufacturer A-constants or experimentally-derived A-constants posted online at the User Group for Laser Intereference Biometry [2]. This appendix provides the same assessments conducted with optimized A-constants for each dataset. A-constants were selected by obtaining post-operative refraction predictions from Haigis, Hoffer Q, Holladay 1 and SRK/T models at different A-constants, and selecting the respective A-constants which produce the closest absolute mean error to zero. These can be shown in Table H.1.

Results for optimized performance of Nallasamy-G compared with Nallasamy, Haigis, HofferQ¡ Holladay1, and SRK/T under the optimzed constants from Table H.1 are given in Table H.2. In the FH5600AS at Aravind dataset, Nall-G provided an MAE of 0.2901, outperformaning Nallasamy (0.374), Haigis (0.3376) and HofferQ (0.3155) formulas, but was outperformed by Holladay1 (0.2841) and SRK/T (0.2837). For the HP760AP* dataset, Nallasamy-G outperformed Nallasamy in MAE (0.288 v 0.2982), as well as Haigis (0.3182) and HofferQ (0.312). It was outperformed by Holladay1 (0.285) and SRK/T (0.2873). In the SN60WF at Aravind and UMich datasets, the lens which Nallasamy-G was trainined on, it outperforms all other models in MAE, performing at 0.2495 in the SN60WF (at Aravind) dataset compared with Nallasamy (0.2842), Haigis (0.3049), HofferQ (0.3052), Holladay1 (0.265) and SRK/T (0.2636). In the UMich dataset, Nallasamy-G provides an MAE of 0.3096, outcompeting Nallasamy (0.3125), Haigis (0.3632), HofferQ (0.4038), Holladay1 (0.3706) and SRK/T (0.376). Although ME for the generalized formula on non-SN60WF lenses is higher (FH5600AS: 0.064, HP760AP: -0.0935)) than for the centered formulas Haigis (FH5600AS: 0.0091, HP760AP*: -0.0023), HofferQ (FH5600AS: 0.0179, HP760AP*: 0.0025), Holladay1 (FH5600AS: 0.0163, HP760AP*: 0.003) and SRK/T (FH5600AS: 0.0161, HP760AP*: -0.0071), this is an unfair comparison, as the latter formulas were optimized but the Nallasamy-G was not optimized via retraining. It is notable that the Nallasamy-G demonstrates much better ME and MAEs

compared to the original Nallasamy Formula, its closest comparable model. Furthermore, all results are comparable with the other models and never underperform all models listed.

Note that Nallasamy-G also outperforms Nallasamy formula in MedAE, with a Median Absolute Error of 0.2271 with the FH5600AS dataset compared with Nallasamy MedAE of 0.3082. With the HP760AP* dataset, Nallasamy-G performs a MedAE of 0.2449 compared with 0.2495 with Nallasamy. With the SN60WF (at Aravind Dataset), Nallasamy-G performs with a MedAE of 0.2093, while Nallasamy is 0.2402. Finally, on the UMich dataset, which both models were trained on, Nallasamy-G performs with a MedAE of 0.2403 and Nallasamy 0.2423. Nallasamy-G also provides a higher proportion of Absolute Errors $<$ 0.5 D, with the following percentages (FH5600AS: 0.8513, HP760AP*: 0.8158, SN60WF: 0.8843, UMich: 0.8016). This contrasts with the lower or equivalent results from Nallasamy (FH5600AS: 0.7504, HP760AP*: 0.8217, SN60WF: 0.8487, UMich: 0.8016). Therefore, even with mean errors (ME) further off from zero than the Nallasamy Formula, Nallasamy-G still confers lower error as demonstrated by MAE, MedAE, and AE < 0.5.

Correlation performance of optimized IOL formulas are given in Table H.1. Although correlations were shown previously in Chapter 6, the results with optimized values demonstrate that correlations for Nallasamy-G and Nallasamy remain the highest among compared IOL formulas. Correlations are illustrated in Figure H.2. A graph of MAE to A-constant for each dataset to demonstrate the optimization results is given in Figure H.3.

Performance with optimized A-constants on Nallasamy-G (Our Method), Nallasamy, Haigis, HofferQ, Holladay1, and SRK/T are given in Figures H.4, H.5, H.6, and H.7.

Results from this appendix demonstrate the effective use of Nallasamy-G even under optimized constants, demonstrating competitive and robust results compared to the models shown. Notably, Nallasamy-G outperforms or matches the Nallasamy Formula in MAE, MedAE and AE$<$ 0.5 even when optimized constants are given.

Dataset	IOL Formula	best A-constant	best ME	best MAE
FH5600AS	Holladay	1.118	0.0006	0.2491
	SRKT	117.883	-0.0001	0.2438
	HofferQ	4.94	-0.0002	0.2783
	Haigis	0.629	-0.0002	0.3024
HP760AP	Holladay	1.76	-0.0003	0.2519
	SRKT	118.9	-0.0003	0.2542
	HofferQ	5.578	-0.0005	0.2798
	Haigis	1.296	0.0006	0.2872
SN60WF	Holladay	1.752	0.0001	0.2364
	SRKT	118.903	-0.0003	0.2325
	HofferQ	5.572	0.0005	0.2756
	Haigis	1.303	0.0001	0.2756
UMich*	Holladay	1.863	0.0004	0.3256
	SRKT	119.092	0.0003	0.3295
	HofferQ	5.725	0.0005	0.3605
	Haigis	-0.727	-0.0003	0.324

Table H.1: Post-hoc optimized A-constants for each tested dataset. The UMich dataset is marked with * because the optimized values used in the rest of the analyses in this section are based on the optimized constants from [106] which contained a larger set of UMich patients than this test set. However, the constants for the test set are similar, reflecting a good fit. Constants a1 and a2 for Haigis were given in accordance with those suggested in ULIB; that is, default constants for FH5600AS and SN60WF (India), a1=0.4,a2=0.1. We assigned HP760AP default constants as well. The UMich SN60WF constants were assigned as a1=0.234 and a2=0.217 as suggested by ULIB.

DS	Formula	MAE	ME	MedAE	STD	AE< 0.5	m	FPI
A	**Nall-G**	**0.2901**	**0.064**	**0.2271**	**0.4312**	**0.8513**	**0.1326**	**0.3166**
	Nall.	0.374	0.2873	0.3082	0.4373	0.7504	0.101	0.3238
	Haigis	0.3376	0.0091	0.2633	0.489	0.792	-0.0508	0.3963
	HofferQ	0.3155	0.0179	0.2368	0.4677	0.8159	-0.2731	0.2146
	Holl.	0.2841	0.0163	0.2023	0.4396	0.8522	-0.0546	0.4235
	SRK/T	0.2837	0.0161	0.2108	0.4371	0.8513	-0.1043	0.3489
B	**Nall-G**	**0.288**	**-0.0935**	**0.2449**	**0.3549**	**0.8158**	**0.1793**	**0.2764**
	Nall.	0.2982	0.1382	0.2495	0.3457	0.8217	0.1238	0.3278
	Haigis	0.3182	-0.0023	0.2748	0.3956	0.8024	-0.1828	0.267
	HofferQ	0.312	0.0025	0.2679	0.3959	0.7964	-0.3462	0.1858
	Holl.	0.285	0.003	0.2331	0.3651	0.8276	-0.1308	0.3211
	SRK/T	0.2873	-0.0071	0.2241	0.3734	0.8053	-0.2215	0.2467
C	**Nall-G**	**0.2495**	**-0.0149**	**0.2093**	**0.3182**	**0.8843**	**-0.0856**	**0.3977**
	Nall.	0.2842	0.1536	0.2402	0.3189	0.8487	-0.1179	0.3429
	Haigis	0.3049	-0.0095	0.2498	0.3917	0.8071	-0.4557	0.1553
	HofferQ	0.3052	-0.0173	0.2541	0.388	0.8234	-0.7033	0.1125
	Holl.	0.265	-0.0131	0.22	0.3402	0.865	-0.5362	0.1413
	SRK/T	0.2636	-0.0123	0.2078	0.3408	0.8569	-0.6225	0.1259
D	**Nall-G**	**0.3096**	**-0.0293**	**0.2403**	**0.4135**	**0.8016**	**-0.049**	**0.4181**
	Nall.	0.3125	-0.0152	0.2423	0.4179	0.8016	0.0331	0.4467
	Haigis	0.3632	-0.0237	0.2894	0.4685	0.7468	-0.226	0.2295
	HofferQ	0.4038	-0.0091	0.3311	0.5174	0.7029	-0.9513	0.0849
	Holl.	0.3706	-0.0207	0.298	0.4864	0.7398	-0.7733	0.1013
	SRK/T	0.376	-0.0144	0.2999	0.4846	0.7318	-0.4858	0.1427

Table H.2: Optimized results of our generalized formula vs other well-known IOL formulas for the following Datasets (DS): A) FH5600AS, B) HP760AP*, C) SN60WF, D) UMich.

A-Constant	**Dataset**	**Ours**	**Nall.**	**Haigis**	**Hoff.**	**Holl.**	**SRK/T**
Optimized	FH5600AS	**0.2607**	**0.2682**	0.2161	0.1834	0.1807	0.158
	HP760AP	**0.2856**	**0.3336**	0.2387	0.18	0.175	0.0946
	SN60WF	**0.3572**	**0.3742**	0.19	0.1644	0.2125	0.1848
	UMich	**0.8966**	**0.8943**	0.8696	0.8419	0.856	0.8553

Table H.3: Correlations of IOL formula predictions to ground truth post-operative refractions under different A-constants and different datasets. The column "A-Constant" refers to the A-constant used in the analysis: Optimized. Note that both our formula and the Nallasamy formula demonstrate the strongest correlations with post-operative refrection.

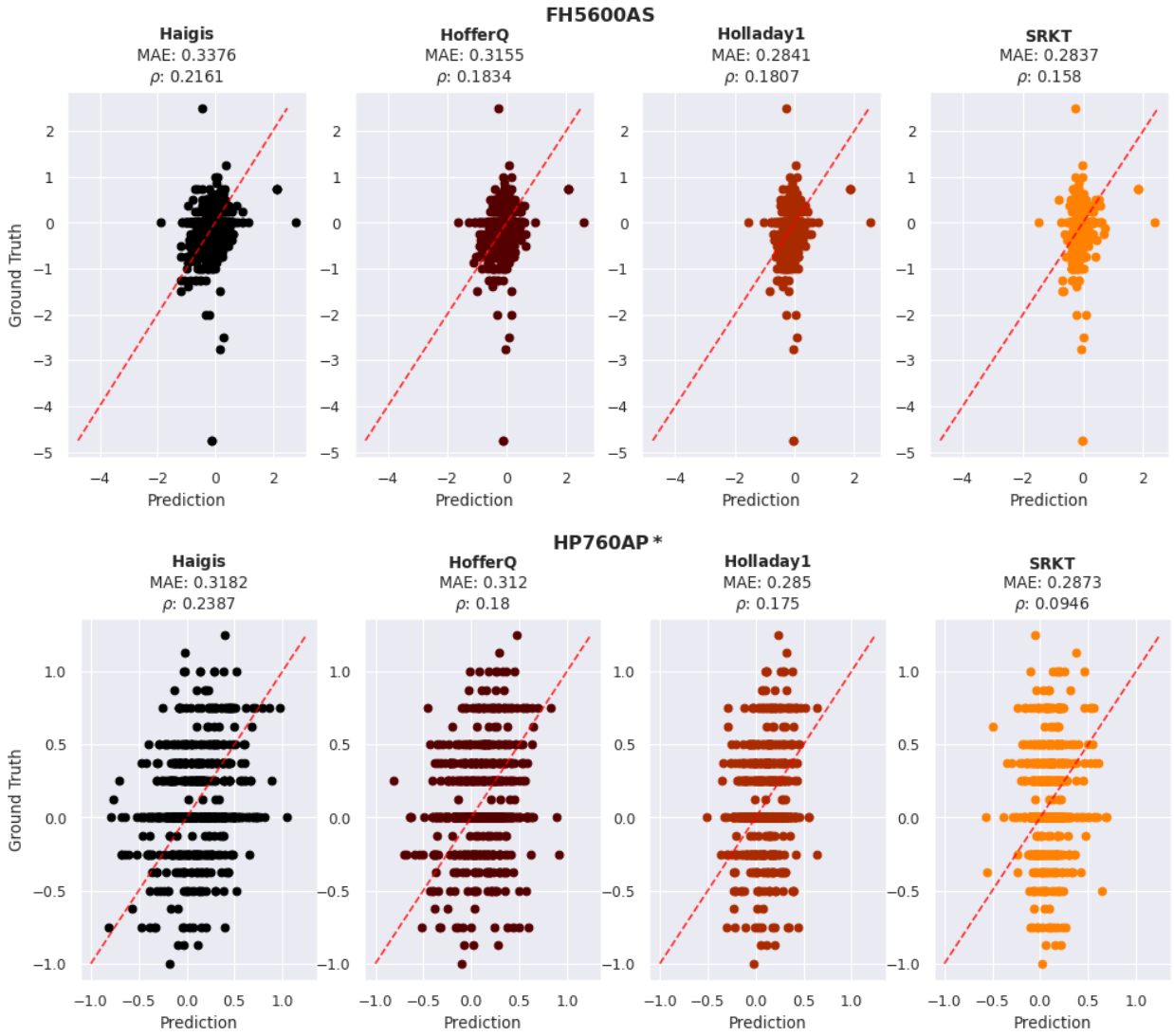

Figure H.1: Correlations of IOL prediction formulas against the true post-operative refractions for the FH5600AS and HP760AP* datasets. These predictions are constructed with optimized A-constants.

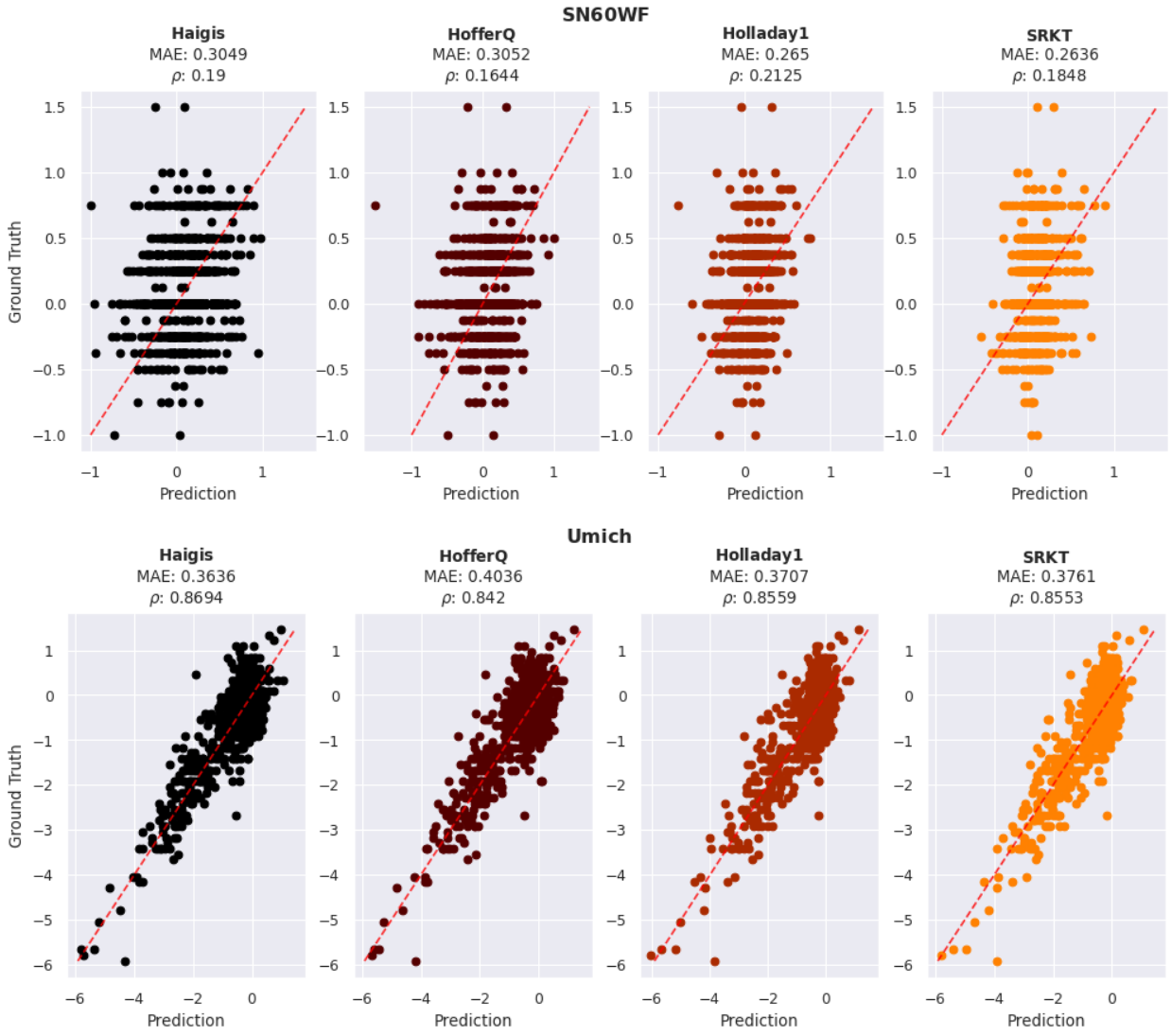

Figure H.2: Correlations of IOL prediction formulas against the true post-operative refractions for the SN60WF and UMich datasets. These predictions are constructed with optimized A-constants.

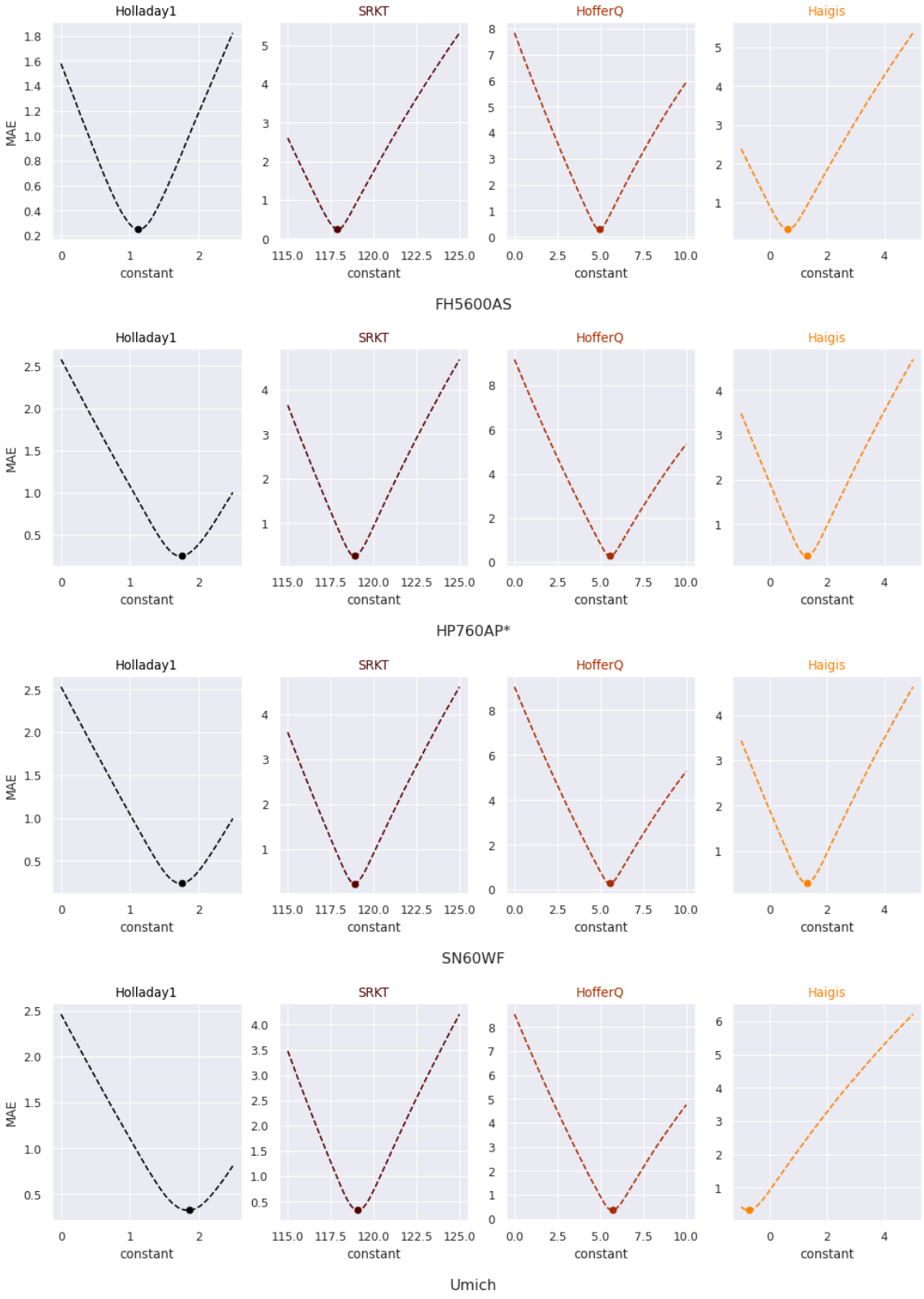

Figure H.3: Optimized constants for each dataset

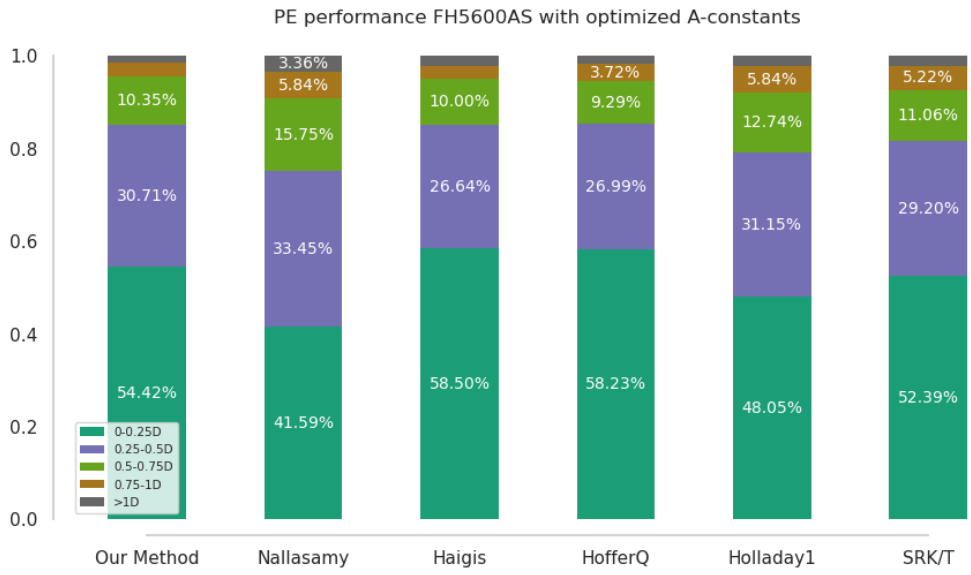

Figure H.4: FH5600AS prediction error breakdown by dioptre (D).

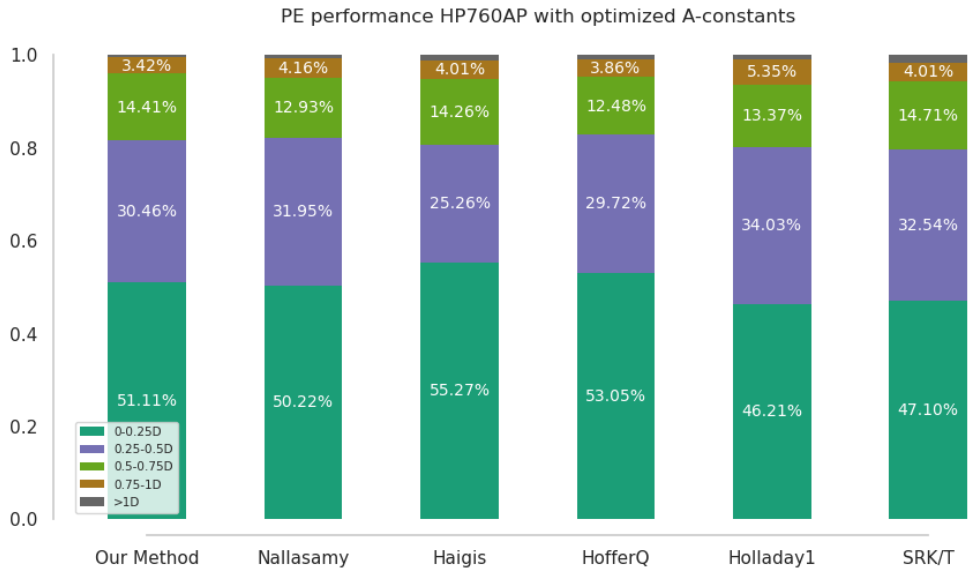

Figure H.5: HP760AP prediction error breakdown by dioptre (D).

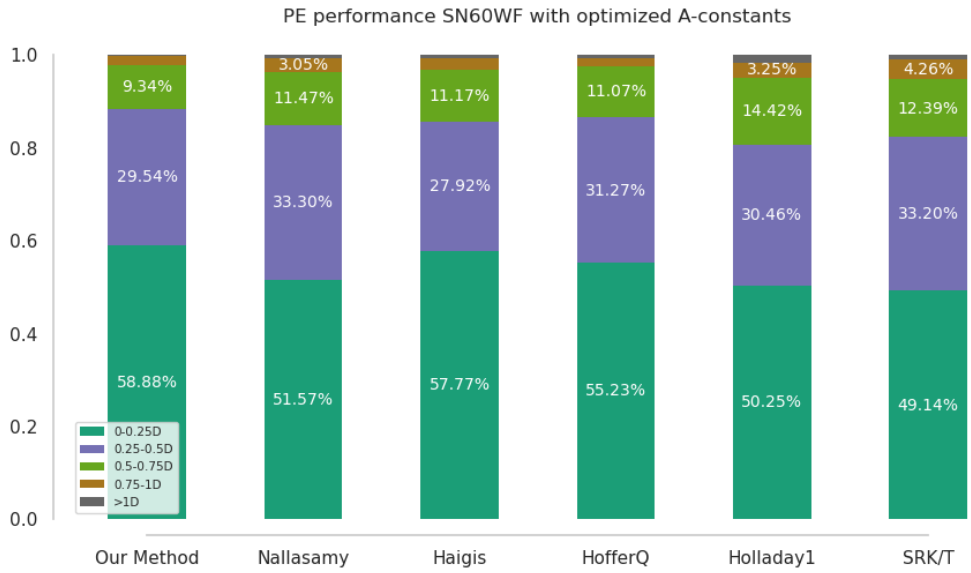

Figure H.6: SN60WF (Aravind) prediction error breakdown by dioptre (D).

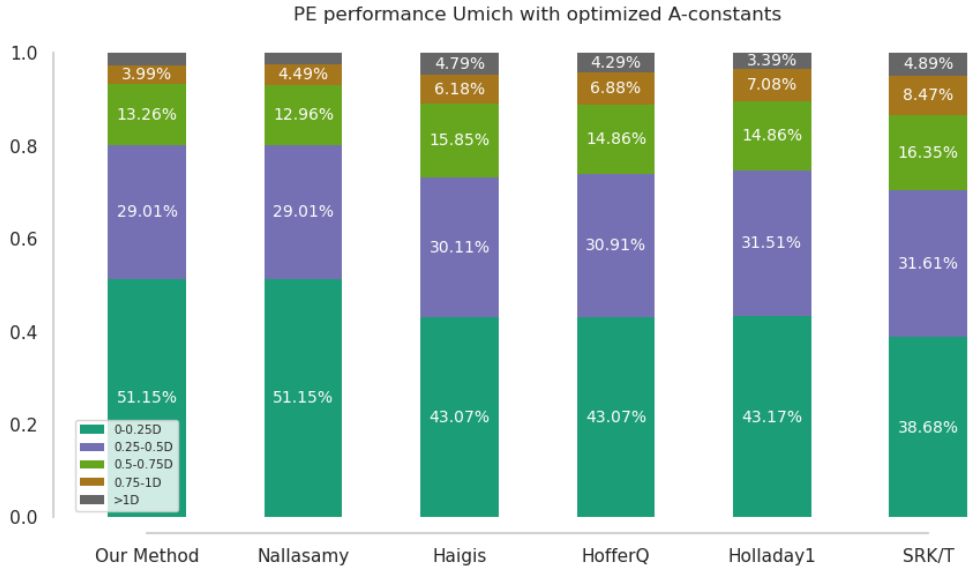

Figure H.7: UMich (SN60WF at University of Michigan) prediction error breakdown by dioptre (D).

APPENDIX I

Analysis of Our Generalized Model as a Predictor of IOL Power

I.1 Introduction

Many IOL formulas are are used to predict the power of an IOL given a target post-operative refraction. In [107], two new metrics called MAEPI and CIR were constructed for evaluating the accuracy of a IOL power prediction model. In this analysis, we reverse our generalized model and assess its ability to predict the power of the implanted intraocular lens given the known post-operative refraction of each patient. To assess the model's IOL power prediction capabilities, we calculate MAEPI and CIR(0), CIR(0.5), and CIR(1).

I.2 Methods

I.2.1 Preprocessing

In order to reverse the model, patients are assessed sequentially in an iterative framework. In each iteration, one patient's AL, ACD, CCT, White-to-white (WTW), LT, Keratometry of the Left Eye (K1), Keratometry of the Right Eye (K2), age at surgery and patient sex are extracted as input to the predictive model. Eighty-one potential IOL powers are generated for a single patient and entered into the model with the patient's data for a total of 81 entries into the model. After receiving the model's output predicted post-operative refraction, and IOL power is chosen based on which power returned the post-operative refraction closest to the patient's true refraction.

IOL power prediction for our generalized model (here referred to as "Nallasamy-II") is compared with the original Nallasamy formula (here referred to as "Nallasamy-I") and PearlDGS, a state-of-the-art machine learning-based IOL power prediction tool which has a model specifically constructed for the SN60WF lens. Our assessment includes prediction both

with the dataset of SN60WF lens implants obtained from University of Michigan Kellogg Eye Center (called "Umich") in Ann Arbor, USA, and the dataset of SN60WF lens implants obtained the Aravind Eye Center in Chennai, Tamil Nadu.

I.2.2 Statistical Assessment

In this investigative study, MAEPI and CIR are evaluated based on the IOL power predicted. MAEPI is calculated with the following equation:

$$MAEPI = \frac{\sum_{i=1}^{n} |p_i - \hat{p}_i|}{n}$$

CIR is calculated as the proportion of predicted IOL powers which fall within 0D, 0.5D or 1.0D from the true implanted IOL. The CIR equations are calculated with the following equations for an error of d dioptres:

$$CIR(d) = \frac{\sum_{i=1}^{n} I(|p_i - \hat{p}_i| = d)}{n} \times 100$$

, where the function $I(\cdot)$ represents an indicator function for whether or not the predicted lens power $\hat{p}_i$ falls within d dioptres of the true IOL power.

It is important to note that the MAEPI ranges from $[0,\ 13.719)$ based on assessments with random values in [107]. The best MAEPI value would be 0 and better predictors of IOL power have lower MAEPI values. By constrast, CIR will range from $[0, 100]$ and a larger value means a better IOL power predictor. A perfect predictor will have a value of 100 for any $CIR(d)$ for all dioptre errors d.

Figures assessing predictions and predicted error are constructed with matplotlib in Python 3.8.

I.3 Results

A table of results for MAEPI, CIR(0), CIR(0.5) and CIR(1) are contained in I.1. The The results indicate that PearlDGS is the best predictor of IOL power with an MAEPI of (0.4412) in the UMich dataset and 0.3452 in the SN60WF (Aravind) dataset, followed by the Nallasamy Formula (Nall-I) with an MAEPI of 0.4521 in the UMich dataset and 0.3574 in the SN60WF (Aravind) dataset and our generalized model (Nall-G), with an MAEPI of 0.5578 in the UMich dataset and 0.6071 in the SN60WF (Aravind) dataset. While the Nallasamy Formula and PearlDGS show a reduction of error in IOL power prediction with the SN60WF (Aravind) dataset compared with the UMich dataset, our generalized model performs worse.

CIR(0) results are given as 33.001 for the Nall-G (Umich), 29.9492 for the Nall-G (SN60WF at Aravind), 35.3939 for the Nall-I (Umich), 41.7259 for the Nall-I (SN60WF at Aravind), 36.989 for PearlDGS (Umich) and 42.0305 for PearlDGS (SN60WF at Aravind). The low values for the Nall-G indicate that our generalized model is the least likely of the three models to correctly predict the IOL lens power, and CIR(1) demonstrates that less than 90% of Nall-G predictions for IOL power are even within 1D, while Nall-G provides predictions within 1D of the correct lens power for 95.81% (UMich) and 98.58% (SN60WF at Aravind) of their datasets, respectively. PearlDGS performs slightly better, with CIR(1) results of 95.5135 and 99.1878 for the UMich and SN60WF at Aravind datasets, respectively.

To better understand the poor comparative performance of our generalized Nall-G model compared with the original Nall-I model, we plot IOL power prediction error of the Nall-G by targeted refraction (the patient's true post-operative refraction) in figure I.1, and compare this with the same assessment for Nall-I (figure I.2). Most notable is that Nall-G contains errors up to 4.5D and exhibits a pattern of exclusively high error when targeted refraction is above 0D. This pattern is less obvious in the Nall-I model.

For further analysis, cases with high error were assessed individually. Graphs of IOL power against predicted post-operative refraction of our generalized Nall-G were compared with those for the Nall-I. Graphs for IOL power against predicted refraction were similar for every patient in the dataset for both Nall-G and Nall-I models. We exhibit a result graph of a patient with one of the largest fail cases for Nall-G in figure I.3 and this same patient for Nall-I in figure I.4. For this patient, target refraction was 1.0886 and the predicted power of the lens given by Nall-G was 12.5D. By contrast, Nall-I predicted a 15.0D lens. The true IOL power implanted was 17.0D. In the graphs, the Nall-G demonstrates a signifantly small slope for in the predicted refraction range greater than 0. while Nall-I demonstrates a near-linear curve across all refractions and IOL powers.

Lastly, predicted refraction error (x-axis) against true refraction (y-axis) is given for Nall-G in figure I.5 and Nall-I in figure I.6. Nall-G and Nall-I graphs look very similar when assessing predicted refraction error given the true IOL lens power.

I.4 Conclusion

Because the paradigm of predicting IOL power given a target post-operative refraction can be a useful tool for clinical use, this small study sought to assess the ability of our generalized formula to work in a reverse fashion where IOL power is predicted given a target refraction. However, based on calculations of MAEPI and CIR, it has been determined that our pretrained generalized formula (here shortened to Nall-G) functions poorly in a reverse

framework where IOL power is predicted using a targeted refraction. With this information and the low MAEs given in chapter 4, we can conclude that our Nall-G formula is exceptional at predicting postoperative refraction but cannot map IOL power back to refraction. Among the largest contributors to a high MAEPI is Nall-G's comparative difficulty in predicting IOL power for positive refractions. This is confirmed with the model's MAEPI for SN60WF at Aravind given in table I.1, which is higher than the performance of the UMich dataset. It is also known from chapter 3 figure 3.1 that the SN60WF dataset contains more positive predictions than the UMich dataset. The Aravind dataset contains more positive refractions because every patient targeted emmetropia (0D), while many patients at the University of Michigan targeted close-up reading ability, requiring a negative refraction. Therefore, the UMIch dataset from fig 3.1 contains a significantly higher proportion of patients who have negative post-operative refractions, leading to a lack of data on positive refractions in the training set.

To understand why the Nall-G has difficulty in assessing positive refractions despite its high MAE results and generalizability shown in chapter 4, curves of IOL power versus the Nall-G's refraction was given in figure I.3. In the figure, it becomes clear that Nall-G has not developed a strong relationship between postoperative refraction and IOL power, likely due to the lack of data. As a result, to reduce MAE, the Nall-G has zeroed out the slope of low-power IOLs to some approximate averages in post-operative refraction. While this approach has been learned to reduce MAE, it is not a good method for predicting the reverse case of IOL power, because many powers can relate to a similar predicted refraction, thus increasing the probability of error. Note that in figure I.4, the Nall-I does not demonstrate the near-zero slopes for low IOL powers. This is likely due to a lack of features which results in a relatively poorer fit to post-operative refraction (as reflected in MAE in chapter 4, but a better fit for IOL power prediction.

Although it may be reasonably assumed that predictors which map a set of data X to an outcome Y would also be capable of reverse mapping Y to X like a mathematical equation, this bi-directional assumption does not often hold true in machine learning cases. Bi-directionality is a feature of a trained mapping function that often has to be intentionally integrated into the design of the model. This can be demonstrated by the Nall-G formula, which from chapter 4 is shown to be an excellent generalizable model for post-operative refraction prediction but functions as a poor IOL power predictor in the reverse. In fact, we hypothesize that the very presence of additional features which makes the model generalizable are also the cause of the overfitting to post-operative refraction that make the model irreversible.

Formula	Dataset	MAEPI	CIR(0)	CIR(0.5)	CIR(1)
Nall-G	Umich	0.5578	33.001	75.0748	89.6311
	SN60WF	0.6071	29.9492	70.2538	87.6142
Nall-I	Umich	0.4521	35.3939	80.3589	95.8126
	SN60WF	0.3574	41.7259	88.4264	98.5787
PearlDGS	Umich	0.4412	36.989	81.3559	95.5135
	SN60WF	0.3452	42.0305	90.1523	99.1878

Table I.1: MAEPI, CIR(0), CIR(0.5), and CIR(1) performance comparing our generalized method (Nall-G) to the original Nallasamy Formula (Nall-I) and another machine learning-based predictor of IOL power.

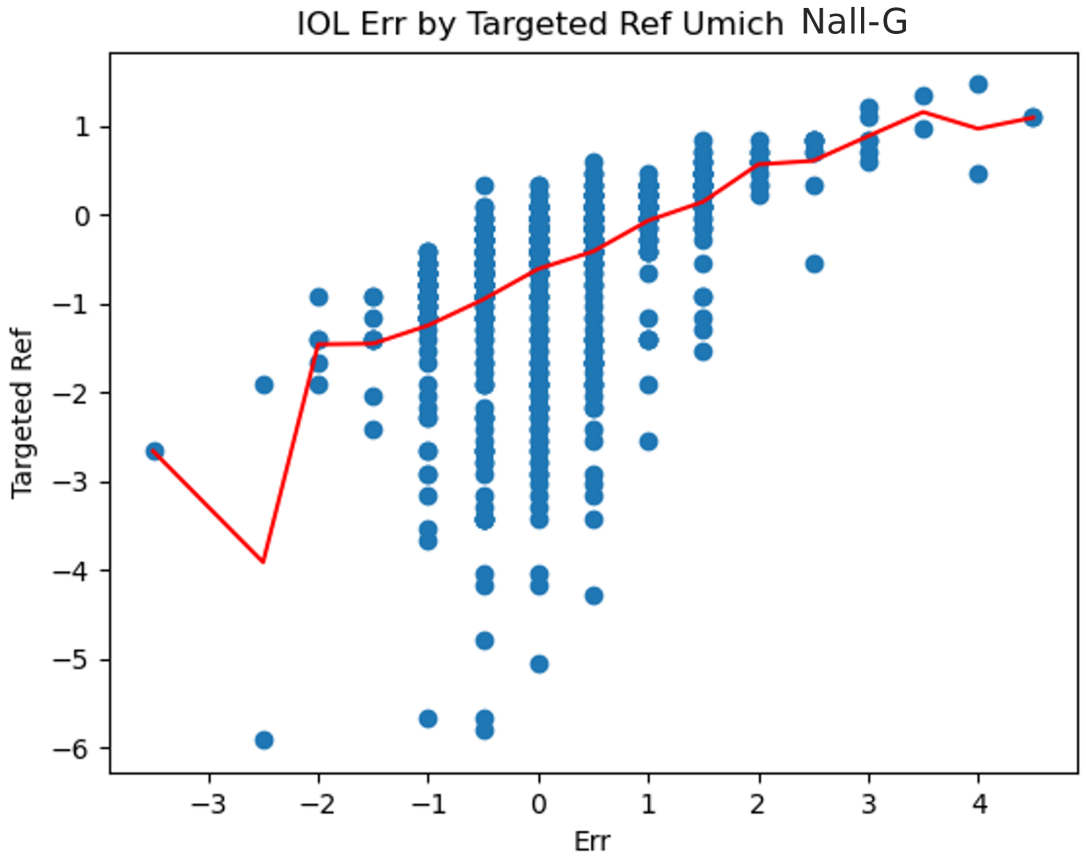

Figure I.1: Error in IOL power prediction by target refraction for our generalized model (Nallasamy-II)

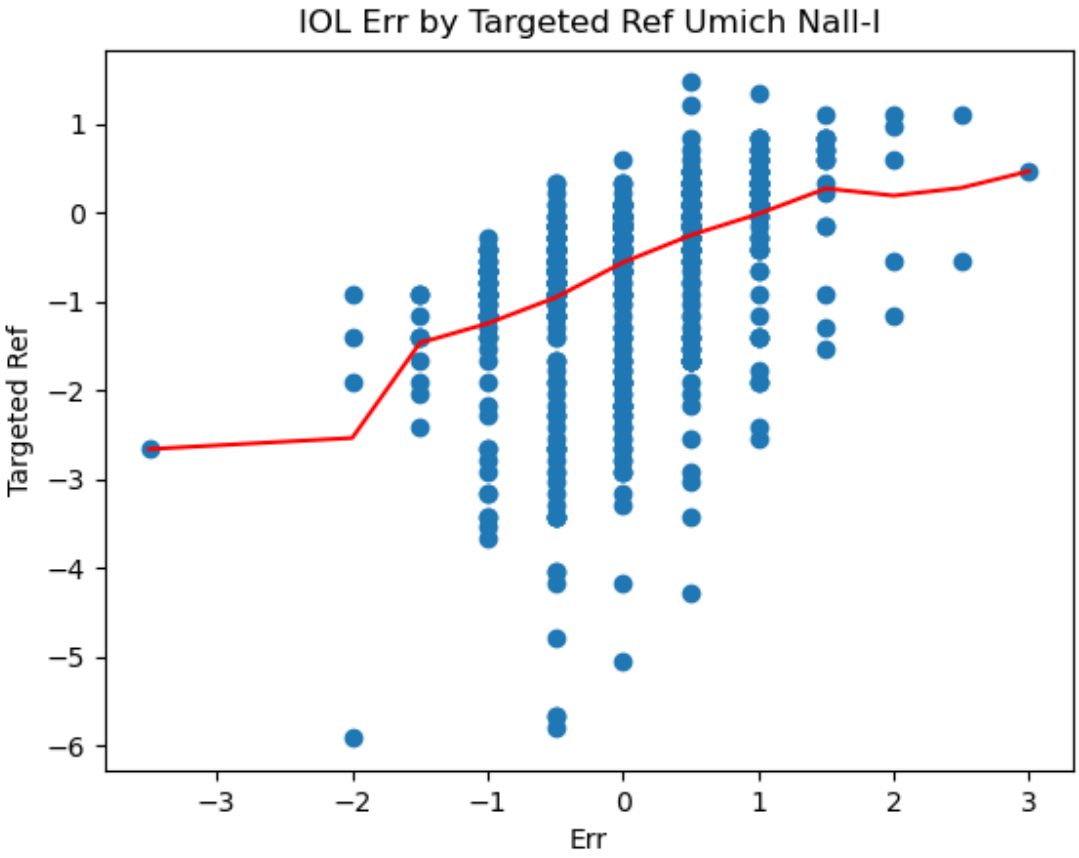

Figure I.2: Error in IOL power prediction by target refraction for the Nallasamy Formula (Nallasamy-I)

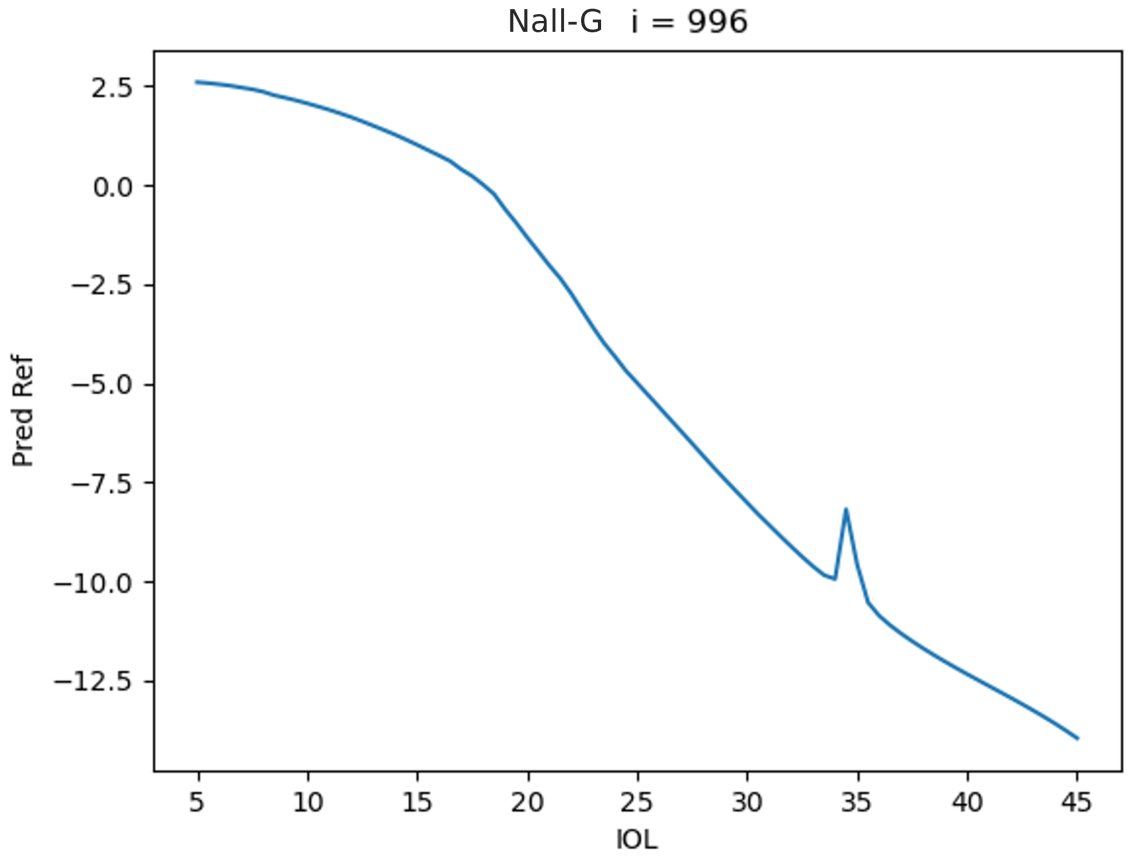

Figure I.3: Input IOL power vs the predicted refraction output by our generalized model (Nallasamy-II). This graph shows the output for a single patient (identified only as patient 996) who had an actual implant of 17.0D and a post-operative refraction 1.0886. The Nallasamy-II predicts she needs a 12.5D lens for a total error or 4.5D.

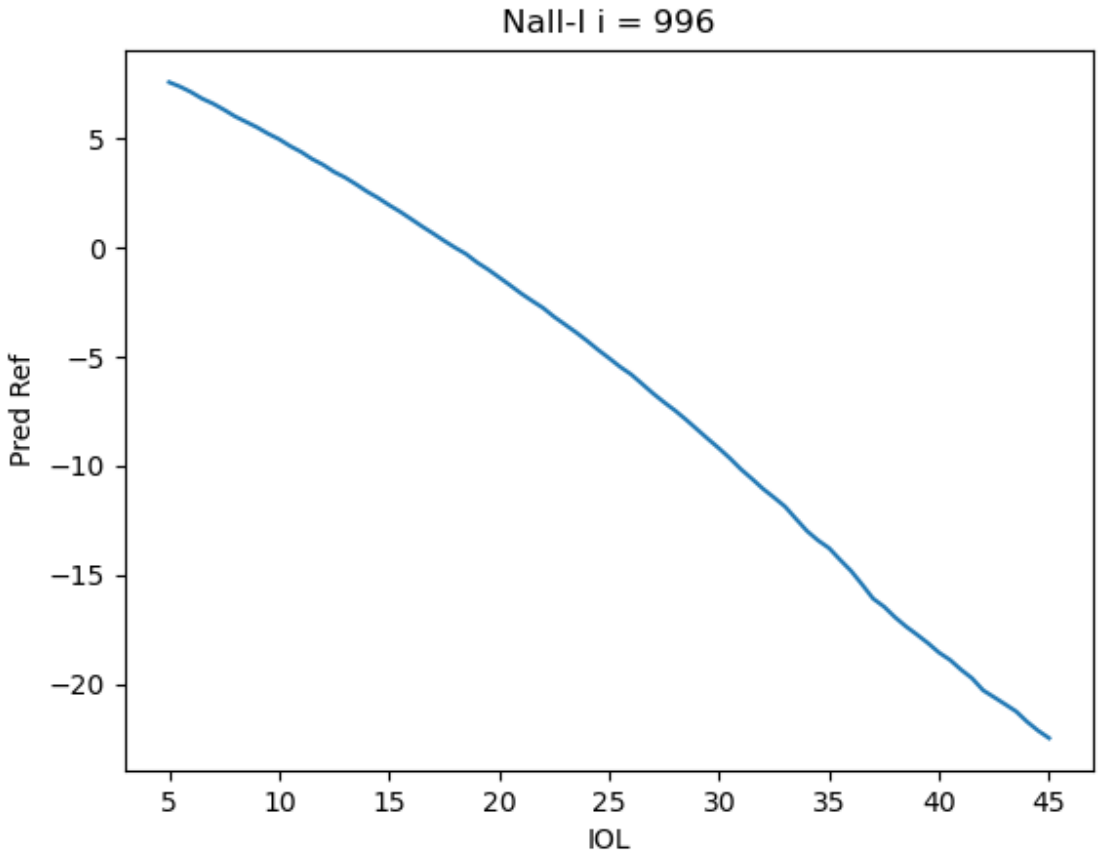

Figure I.4: Input IOL power vs the predicted refraction output by the Nallasamy Formula (Nallasamy-I). This graph shows the output for a single patient (identified only as patient 996) who had an actual implant of 17.0D and a post-operative refraction 1.0886. The Nallasamy-I predicts she needs a 15.0D lens for a total error or 2.0D.

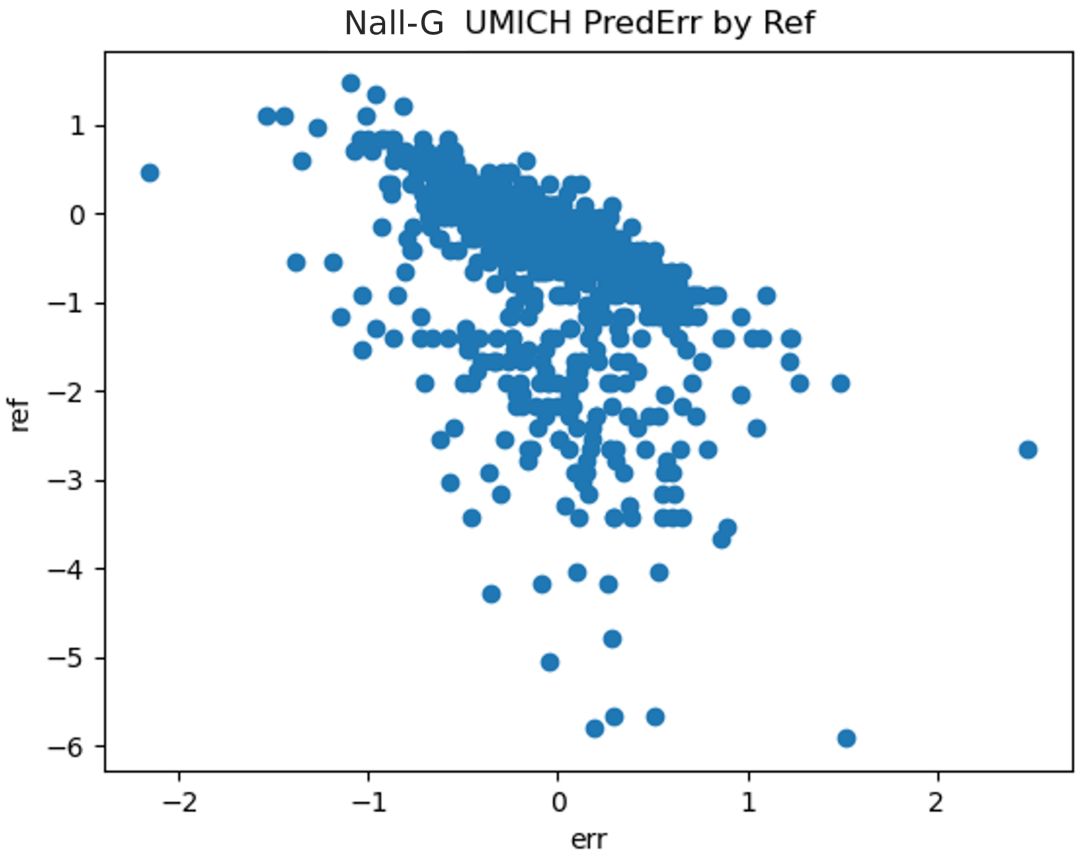

Figure I.5: A scatter plot of post-operative refraction prediction error by ground truth postoperative refraction for our generalized formula (Nallasamy-II)

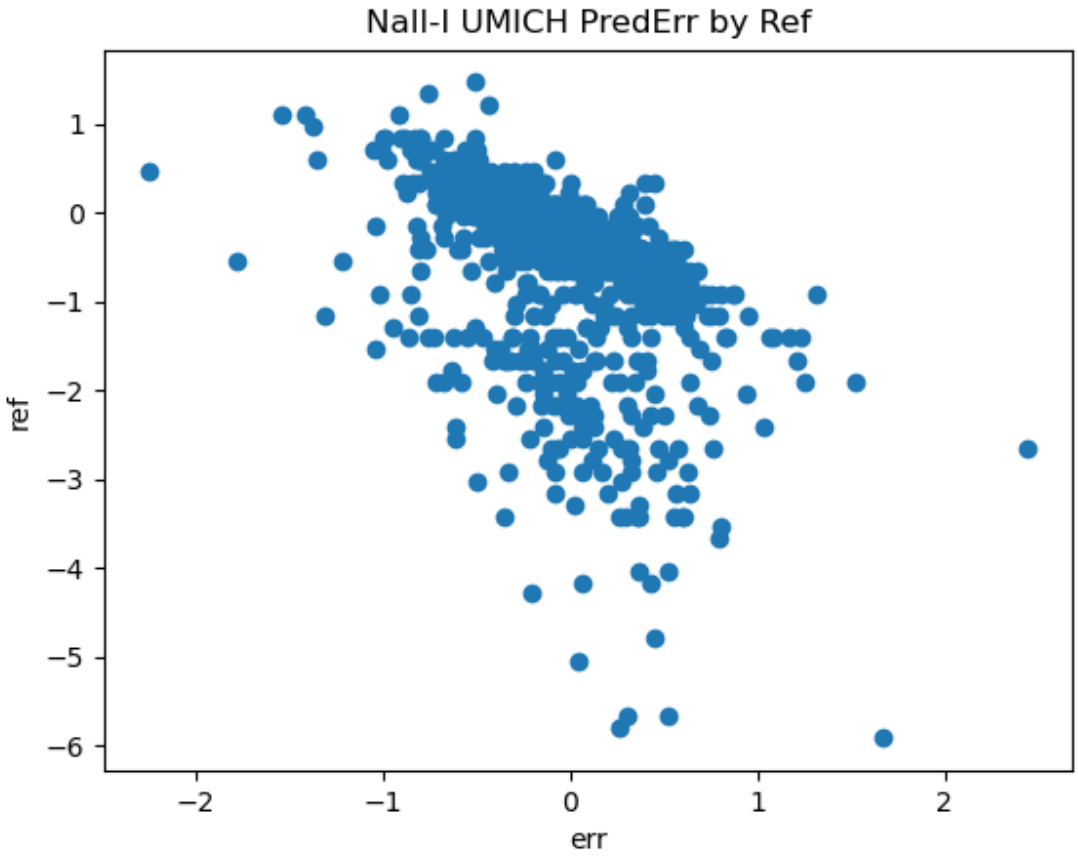

Figure I.6: A scatter plot of post-operative refraction prediction error by ground truth postoperative refraction for the Nallasamy Formula (Nallasamy-I)

BIBLIOGRAPHY

[1] Exposure to solar ultraviolet (uv) radiation data by country. https://apps.who.int/gho/data/view.main.35300. Accessed: 2023-12-18.

[2] Ulib (user group for laser intereference biometry). http://ocusoft.de/ulib/c1.htm. Accessed: 2024-03-01.

[3] Moloud Abdar, Maryam Samami, Sajjad Dehghani Mahmoodabad, Thang Doan, Bogdan Mazoure, Reza Hashemifesharaki, Li Liu, Abbas Khosravi, U. Rajendra Acharya, Vladimir Makarenkov, and Saeid Nahavandi. Uncertainty quantification in skin cancer classification using three-way decision-based bayesian deep learning. *Computers in Biology and Medicine*, 135:104418, August 2021.

[4] Khalia Ackermann, Jannah Baker, Malcolm Green, and et al. Computerized clinical decision support systems for the early detection of sepsis among adult inpatients: Scoping review. *Journal of Medical Internet Research*, 24(2):e31083, February 2022.

[5] Adewole S. Adamson and H. Gilbert Welch. Machine learning and the cancer-diagnosis problem — no gold standard. *New England Journal of Medicine*, 381(24):2285–2287, December 2019.

[6] Elsa Aghaian, Joyce E. Choe, Shan Lin, and Robert L. Stamper. Central corneal thickness of caucasians, chinese, hispanics, filipinos, african americans, and japanese in a glaucoma clinic. *Ophthalmology*, 111(12):2211–2219, December 2004.

[7] Tahra AlMahmoud, David Priest, Rejean Munger, and W. Bruce Jackson. Correlation between refractive error, corneal power, and thickness in a large population with a wide range of ametropia. *Investigative Opthalmology & Visual Science*, 52(3):1235, March 2011.

[8] Deepak Anand, Darshan Tank, Harshvardhan Tibrewal, and Amit Sethi. Self-supervision vs. transfer learning: Robust biomedical image analysis against adversarial attacks. In *2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI)*. IEEE, April 2020.

[9] Jessica S. Ancker, Alison Edwards, Sarah Nosal, Diane Hauser, Elizabeth Mauer, and Rainu Kaushal. Effects of workload, work complexity, and repeated alerts on alert fatigue in a clinical decision support system. *BMC Medical Informatics and Decision Making*, 17(1), April 2017.

[10] Chiaki Asao, Yukunori Korogi, Mika Kitajima, Toshinori Hirai, Yuji Baba, Keishi Makino, Masato Kochi, Shoji Morishita, and Yasuyuki Yamashita. Diffusion-weighted imaging in the follow-up of treated high-grade gliomas: Tumor recurrence versus radiation injury. *American Journal of Neuroradiology*, 26(6):1455–1460, June 2005.

[11] Emanuel A. Azcona, Pierre Besson, Yunan Wu, Arjun Punjabi, Adam Martersteck, Amil Dravid, Todd B. Parrish, S. Kathleen Bandt, and Aggelos K. Katsaggelos. Interpretation of brain morphology in association to alzheimer's disease dementia classification using graph convolutional networks on triangulated meshes. In *Shape in Medical Imaging*, pages 95–107. Springer International Publishing, New York, 2020.

[12] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. In Yoshua Bengio and Yann LeCun, editors, *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015.

[13] Spyridon Bakas, Hamed Akbari, Aristeidis Sotiras, Michel Bilello, Martin Rozycki, Justin S. Kirby, John B. Freymann, Keyvan Farahani, and Christos Davatzikos. Advancing the cancer genome atlas glioma MRI collections with expert segmentation labels and radiomic features. *Scientific Data*, 4(1), September 2017.

[14] Spyridon Bakas, Mauricio Reyes, Andras Jakab, Stefan Bauer, Markus Rempfler, Alessandro Crimi, Russell Takeshi Shinohara, Christoph Berger, Sung Min Ha, Martin Rozycki, Marcel Prastawa, Esther Alberts, Jana Lipkova, John Freymann, Justin Kirby, Michel Bilello, Hassan Fathallah-Shaykh, Roland Wiest, Jan Kirschke, Benedikt Wiestler, Rivka Colen, Aikaterini Kotrotsou, Pamela Lamontagne, Daniel Marcus, Mikhail Milchenko, Arash Nazeri, Marc-Andre Weber, Abhishek Mahajan, Ujjwal Baid, Elizabeth Gerstner, Dongjin Kwon, Gagan Acharya, Manu Agarwal, Mahbubul Alam, Alberto Albiol, Antonio Albiol, Francisco J. Albiol, Varghese Alex, Nigel Allinson, Pedro H. A. Amorim, Abhijit Amrutkar, Ganesh Anand, Simon Andermatt, Tal Arbel, Pablo Arbelaez, Aaron Avery, Muneeza Azmat, Pranjal B., W Bai, Subhashis Banerjee, Bill Barth, Thomas Batchelder, Kayhan Batmanghelich, Enzo Battistella, Andrew Beers, Mikhail Belyaev, Martin Bendszus, Eze Benson, Jose Bernal, Halandur Nagaraja Bharath, George Biros, Sotirios Bisdas, James Brown, Mariano Cabezas, Shilei Cao, Jorge M. Cardoso, Eric N Carver, Adrià Casamitjana, Laura Silvana Castillo, Marcel Catà, Philippe Cattin, Albert Cerigues, Vinicius S. Chagas, Siddhartha Chandra, Yi-Ju Chang, Shiyu Chang, Ken Chang, Joseph Chazalon, Shengcong Chen, Wei Chen, Jefferson W Chen, Zhaolin Chen, Kun Cheng, Ahana Roy Choudhury, Roger Chylla, Albert Clérigues, Steven Colleman, Ramiro German Rodriguez Colmeiro, Marc Combalia, Anthony Costa, Xiaomeng Cui, Zhenzhen Dai, Lutao Dai, Laura Alexandra Daza, Eric Deutsch, Changxing Ding, Chao Dong, Shidu Dong, Wojciech Dudzik, Zach Eaton-Rosen, Gary Egan, Guilherme Escudero, Théo Estienne, Richard Everson, Jonathan Fabrizio, Yong Fan, Longwei Fang, Xue Feng, Enzo Ferrante, Lucas Fidon, Martin Fischer, Andrew P. French, Naomi Fridman, Huan Fu, David Fuentes, Yaozong Gao, Evan Gates, David Gering, Amir Gholami, Willi Gierke, Ben Glocker, Mingming Gong, Sandra González-Villá, T. Grosges, Yuanfang

Guan, Sheng Guo, Sudeep Gupta, Woo-Sup Han, Il Song Han, Konstantin Harmuth, Huiguang He, Aura Hernández-Sabaté, Evelyn Herrmann, Naveen Himthani, Winston Hsu, Cheyu Hsu, Xiaojun Hu, Xiaobin Hu, Yan Hu, Yifan Hu, Rui Hua, Teng-Yi Huang, Weilin Huang, Sabine Van Huffel, Quan Huo, Vivek HV, Khan M. Iftekharuddin, Fabian Isensee, Mobarakol Islam, Aaron S. Jackson, Sachin R. Jambawalikar, Andrew Jesson, Weijian Jian, Peter Jin, V Jeya Maria Jose, Alain Jungo, B Kainz, Konstantinos Kamnitsas, Po-Yu Kao, Ayush Karnawat, Thomas Kellermeier, Adel Kermi, Kurt Keutzer, Mohamed Tarek Khadir, Mahendra Khened, Philipp Kickingereder, Geena Kim, Nik King, Haley Knapp, Urspeter Knecht, Lisa Kohli, Deren Kong, Xiangmao Kong, Simon Koppers, Avinash Kori, Ganapathy Krishnamurthi, Egor Krivov, Piyush Kumar, Kaisar Kushibar, Dmitrii Lachinov, Tryphon Lambrou, Joon Lee, Chengen Lee, Yuehchou Lee, M Lee, Szidonia Lefkovits, Laszlo Lefkovits, James Levitt, Tengfei Li, Hongwei Li, Wenqi Li, Hongyang Li, Xiaochuan Li, Yuexiang Li, Heng Li, Zhenye Li, Xiaoyu Li, Zeju Li, XiaoGang Li, Wenqi Li, Zheng-Shen Lin, Fengming Lin, Pietro Lio, Chang Liu, Boqiang Liu, Xiang Liu, Mingyuan Liu, Ju Liu, Luyan Liu, Xavier Llado, Marc Moreno Lopez, Pablo Ribalta Lorenzo, Zhentai Lu, Lin Luo, Zhigang Luo, Jun Ma, Kai Ma, Thomas Mackie, Anant Madabushi, Issam Mahmoudi, Klaus H. Maier-Hein, Pradipta Maji, CP Mammen, Andreas Mang, B. S. Manjunath, Michal Marcinkiewicz, S McDonagh, Stephen McKenna, Richard McKinley, Miriam Mehl, Sachin Mehta, Raghav Mehta, Raphael Meier, Christoph Meinel, Dorit Merhof, Craig Meyer, Robert Miller, Sushmita Mitra, Aliasgar Moiyadi, David Molina-Garcia, Miguel A. B. Monteiro, Grzegorz Mrukwa, Andriy Myronenko, Jakub Nalepa, Thuyen Ngo, Dong Nie, Holly Ning, Chen Niu, Nicholas K Nuechterlein, Eric Oermann, Arlindo Oliveira, Diego D. C. Oliveira, Arnau Oliver, Alexander F. I. Osman, Yu-Nian Ou, Sebastien Ourselin, Nikos Paragios, Moo Sung Park, Brad Paschke, J. Gregory Pauloski, Kamlesh Pawar, Nick Pawlowski, Linmin Pei, Suting Peng, Silvio M. Pereira, Julian Perez-Beteta, Victor M. Perez-Garcia, Simon Pezold, Bao Pham, Ashish Phophalia, Gemma Piella, G. N. Pillai, Marie Piraud, Maxim Pisov, Anmol Popli, Michael P. Pound, Reza Pourreza, Prateek Prasanna, Vesna Prkovska, Tony P. Pridmore, Santi Puch, Élodie Puybareau, Buyue Qian, Xu Qiao, Martin Rajchl, Swapnil Rane, Michael Rebsamen, Hongliang Ren, Xuhua Ren, Karthik Revanuru, Mina Rezaei, Oliver Rippel, Luis Carlos Rivera, Charlotte Robert, Bruce Rosen, Daniel Rueckert, Mohammed Safwan, Mostafa Salem, Joaquim Salvi, Irina Sanchez, Irina Sánchez, Heitor M. Santos, Emmett Sartor, Dawid Schellingerhout, Klaudius Scheufele, Matthew R. Scott, Artur A. Scussel, Sara Sedlar, Juan Pablo Serrano-Rubio, N. Jon Shah, Nameetha Shah, Mazhar Shaikh, B. Uma Shankar, Zeina Shboul, Haipeng Shen, Dinggang Shen, Linlin Shen, Haocheng Shen, Varun Shenoy, Feng Shi, Hyung Eun Shin, Hai Shu, Diana Sima, M Sinclair, Orjan Smedby, James M. Snyder, Mohammadreza Soltaninejad, Guidong Song, Mehul Soni, Jean Stawiaski, Shashank Subramanian, Li Sun, Roger Sun, Jiawei Sun, Kay Sun, Yu Sun, Guoxia Sun, Shuang Sun, Yannick R Suter, Laszlo Szilagyi, Sanjay Talbar, Dacheng Tao, Dacheng Tao, Zhongzhao Teng, Siddhesh Thakur, Meenakshi H Thakur, Sameer Tharakan, Pallavi Tiwari, Guillaume Tochon, Tuan Tran, Yuhsiang M. Tsai, Kuan-Lun Tseng, Tran Anh Tuan, Vadim Turlapov, Nicholas Tustison, Maria Vakalopoulou, Sergi Valverde, Rami Vanguri, Evgeny Vasiliev, Jonathan Ventura, Luis Vera, Tom Vercauteren, C. A. Ver-

rastro, Lasitha Vidyaratne, Veronica Vilaplana, Ajeet Vivekanandan, Guotai Wang, Qian Wang, Chiatse J. Wang, Weichung Wang, Duo Wang, Ruixuan Wang, Yuanyuan Wang, Chunliang Wang, Guotai Wang, Ning Wen, Xin Wen, Leon Weninger, Wolfgang Wick, Shaocheng Wu, Qiang Wu, Yihong Wu, Yong Xia, Yanwu Xu, Xiaowen Xu, Peiyuan Xu, Tsai-Ling Yang, Xiaoping Yang, Hao-Yu Yang, Junlin Yang, Haojin Yang, Guang Yang, Hongdou Yao, Xujiong Ye, Changchang Yin, Brett Young-Moxon, Jinhua Yu, Xiangyu Yue, Songtao Zhang, Angela Zhang, Kun Zhang, Xuejie Zhang, Lichi Zhang, Xiaoyue Zhang, Yazhuo Zhang, Lei Zhang, Jianguo Zhang, Xiang Zhang, Tianhao Zhang, Sicheng Zhao, Yu Zhao, Xiaomei Zhao, Liang Zhao, Yefeng Zheng, Liming Zhong, Chenhong Zhou, Xiaobing Zhou, Fan Zhou, Hongtu Zhu, Jin Zhu, Ying Zhuge, Weiwei Zong, Jayashree Kalpathy-Cramer, Keyvan Farahani, Christos Davatzikos, Koen van Leemput, and Bjoern Menze. Identifying the best machine learning algorithms for brain tumor segmentation, progression assessment, and overall survival prediction in the brats challenge, 2019.

[15] Tadas Baltrusaitis, Chaitanya Ahuja, and Louis-Philippe Morency. Multimodal machine learning: A survey and taxonomy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(2):423–443, February 2019.

[16] Graham D. Barrett. Intraocular lens calculation formulas for new intraocular lens implants. *Journal of Cataract and Refractive Surgery*, 13(4):389–396, July 1987.

[17] Graham D. Barrett. An improved universal theoretical formula for intraocular lens power prediction. *Journal of Cataract and Refractive Surgery*, 19(6):713–720, November 1993.

[18] David Bau, Bolei Zhou, Aditya Khosla, Aude Oliva, and Antonio Torralba. Network dissection: Quantifying interpretability of deep visual representations. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pages 3319–3327. IEEE Computer Society, 2017.

[19] Yohann Benchetrit, Hubert Banville, and Jean-Rémi King. Brain decoding: toward real-time reconstruction of visual perception, 2023.

[20] Y. Bengio, A. Courville, and P. Vincent. Representation learning: A review and new perspectives. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(8):1798–1828, August 2013.

[21] James Betker, Gabriel Goh, Li Jing, Tim Brooks, Jianfeng Wang, Linjie Li, Long Ouyang, Juntang Zhuang, Joyce Lee, Yufei Guo, Wesam Manassra, Prafulla Dhariwal, Casey Chu, Yunxin Jiao, and Aditya Ramesh. Improving image generation with better captions. 2023.

[22] Riddhish Bhalodia, Ali Hatamizadeh, Leo Tam, Ziyue Xu, Xiaosong Wang, Evrim Turkbey, and Daguang Xu. Improving pneumonia localization via cross-attention on medical images and reports. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2021*, pages 571–581. Springer International Publishing, 2021.

[23] Jonas Bianchi, Antônio Carlos de Oliveira Ruellas, João Roberto Gonçalves, and et al. Osteoarthritis of the temporomandibular joint can be diagnosed earlier using biomarkers and machine learning. *Scientific Reports*, 10(1), May 2020.

[24] Jonas Bianchi, João Roberto Gonçalves, Antonio Carlos de Oliveira Ruellas, and et al. Software comparison to analyze bone radiomics from high resolution CBCT scans of mandibular condyles. *Dentomaxillofacial Radiology*, 48(6):20190049, September 2019.

[25] Alexandre Bône, Paul Vernhet, Olivier Colliot, and Stanley Durrleman. Learning joint shape and appearance representations with metamorphic auto-encoders. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2020*, pages 202–211. Springer International Publishing, New York, 2020.

[26] Tiffani J. Bright, Anthony Wong, Ravi Dhurjati, Erin Bristow, Lori Bastian, Remy R. Coeytaux, Gregory Samsa, Vic Hasselblad, John W. Williams, Michael D. Musty, Liz Wing, Amy S. Kendrick, Gillian D. Sanders, and David Lobach. Effect of clinical decision-support systems: A systematic review. *Annals of Internal Medicine*, 157(1):29, July 2012.

[27] Toan Duc Bui, Manh Nguyen, Ngan Le, and Khoa Luu. Flow-based deformation guidance for unpaired multi-contrast MRI image-to-image translation. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2020*, pages 728–737. Springer International Publishing, New York, 2020.

[28] Kaidi Cao, Jing Liao, and Lu Yuan. CariGANs. *ACM Transactions on Graphics*, 37(6):1–14, December 2018.

[29] Eduard Lloret Carbonell, Yiqing Shen, Xin Yang, and Jing Ke. COVID-19 pneumonia classification with transformer from incomplete modalities. In *Lecture Notes in Computer Science*, pages 379–388. Springer Nature Switzerland, Basel, 2023.

[30] L.H.S. Cevidanes, D. Walker, J. Schilling, and et al. 3d osteoarthritic changes in TMJ condylar morphology correlates with specific systemic and local biomarkers of disease. *Osteoarthritis and Cartilage*, 22(10):1657–1667, October 2014.

[31] Geeticka Chauhan, Ruizhi Liao, William Wells, Jacob Andreas, Xin Wang, Seth Berkowitz, Steven Horng, Peter Szolovits, and Polina Golland. Joint modeling of chest radiographs and radiology reports for pulmonary edema assessment. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2020*, pages 529–539. Springer International Publishing, New York, 2020.

[32] Geeticka Chauhan, Ruizhi Liao, William Wells, and et al. Joint modeling of chest radiographs and radiology reports for pulmonary edema assessment, 2020.

[33] Richard J. Chen, Ming Y. Lu, Tiffany Y. Chen, Drew F. K. Williamson, and Faisal Mahmood. Synthetic data in machine learning for medicine and healthcare. *Nature Biomedical Engineering*, 5(6):493–497, June 2021.

[34] Richard J. Chen, Ming Y. Lu, Jingwen Wang, Drew F. K. Williamson, Scott J. Rodig, Neal I. Lindeman, and Faisal Mahmood. Pathomic fusion: An integrated framework for fusing histopathology and genomic features for cancer diagnosis and prognosis. *IEEE Transactions on Medical Imaging*, pages 1–1, 2020.

[35] Xi Chen, Yan Duan, Rein Houthooft, John Schulman, Ilya Sutskever, and Pieter Abbeel. Infogan: Interpretable representation learning by information maximizing generative adversarial nets. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, NIPS'16, page 2180–2188, Red Hook, NY, USA, 2016. Curran Associates Inc.

[36] Zhen Chen, Qingyu Guo, Leo K. T. Yeung, Danny T. M. Chan, Zhen Lei, Hongbin Liu, and Jinqiao Wang. Surgical video captioning with mutual-modal concept alignment. In *Lecture Notes in Computer Science*, pages 24–34. Springer Nature Switzerland, New York, 2023.

[37] Phillip M. Cheng and Harshawn S. Malhi. Transfer learning with convolutional neural networks for classification of abdominal ultrasound images. *Journal of Digital Imaging*, 30(2):234–243, November 2016.

[38] Eleni Chiou, Francesco Giganti, Shonit Punwani, Iasonas Kokkinos, and Eleftheria Panagiotaki. Harnessing uncertainty in domain adaptation for MRI prostate lesion segmentation. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2020*, pages 510–520. Springer International Publishing, New York, 2020.

[39] Edward Choi, Siddharth Biswal, Bradley Malin, Jon Duke, Walter F. Stewart, and Jimeng Sun. Generating multi-label discrete patient records using generative adversarial networks. In Finale Doshi-Velez, Jim Fackler, David Kale, Rajesh Ranganath, Byron Wallace, and Jenna Wiens, editors, *Proceedings of the 2nd Machine Learning for Healthcare Conference*, volume 68 of *Proceedings of Machine Learning Research*, pages 286–305. PMLR, 18–19 Aug 2017.

[40] Kenneth Clark, Bruce Vendt, Kirk Smith, John Freymann, Justin Kirby, Paul Koppel, Stephen Moore, Stanley Phillips, David Maffitt, Michael Pringle, Lawrence Tarbox, and Fred Prior. The cancer imaging archive (TCIA): Maintaining and operating a public information repository. *Journal of Digital Imaging*, 26(6):1045–1057, July 2013.

[41] Can Cui, Han Liu, Quan Liu, Ruining Deng, Zuhayr Asad, Yaohong Wang, Shilin Zhao, Haichun Yang, Bennett A. Landman, and Yuankai Huo. Survival prediction of brain cancer with incomplete radiology, pathology, genomic, and demographic data. In *Lecture Notes in Computer Science*, pages 626–635. Springer Nature Switzerland, Basel, 2022.

[42] Laura Daza, Angela Castillo, María Escobar, Sergio Valencia, Bibiana Pinzón, and Pablo Arbeláez. LUCAS: LUng CAncer screening with multimodal biomarkers. In *Multimodal Learning for Clinical Decision Support and Clinical Image-Based Procedures*, pages 115–124. Springer International Publishing, New York, 2020.

[43] Hans-Peter de Ruiter, Joan Liaschenko, and Jan Angus. Problems with the electronic health record. *Nursing Philosophy*, 17(1):49–58, November 2015.

[44] Pierre Delanaye and Hans Pottel. Estimating glomerular filtration rate in african american individuals. *JAMA Internal Medicine*, 180(11):1549, November 2020.

[45] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, June 2009.

[46] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *North American Chapter of the Association for Computational Linguistics*, 2019.

[47] Federico Di Mattia, Paolo Galeone, Michele De Simoni, and Emanuele Ghelfi. A Survey on GANs for Anomaly Detection. *arXiv e-prints*, page arXiv:1906.11632, June 2019.

[48] Daqiang Dong, Guanghui Fu, Jianqiang Li, Yan Pei, and Yueda Chen. An unsupervised domain adaptation brain CT segmentation method across image modalities and diseases. *Expert Systems with Applications*, 207:118016, November 2022.

[49] Michele Donini, Joao M. Monteiro, Massimiliano Pontil, John Shawe-Taylor, and Janaina Mourao-Miranda. A multimodal multiple kernel learning approach to alzheimer's disease detection. In *2016 IEEE 26th International Workshop on Machine Learning for Signal Processing (MLSP)*, New York, September 2016. IEEE.

[50] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021.

[51] Loïc Duron, Daniel Balvay, Saskia Vande Perre, Afef Bouchouicha, Julien Savatovsky, Jean-Claude Sadik, Isabelle Thomassin-Naggara, Laure Fournier, and Augustin Lecler. Gray-level discretization impacts reproducible mri radiomics texture features. *PLOS ONE*, 14(3):e0213459, March 2019.

[52] Joann G. Elmore, Carolyn K. Wells, Carol H. Lee, Debra H. Howard, and Alvan R. Feinstein. Variability in radiologists' interpretations of mammograms. *New England Journal of Medicine*, 331(22):1493–1499, December 1994.

[53] R. Scott Evans, Stanley L. Pestotnik, David C. Classen, Terry P. Clemmer, Lindell K. Weaver, James F. Orme, James F. Lloyd, and John P. Burke. A computer-assisted management program for antibiotics and other antiinfective agents. *New England Journal of Medicine*, 338(4):232–238, January 1998.

[54] Tao Fang, Yu Qi, and Gang Pan. Reconstructing perceptive images from brain activity by shape-semantic gan. In *Proceedings of the 34th International Conference on*

Neural Information Processing Systems, NIPS'20, Red Hook, NY, USA, 2020. Curran Associates Inc.

[55] US Food and Drug Administration. Clinical decision support software: Guidance for industry and food and drug administration staff, Sept 2022.

[56] Charles P. Friedman, Guido G. Gatti, Timothy M. Franz, Gwendolyn C. Murphy, Fredric M. Wolf, Paul S. Heckerling, Paul L. Fine, Thomas M. Miller, and Arthur S. Elstein. Do physicians know when their diagnoses are correct?: Implications for decision support and error reduction. *Journal of General Internal Medicine*, 20(4):334–339, April 2005.

[57] Chamuditha Jayanga Galappaththige, Gayal Kuruppu, and Muhammad Haris Khan. Generalizing to unseen domains in diabetic retinopathy classification. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 7685–7695, January 2024.

[58] Mohsen Ghafoorian, Alireza Mehrtash, Tina Kapur, Nico Karssemeijer, Elena Marchiori, Mehran Pesteie, Charles R. G. Guttmann, Frank-Erik de Leeuw, Clare M. Tempany, Bram van Ginneken, Andriy Fedorov, Purang Abolmaesumi, Bram Platel, and William M. Wells. Transfer learning for domain adaptation in mri: Application in brain lesion segmentation. In Maxime Descoteaux, Lena Maier-Hein, Alfred Franz, Pierre Jannin, D. Louis Collins, and Simon Duchesne, editors, *Medical Image Computing and Computer Assisted Intervention - MICCAI 2017*, pages 516–524, New York, 2017. Springer International Publishing.

[59] Isabella Gollini, Binbin Lu, Martin Charlton, Christopher Brunsdon, and Paul Harris. Gwmodel: An r package for exploring spatial heterogeneity using geographically weighted models. *Journal of Statistical Software*, 63(17), 2015.

[60] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *Communications of the ACM*, 63(11):139–144, October 2020.

[61] Aparna Gopalakrishnan, Jameel Rizwana Hussaindeen, Viswanathan Sivaraman, Meenakshi Swaminathan, Yee Ling Wong, James A. Armitage, Alex Gentle, and Simon Backhouse. Myopia and its association with near work, outdoor time, and housing type among schoolchildren in south india. *Optometry and Vision Science*, 100(1):105–110, December 2022.

[62] Nicolas Guibert, Julien Mazieres, Myriam Delaunay, Anne Casanova, Magali Farella, Laura Keller, Gilles Favre, and Anne Pradines. Monitoring of KRAS-mutated ctDNA to discriminate pseudo-progression from true progression during anti-PD-1 treatment of lung adenocarcinoma. *Oncotarget*, 8(23):38056–38060, June 2017.

[63] Pengfei Guo, Puyang Wang, Jinyuan Zhou, Vishal M. Patel, and Shanshan Jiang. Lesion mask-based simultaneous synthesis of anatomic and molecular MR images using

a GAN. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2020*, pages 104–113. Springer International Publishing, New York, 2020.

[64] Rohit Gupta, Anurag Sharma, and Anupam Kumar. Super-resolution using GANs for medical imaging. *Procedia Computer Science*, 173:28–35, 2020.

[65] Gavriel Habib, Nahum Kiryati, Miri Sklair-Levy, Anat Shalmon, Osnat Halshtok Neiman, Renata Faermann Weidenfeld, Yael Yagil, Eli Konen, and Arnaldo Mayer. Automatic breast lesion classification by joint neural analysis of mammography and ultrasound. In *Multimodal Learning for Clinical Decision Support and Clinical Image-Based Procedures*, pages 125–135. Springer International Publishing, New York, 2020.

[66] Md. Akmal Haidar and Mehdi Rezagholizadeh. Textkd-gan: Text generation using knowledge distillation and generative adversarial networks. In Marie-Jean Meurs and Frank Rudzicz, editors, *Advances in Artificial Intelligence*, pages 107–118, New York, 2019. Springer International Publishing.

[67] Wolfgang Haigis. Intraocular lens calculation after refractive surgery for myopia: Haigis-l formula. *Journal of Cataract and Refractive Surgery*, 34(10):1658–1663, October 2008.

[68] Mohammad Hamghalam, Alejandro F. Frangi, Baiying Lei, and Amber L. Simpson. Modality completion via gaussian process prior variational autoencoders for multi-modal glioma segmentation. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2021*, pages 442–452. Springer International Publishing, New York, 2021.

[69] A. Han, A. Isaacson, and P. Muennig. The promise of big data for precision population health management in the US. *Public Health*, 185:110–116, August 2020.

[70] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, June 2016.

[71] Patrick A. Hein, Clifford J. Eskey, Jeffrey F. Dunn, and Eugen B. Hug. Diffusion-weighted imaging in the follow-up of treated high-grade gliomas: Tumor recurrence versus radiation injuary. *American Journal of Neuroradiology*, 25(12):201–209, February 2004.

[72] Danliang Ho, Iain Bee Huat Tan, and Mehul Motani. Predictive models for colorectal cancer recurrence using multi-modal healthcare data. In *Proceedings of the Conference on Health, Inference, and Learning*, New York, April 2021. ACM.

[73] Raymond D. Hobbs, Zeina Habib, Dalal Alromaihi, Leila Idi, Nayana Parikh, Frank Blocki, and D. Sudhaker Rao. Severe vitamin d deficiency in arab-american women living in dearborn, michigan. *Endocrine Practice*, 15(1):35–40, January 2009.

[74] Kenneth J. Hoffer. The hoffer q formula: A comparison of theoretic and regression formulas. *Journal of Cataract and Refractive Surgery*, 19(6):700–712, November 1993.

[75] Kenneth J. Hoffer and Giacomo Savini. Update on intraocular lens power calculation study protocols. *Ophthalmology*, 128(11):e115–e120, November 2021.

[76] Minhao Hu, Matthis Maillard, Ya Zhang, Tommaso Ciceri, Giammarco La Barbera, Isabelle Bloch, and Pietro Gori. Knowledge distillation from multi-modal to mono-modal segmentation networks. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2020*, pages 772–781. Springer International Publishing, New York, 2020.

[77] Minhao Hu, Tao Song, Yujun Gu, Xiangde Luo, Jieneng Chen, Yinan Chen, Ya Zhang, and Shaoting Zhang. Fully test-time adaptation for image segmentation. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2021*, pages 251–260. Springer International Publishing, New York, 2021.

[78] Shengye Hu, Yanyan Shen, Shuqiang Wang, and Baiying Lei. Brain MR to PET synthesis via bidirectional generative adversarial network. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2020*, pages 698–707. Springer International Publishing, New York, 2020.

[79] Zhizhong Huang, Shouzhen Chen, Junping Zhang, and Hongming Shan. Pfa-gan: Progressive face aging with generative adversarial network. *IEEE Transactions on Information Forensics and Security*, 16:2031–2045, 2021.

[80] Benjamin Q. Huynh, Hui Li, and Maryellen L. Giger. Digital mammographic tumor classification using transfer learning from deep convolutional neural networks. *Journal of Medical Imaging*, 3(3):034501, August 2016.

[81] Jeremy A. Irvin, Pranav Rajpurkar, Michael Ko, Yifan Yu, Silviana Ciurea-Ilcus, Chris Chute, Henrik Marklund, Behzad Haghgoo, Robyn L. Ball, Katie S. Shpanskaya, Jayne Seekins, David Andrew Mong, Safwan S. Halabi, Jesse K. Sandberg, Ricky Jones, David B. Larson, C. Langlotz, Bhavik N. Patel, Matthew P. Lungren, and A. Ng. Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In *AAAI Conference on Artificial Intelligence*, 2019.

[82] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A. Efros. Image-to-image translation with conditional adversarial networks. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, July 2017.

[83] Mina Jafari, Susan Francis, Jonathan M. Garibaldi, and Xin Chen. LMISA: a lightweight multi-modality image segmentation network via domain adaptation using gradient magnitude and shape constraint. *Medical Image Analysis*, 81:102536, October 2022.

[84] Yinghong Ji, Lei Cai, Tianyu Zheng, Hongfei Ye, Xianfang Rong, Jun Rao, and Yi Lu. The mechanism of uvb irradiation induced-apoptosis in cataract. *Molecular and Cellular Biochemistry*, 401(1–2):87–95, December 2014.

[85] Jue Jiang and Harini Veeraraghavan. Unified cross-modality feature disentangler for unsupervised multi-domain MRI abdomen organs segmentation. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2020*, pages 347–358. Springer International Publishing, New York, 2020.

[86] R.I. John and P.R. Innocent. Modeling uncertainty in clinical diagnosis using fuzzy logic. *IEEE Transactions on Systems, Man and Cybernetics, Part B (Cybernetics)*, 35(6):1340–1350, December 2005.

[87] Alistair Johnson, Tom Pollard, and Roger Mark. Mimic-iii clinical database demo, 2019.

[88] Alistair E.W. Johnson, Tom J. Pollard, Lu Shen, Li wei H. Lehman, 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(1), May 2016.

[89] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *Computer Vision – ECCV 2016*, pages 694–711. Springer International Publishing, New York, 2016.

[90] Sanil Joseph, Tiruvengada Krishnan, Ravilla D. Ravindran, Giovanni Maraini, Monica Camparini, Usha Chakravarthy, Thulasiraj D. Ravilla, Andrew Hutchings, and Astrid E. Fletcher. Prevalence and risk factors for myopia and other refractive errors in an adult population in southern india. *Ophthalmic and Physiological Optics*, 38(3):346–358, March 2018.

[91] Bardia Khosravi, Pouria Rouzrokh, Hilal Maradit Kremers, Dirk R. Larson, Quinn J. Johnson, Shahriar Faghani, Walter K. Kremers, Bradley J. Erickson, Rafael J. Sierra, Michael J. Taunton, and Cody C. Wyles. Patient-specific hip arthroplasty dislocation risk calculator: An explainable multimodal machine learning–based approach. *Radiology: Artificial Intelligence*, 4(6), November 2022.

[92] D.H. Kim and T. MacKinnon. Artificial intelligence in fracture detection: transfer learning from deep convolutional neural networks. *Clinical Radiology*, 73(5):439–445, May 2018.

[93] Thomas N. Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. In *International Conference on Learning Representations*, 2017.

[94] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 25, Red Hook, NY, 2012. Curran Associates, Inc.

[95] Tim J Kruser, Minesh P Mehta, and H Ian Robins. Pseudoprogression after glioma therapy: a comprehensive review. *Expert Review of Neurotherapeutics*, 13(4):389–403, April 2013.

[96] Yongchan Kwon, Joong-Ho Won, Beom Joon Kim, and Myunghee Cho Paik. Uncertainty quantification using bayesian neural networks in classification: Application to biomedical image segmentation. *Computational Statistics &Data Analysis*, 142:106816, February 2020.

[97] Dennis Lam, Srinivas K. Rao, Vineet Ratra, Yizhi Liu, Paul Mitchell, Jonathan King, Marie-José Tassignon, Jost Jonas, Chi P. Pang, and David F. Chang. Cataract. *Nature Reviews Disease Primers*, 1(1), June 2015.

[98] John Lambert, Ozan Sener, and Silvio Savarese. Deep learning under privileged information using heteroscedastic dropout. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[99] Christian Ledig, Lucas Theis, Ferenc Huszár, Jose Caballero, Andrew Cunningham, Alejandro Acosta, Andrew Aitken, Alykhan Tejani, Johannes Totz, Zehan Wang, and Wenzhe Shi. Photo-realistic single image super-resolution using a generative adversarial network. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 105–114, 2017.

[100] Joonsang Lee, Nicholas Wang, Sevcan Turk, Shariq Mohammed, Remy Lobo, John Kim, Eric Liao, Sandra Camelo-Piragua, Michelle Kim, Larry Junck, Jayapalli Bapuraj, Ashok Srinivasan, and Arvind Rao. Discriminating pseudoprogression and true progression in diffuse infiltrating glioma using multi-parametric MRI data through deep learning. *Scientific Reports*, 10(1), November 2020.

[101] Christian Leibig, Vaneeda Allken, Murat Seçkin Ayhan, Philipp Berens, and Siegfried Wahl. Leveraging uncertainty information from deep neural networks for disease detection. *Scientific Reports*, 7(1), December 2017.

[102] Amaury Leroy, Alexandre Cafaro, Grégoire Gessain, Anne Champagnac, Vincent Grégoire, Eric Deutsch, Vincent Lepetit, and Nikos Paragios. StructuRegNet: Structure-guided multimodal 2d-3d registration. In *Lecture Notes in Computer Science*, pages 771–780. Springer Nature Switzerland, Basel, 2023.

[103] Chuan Li and Michael Wand. Precomputed real-time texture synthesis with markovian generative adversarial networks. In *Computer Vision – ECCV 2016*, pages 702–716. Springer International Publishing, New York, 2016.

[104] Dan Li, Dacheng Chen, Baihong Jin, Lei Shi, Jonathan Goh, and See-Kiong Ng. MAD-GAN: Multivariate anomaly detection for time series data with generative adversarial networks. In *Artificial Neural Networks and Machine Learning – ICANN 2019: Text and Time Series*, pages 703–716. Springer International Publishing, New York, 2019.

[105] Thomas Z. Li, John M. Still, Kaiwen Xu, Ho Hin Lee, Leon Y. Cai, Aravind R. Krishnan, Riqiang Gao, Mirza S. Khan, Sanja Antic, Michael Kammer, Kim L. Sandler, Fabien Maldonado, Bennett A. Landman, and Thomas A. Lasko. Longitudinal multimodal transformer integrating imaging and latent clinical signatures from routine

EHRs for pulmonary nodule classification. In *Lecture Notes in Computer Science*, pages 649–659. Springer Nature Switzerland, Basel, 2023.

[106] Tingyang Li, Joshua Stein, and Nambi Nallasamy. Evaluation of the nallasamy formula: a stacking ensemble machine learning method for refraction prediction in cataract surgery. *British Journal of Ophthalmology*, 107(8):1066–1071, April 2022.

[107] Tingyang Li, Joshua D. Stein, and Nambi Nallasamy. Maepi and cir: New metrics for robust evaluation of the prediction performance of ai-based iol formulas. *Translational Vision Science & Technology*, 12(3):29, March 2023.

[108] Yan Li, Yiqi Ma, Zijun Wu, Ruoxi Xie, Fanxin Zeng, Huawei Cai, Su Lui, Bin Song, Lei Chen, and Min Wu. Advanced imaging techniques for differentiating pseudoprogression and tumor recurrence after immunotherapy for glioblastoma. *Frontiers in Immunology*, 12, November 2021.

[109] Wentong Liao, Kai Hu, Michael Ying Yang, and Bodo Rosenhahn. Text to image generation with semantic-spatial aware GAN. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, June 2022.

[110] Weiming Lin, Tong Tong, Qinquan Gao, Di Guo, Xiaofeng Du, Yonggui Yang, Gang Guo, Min Xiao, Min Du, Xiaobo Qu, and Alzheimer's Disease Neuroimaging Initiative. Convolutional neural networks-based MRI image analysis for the alzheimer's disease prediction from mild cognitive impairment. *Front. Neurosci.*, 12:777, November 2018.

[111] Hanwen Liu, Pablo Navarrete Michelini, and Dan Zhu. Artsy-GAN: A style transfer system with improved quality, diversity and performance. In *2018 24th International Conference on Pattern Recognition (ICPR)*. IEEE, August 2018.

[112] Zecheng Liu, Jia Wei, Rui Li, and Jianlong Zhou. SFusion: Self-attention based n-to-one multimodal fusion block. In *Lecture Notes in Computer Science*, pages 159–169. Springer Nature Switzerland, Basel, 2023.

[113] Ming Y. Lu, Drew F. K. Williamson, Tiffany Y. Chen, Richard J. Chen, Matteo Barbieri, and Faisal Mahmood. Data-efficient and weakly supervised computational pathology on whole-slide images. *Nature Biomedical Engineering*, 5(6):555–570, March 2021.

[114] Ming Y. Lu, Drew F. K. Williamson, Tiffany Y. Chen, Richard J. Chen, Matteo Barbieri, and Faisal Mahmood. Data-efficient and weakly supervised computational pathology on whole-slide images. *Nature Biomedical Engineering*, 5(6):555–570, March 2021.

[115] Steven A. Lubitz, Anthony Z. Faranesh, Caitlin Selvaggi, Steven J. Atlas, David D. McManus, Daniel E. Singer, Sherry Pagoto, Michael V. McConnell, Alexandros Pantelopoulos, and Andrea S. Foulkes. Detection of atrial fibrillation in a large population using wearable devices: The Fitbit heart study. *Circulation*, 146(19):1415–1424, November 2022.

[116] Junyan Lyu, Yiqi Zhang, Yijin Huang, Li Lin, Pujin Cheng, and Xiaoying Tang. Aadg: Automatic augmentation for domain generalization on retinal image segmentation. *IEEE Transactions on Medical Imaging*, 41(12):3699–3711, 2022.

[117] Yiming Ma, Qiwei Wang, Qian Dong, Lei Zhan, and Jingdong Zhang. How to differentiate pseudoprogression from true progression in cancer patients treated with immunotherapy. *Am J Cancer Res*, 9(8), August 2019.

[118] Dwarikanath Mahapatra, Behzad Bozorgtabar, Sajini Hewavitharanage, and Rahil Garnavi. Image super resolution using generative adversarial networks and local saliency maps for retinal image analysis. In *Medical Image Computing and Computer Assisted Intervention - MICCAI 2017*, pages 382–390. Springer International Publishing, New York, 2017.

[119] Huanru Henry Mao. A survey on self-supervised pre-training for sequential transfer learning in neural networks, 2020.

[120] Jiaqi Meng, Ling Wei, Wenwen He, Jiao Qi, Yi Lu, and Xiangjia Zhu. Lens thickness and associated ocular biometric factors among cataract patients in shanghai. *Eye and Vision*, 8(1), May 2021.

[121] Bjoern H. Menze, Andras Jakab, Stefan Bauer, Jayashree Kalpathy-Cramer, Keyvan Farahani, Justin Kirby, Yuliya Burren, Nicole Porz, Johannes Slotboom, Roland Wiest, Levente Lanczi, Elizabeth Gerstner, Marc-Andre Weber, Tal Arbel, Brian B. Avants, Nicholas Ayache, Patricia Buendia, D. Louis Collins, Nicolas Cordier, Jason J. Corso, Antonio Criminisi, Tilak Das, Herve Delingette, Cagatay Demiralp, Christopher R. Durst, Michel Dojat, Senan Doyle, Joana Festa, Florence Forbes, Ezequiel Geremia, Ben Glocker, Polina Golland, Xiaotao Guo, Andac Hamamci, Khan M. Iftekharuddin, Raj Jena, Nigel M. John, Ender Konukoglu, Danial Lashkari, Jose Antonio Mariz, Raphael Meier, Sergio Pereira, Doina Precup, Stephen J. Price, Tammy Riklin Raviv, Syed M. S. Reza, Michael Ryan, Duygu Sarikaya, Lawrence Schwartz, Hoo-Chang Shin, Jamie Shotton, Carlos A. Silva, Nuno Sousa, Nagesh K. Subbanna, Gabor Szekely, Thomas J. Taylor, Owen M. Thomas, Nicholas J. Tustison, Gozde Unal, Flor Vasseur, Max Wintermark, Dong Hye Ye, Liang Zhao, Binsheng Zhao, Darko Zikic, Marcel Prastawa, Mauricio Reyes, and Koen Van Leemput. The multimodal brain tumor image segmentation benchmark (BRATS). *IEEE Transactions on Medical Imaging*, 34(10):1993–2024, October 2015.

[122] Rachel Metz. Ai won an art contest, and artists are furious. *CNN Business*, Sep 2022.

[123] S. Mohammed, V. Ravikumar, E. Warner, S.H. Patel, S. Bakas, A. Rao, and R. Jain. Quantifying t2-FLAIR mismatch using geographically weighted regression and predicting molecular status in lower-grade gliomas. *American Journal of Neuroradiology*, 43(1):33–39, November 2021.

[124] S Mohammed, V Ravikumar, E Warner, SH Patel, S Bakas, A Rao, and R Jain. Quantifying t2-flair mismatch using geographically weighted regression and predict-

ing molecular status in lower-grade gliomas. *American Journal of Neuroradiology*, 43(1):33–39, 2022.

[125] Shariq Mohammed, Tingyang Li, Xing D Chen, Elisa Warner, Anand Shankar, Maria Fernanda Abalem, Thiran Jayasundera, Thomas W Gardner, and Arvind Rao. Density-based classification in diabetic retinopathy through thickness of retinal layers from optical coherence tomography. *Scientific reports*, 10(1):15937, 2020.

[126] Federico Monti, Fabrizio Frasca, Davide Eynard, Damon Mannion, and Michael M. Bronstein. Fake news detection on social media using geometric deep learning, 2019.

[127] Karel G.M. Moons, Douglas G. Altman, Johannes B. Reitsma, John P.A. Ioannidis, Petra Macaskill, Ewout W. Steyerberg, Andrew J. Vickers, David F. Ransohoff, and Gary S. Collins. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (tripod): Explanation and elaboration. *Annals of Internal Medicine*, 162(1):W1–W73, January 2015.

[128] Mehdi Moradi, Tanveer Syeda-Mahmood, and Soheil Hor. Tree-based transforms for privileged learning. In *Machine Learning in Medical Imaging*, pages 188–195. Springer International Publishing, New York, 2016.

[129] Majid Moshirfar, Michael V McCaughey, and Luis Santiago-Caban. Corrective techniques and future directions for treatment of residual refractive error following cataract surgery. *Expert Review of Ophthalmology*, 9(6):529–537, October 2014.

[130] ME Mullins, GD Barest, PW Schaefer, FH Hochberg, RG Gonzalez, and MH Lev. Radiation necrosis versus glioma recurrence: conventional mr imaging clues to diagnosis. *Neurology*, 26(8):1967–1972, September 2005.

[131] Kristian Naeser, Jannik Boberg-Ans, and Ralph Bargum. Biometry of the posterior lens capsule: A new method to predict pseudophakic anterior chamber depth. *Journal of Cataract and Refractive Surgery*, 16(2):202–206, March 1990.

[132] Theresa Neubauer, Maria Wimmer, Astrid Berg, David Major, Dimitrios Lenis, Thomas Beyer, Jelena Saponjski, and Katja Bühler. Soft tissue sarcoma cosegmentation in combined MRI and PET/CT data. In *Multimodal Learning for Clinical Decision Support and Clinical Image-Based Procedures*, pages 97–105. Springer International Publishing, New York, 2020.

[133] Long D. Nguyen, Ruihan Gao, Dongyun Lin, and Zhiping Lin. Biomedical image classification based on a feature concatenation and ensemble of deep CNNs. *Journal of Ambient Intelligence and Humanized Computing*, March 2019.

[134] Luke Oakden-Rayner and Lyle John Palmer. *Artificial Intelligence in Medicine: Validation and Study Design*, page 83–104. Springer International Publishing, 2019.

[135] Sangyoon Oh, Min Su Lee, and Byoung-Tak Zhang. Ensemble learning with active example selection for imbalanced biomedical data classification. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 8(2):316–325, March 2011.

[136] Thomas Olsen and Peter Hoffmann. C constant: New concept for ray tracing–assisted intraocular lens power calculation. *Journal of Cataract and Refractive Surgery*, 40(5):764–773, May 2014.

[137] Thomas Olsen, Kirsten Thim, and Leif Corydon. Theoretical versus srk i and srk ii calculation of intraocular lens power. *Journal of Cataract and Refractive Surgery*, 16(2):217–225, March 1990.

[138] Miki Kamikawatoko Omoto, Hidemasa Torii, Ken Hayashi, Masahiko Ayaki, Kazuo Tsubota, and Kazuno Negishi. Ratio of axial length to corneal radius in japanese patients and accuracy of intraocular lens power calculation based on biometric data. *American Journal of Ophthalmology*, 218:320–329, October 2020.

[139] OpenAI. Gpt-4 technical report, 2023.

[140] Jonas Oppenlaender. The creativity of text-to-image generation. In *Proceedings of the 25th International Academic Mindtrek Conference*, New York, November 2022. ACM.

[141] Marisa Palazzo, Francesco Vizzarri, Lubomir Ondruška, Michele Rinaldi, Luigi Pacente, Germano Guerra, Francesco Merolla, Ciro Caruso, and Ciro Costagliola. Corneal uv protective effects of a topical antioxidant formulation: A pilot study on in vivo rabbits. *International Journal of Molecular Sciences*, 21(15):5426, July 2020.

[142] Sveinn Palsson, Eirikur Agustsson, Radu Timofte, and Luc Van Gool. Generative adversarial style transfer networks for face aging. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2018.

[143] Sang Jun Park, Ju Hyun Lee, Se Woong Kang, Joon Young Hyon, and Kyu Hyung Park. Cataract and cataract surgery: Nationwide prevalence and clinical determinants. *Journal of Korean Medical Science*, 31(6):963, 2016.

[144] Niki J. Parmar, Ashish Vaswani, Jakob Uszkoreit, Lukasz Kaiser, Noam Shazeer, Alexander Ku, and Dustin Tran. Image transformer. In *International Conference on Machine Learning (ICML)*, 2018.

[145] Puskar Pattanayak, Evrim B. Turkbey, and Ronald M. Summers. Comparative evaluation of three software packages for liver and spleen segmentation and volumetry. *Academic Radiology*, 24(7):831–839, July 2017.

[146] Chenhao Pei, Fuping Wu, Mingjing Yang, Lin Pan, Wangbin Ding, Jinwei Dong, Liqin Huang, and Xiahai Zhuang. Multi-source domain adaptation for medical image segmentation. *IEEE Transactions on Medical Imaging*, page 1–1, 2023.

[147] Alexander C. Perino, Santosh E. Gummidipundi, Justin Lee, Haley Hedlin, Ariadna Garcia, Todd Ferris, Vidhya Balasubramanian, Rebecca M. Gardner, Lauren Cheung, Grace Hung, Christopher B. Granger, Peter Kowey, John S. Rumsfeld, Andrea M. Russo, Mellanie True Hills, Nisha Talati, Divya Nag, David Tsay, Sumbul Desai, Manisha Desai, Kenneth W. Mahaffey, Mintu P. Turakhia, and Marco V. Perez. Arrhythmias other than atrial fibrillation in those with an irregular pulse detected with a

smartwatch: Findings from the Apple heart study. *Circulation: Arrhythmia and Electrophysiology*, 14(10), October 2021.

[148] Esteban Piacentino, Alvaro Guarner, and Cecilio Angulo. Generating synthetic ECGs using GANs for anonymizing healthcare data. *Electronics*, 10(4):389, February 2021.

[149] Tingting Qiao, Jing Zhang, Duanqing Xu, and Dacheng Tao. Mirrorgan: Learning text-to-image generation by redescription. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[150] Jiezhong Qiu, Jian Tang, Hao Ma, Yuxiao Dong, Kuansan Wang, and Jie Tang. Deepinf. *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, Jul 2018.

[151] Adalberto Claudio Quiros, Roderick Murray-Smith, and Ke Yuan. Pathologygan: Learning deep representations of cancer tissue. In Tal Arbel, Ismail Ben Ayed, Marleen de Bruijne, Maxime Descoteaux, Herve Lombaert, and Christopher Pal, editors, *Proceedings of the Third Conference on Medical Imaging with Deep Learning*, volume 121 of *Proceedings of Machine Learning Research*, pages 669–695. PMLR, 06–08 Jul 2020.

[152] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763, Maastricht, 18–24 Jul 2021. PMLR.

[153] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. 2022.

[154] Carolyn Rando and Tony Waldron. TMJ osteoarthritis: A new approach to diagnosis. *American Journal of Physical Anthropology*, 148(1):45–53, February 2012.

[155] Anoop Rao and Jonathan Palma. Clinical decision support in the neonatal ICU. *Seminars in Fetal and Neonatal Medicine*, page 101332, March 2022.

[156] Scott A. Read, Michael J. Collins, and Stephen J. Vincent. Light exposure and eye growth in childhood. *Investigative Opthalmology & Visual Science*, 56(11):6779, October 2015.

[157] Narathip Reamaroon, Michael W. Sjoding, Kaiwen Lin, Theodore J. Iwashyna, and Kayvan Najarian. Accounting for label uncertainty in machine learning for detection of acute respiratory distress syndrome. *IEEE Journal of Biomedical and Health Informatics*, 23(1):407–415, January 2019.

[158] Sandeep Reddy. Explainability and artificial intelligence in medicine. *The Lancet Digital Health*, 4(4):e214–e215, April 2022.

[159] Scott Reed, Zeynep Akata, Xinchen Yan, Lajanugen Logeswaran, Bernt Schiele, and Honglak Lee. Generative adversarial text to image synthesis. In Maria Florina Balcan and Kilian Q. Weinberger, editors, *Proceedings of The 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pages 1060–1069, New York, New York, USA, 20–22 Jun 2016. PMLR.

[160] John A Retzlaff and etc. *Lens implant power calculation*. SLACK, Thorofare, 3 edition, December 1990.

[161] John A. Retzlaff, Donald R. Sanders, and Manus C. Kraff. Development of the srk/t intraocular lens implant power calculation formula. *Journal of Cataract and Refractive Surgery*, 16(3):333–340, May 1990.

[162] Yair Rivenson, Hongda Wang, Zhensong Wei, Kevin de Haan, Yibo Zhang, Yichen Wu, Harun Günaydın, Jonathan E. Zuckerman, Thomas Chong, Anthony E. Sisk, Lindsey M. Westbrook, W. Dean Wallace, and Aydogan Ozcan. Virtual histological staining of unlabelled tissue-autofluorescence images via deep learning. *Nature Biomedical Engineering*, 3(6):466–477, March 2019.

[163] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Lecture Notes in Computer Science*, pages 234–241. Springer International Publishing, New York, 2015.

[164] Jeffrey D. Rudie, Evan Calabrese, Rachit Saluja, David Weiss, John B. Colby, Soonmee Cha, Christopher P. Hess, Andreas M. Rauschecker, Leo P. Sugrue, and Javier E. Villanueva-Meyer. Longitudinal assessment of posttreatment diffuse glioma tissue volumes with three-dimensional convolutional neural networks. *Radiology: Artificial Intelligence*, 4(5), September 2022.

[165] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)*, 115(3):211–252, 2015.

[166] Stuart Russell and Peter Norvig. *Artificial intelligence*. Pearson, Upper Saddle River, NJ, 3 edition, December 2009.

[167] Elyas Sabeti, Joshua Drews, Narathip Reamaroon, Elisa Warner, Michael W. Sjoding, Jonathan Gryak, and Kayvan Najarian. Learning using partially available privileged information and label uncertainty: Application in detection of acute respiratory distress syndrome. *IEEE Journal of Biomedical and Health Informatics*, 25(3):784–796, March 2021.

[168] Nathan Sanders. A balanced perspective on prediction and inference for data science in industry. *Harvard Data Science Review*, June 2019.

[169] Giacomo Savini, Kenneth J. Hoffer, and Piero Barboni. Influence of corneal asphericity on the refractive outcome of intraocular lens implantation in cataract surgery. *Journal of Cataract and Refractive Surgery*, 41(4):785–789, April 2015.

[170] Eric Schiffman, Richard Ohrbach, Edmond Truelove, and et al. Diagnostic criteria for temporomandibular disorders (DC/TMD) for clinical and research applications: Recommendations of the international RDC/TMD consortium network and orofacial pain special interest group. *Journal of Oral & Facial Pain and Headache*, 28(1):6–27, January 2014.

[171] Steven J. Scrivani, David A. Keith, and Leonard B. Kaban. Temporomandibular disorders. *New England Journal of Medicine*, 359(25):2693–2705, December 2008.

[172] Tawseef Ayoub Shaikh, Rashid Ali, and M. M. Sufyan Beg. Transfer learning privileged information fuels CAD diagnosis of breast cancer. *Machine Vision and Applications*, 31(1-2), February 2020.

[173] H J Shammas. *Intraocular Lens Power Calculations*. SLACK, Thorofare, November 2003.

[174] Hoo-Chang Shin, , Alvin Ihsani, Ziyue Xu, Swetha Mandava, Sharath Turuvekere Sreenivas, Christopher Forster, and Jiook Cha. GANDALF: Generative adversarial networks with discriminator-adaptive loss fine-tuning for alzheimer's disease diagnosis from MRI. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2020*, pages 688–697. Springer International Publishing, New York, 2020.

[175] Christina Silcox, Susan Dentzer, and David W. Bates. AI-enabled clinical decision support software: A "trust and value checklist" for clinicians. *NEJM Catalyst*, 1(6), November 2020.

[176] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *International Conference on Learning Representations*, 2015.

[177] Junho Song, Young Jun Chai, Hiroo Masuoka, Sun-Won Park, Su jin Kim, June Young Choi, Hyoun-Joong Kong, Kyu Eun Lee, Joongseek Lee, Nojun Kwak, Ka Hee Yi, and Akira Miyauchi. Ultrasound image analysis using deep learning algorithm for the diagnosis of thyroid nodules. *Medicine*, 98(15):e15133, April 2019.

[178] Nitish Srivastava and Ruslan Salakhutdinov. Multimodal learning with deep boltzmann machines. *Journal of Machine Learning Research*, 15(84):2949–2980, 2014.

[179] Statista. Distribution of u.s. medical practices by size in 2018, 2018.

[180] Statista. Number of health center sites in the u.s. from 2010 to 2020, 2022.

[181] Jonathan Stubblefield, Mitchell Hervert, Jason L. Causey, Jake A. Qualls, Wei Dong, Lingrui Cai, Jennifer Fowler, Emily Bellis, Karl Walker, Jason H. Moore, Sara Nehring, and Xiuzhen Huang. Transfer learning with chest x-rays for ER patient classification. *Scientific Reports*, 10(1), December 2020.

[182] K. S. Sundar, K. Rajamani, and S. Sai. Exploring image classification of thyroid ultrasound images using deep learning. 2018.

[183] Mujeen Sung, Jinhyuk Lee, Sean S. Yi, Minji Jeon, Sungdong Kim, and Jaewoo Kang. Can language models be biomedical knowledge bases? In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, volume abs/2109.07154 of *Proceedings of Empirical Methods in Natural Language Processing*, pages 4723–4734. Association for Computational Linguistics, Stroudsburg, PA, November 2021.

[184] Reed T. Sutton, David Pincock, Daniel C. Baumgart, Daniel C. Sadowski, Richard N. Fedorak, and Karen I. Kroeker. An overview of clinical decision support systems: benefits, risks, and strategies for success. *npj Digital Medicine*, 3(1), February 2020.

[185] Prashant D. Tailor, Timothy T. Xu, Shreya Tailor, Collin Asheim, and Timothy W. Olsen. Trends in myopia and high myopia from 1966 to 2019 in olmsted county, minnesota. *American Journal of Ophthalmology*, 259:35–44, March 2024.

[186] Yu Takagi and Shinji Nishimoto. High-resolution image reconstruction with latent diffusion models from human brain activity. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, June 2023.

[187] E. Tanaka, M.S. Detamore, and L.G. Mercuri. Degenerative disorders of the temporomandibular joint: Etiology, diagnosis, and treatment. *Journal of Dental Research*, 87(4):296–307, April 2008.

[188] A Tavani. Food and nutrient intake and risk of cataract. *Annals of Epidemiology*, 6(1):41–46, January 1996.

[189] Robert Tinn, Hao Cheng, Yu Gu, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. Fine-tuning large neural language models for biomedical natural language processing. *Patterns*, 4(4):100729, April 2023.

[190] Yao-Hung Hubert Tsai, Shaojie Bai, Paul Pu Liang, J. Zico Kolter, Louis-Philippe Morency, and Ruslan Salakhutdinov. Multimodal transformer for unaligned multimodal language sequences. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, Stroudsberg, PA, 2019. Association for Computational Linguistics.

[191] Huan-Hsin Tseng, Yi Luo, Sunan Cui, Jen-Tzung Chien, Randall K. Ten Haken, and Issam El Naqa. Deep reinforcement learning for automated radiation adaptation in lung cancer. *Medical Physics*, 44(12):6690–6705, November 2017.

[192] Kathryn Tunyasuvunakool, Jonas Adler, Zachary Wu, Tim Green, Michal Zielinski, Augustin Žídek, Alex Bridgland, Andrew Cowie, Clemens Meyer, Agata Laydon, Sameer Velankar, Gerard J. Kleywegt, Alex Bateman, Richard Evans, Alexander Pritzel, Michael Figurnov, Olaf Ronneberger, Russ Bates, Simon A. A. Kohl, Anna Potapenko, Andrew J. Ballard, Bernardino Romera-Paredes, Stanislav Nikolov, Rishub Jain, Ellen Clancy, David Reiman, Stig Petersen, Andrew W. Senior, Koray Kavukcuoglu, Ewan Birney, Pushmeet Kohli, John Jumper, and Demis Hassabis. Highly accurate protein structure prediction for the human proteome. *Nature*, July 2021.

[193] Ashwini Tuppad and Shantala Devi Patil. Machine learning for diabetes clinical decision support: a review. *Advances in Computational Intelligence*, 2(2), April 2022.

[194] United Nationals Environment Program (UNEP). Vital ozone graphics third edition 25th anniversary of the montreal protocol. https://wedocs.unep.org/bitstream/handle/20.500.11822/26558/7771-e-VOGIII_2012.pdf?sequence=1&isAllowed=y. Accessed: 2023-12-18.

[195] Martina Vacalebre, Renato Frison, Carmelo Corsaro, Fortunato Neri, Antonio Santoro, Sabrina Conoci, Elena Anastasi, Maria Cristina Curatolo, and Enza Fazio. Current state of the art and next generation of materials for a customized intraocular lens according to a patient-specific eye power. *Polymers*, 15(6):1590, March 2023.

[196] Martin J. Van den Bent, Michele Reni, Gemma Gatta, and Charles Vecht. Oligodendroglioma. *Critical Reviews in Oncology/Hematology*, 66(3):262–272, June 2008.

[197] Gail A. Van Norman. Drugs, devices, and the fda: Part 2. *JACC: Basic to Translational Science*, 1(4):277–287, June 2016.

[198] Annegreet van Opbroek, M. Arfan Ikram, Meike W. Vernooij, and Marleen de Bruijne. Transfer learning improves supervised image segmentation across imaging protocols. *IEEE Transactions on Medical Imaging*, 34(5):1018–1030, 2015.

[199] Tom van Sonsbeek and Marcel Worring. Towards automated diagnosis with attentive multi-modal learning using electronic health records and chest x-rays. In *Multimodal Learning for Clinical Decision Support and Clinical Image-Based Procedures*, pages 106–114. Springer International Publishing, New York, 2020.

[200] Vladimir Vapnik and Akshay Vashist. A new learning paradigm: Learning using privileged information. *Neural Networks*, 22(5-6):544–557, July 2009.

[201] Thomas Varsavsky, Mauricio Orbes-Arteaga, Carole H. Sudre, Mark S. Graham, Parashkev Nachev, and M. Jorge Cardoso. Test-time unsupervised domain adaptation. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2020*, pages 428–436. Springer International Publishing, New York, 2020.

[202] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. *CoRR*, abs/1706.03762, 2017.

[203] Lingam Vijaya, Ronnie George, Hemamalini Arvind, Satyamangalam Ve Ramesh, Mani Baskaran, Prema Raju, Rashima Asokan, and Lokapavani Velumuri. Central corneal thickness in adult south indians. *Ophthalmology*, 117(4):700–704, April 2010.

[204] Gerome Vivar, Kamilia Mullakaeva, Andreas Zwergal, Nassir Navab, and Seyed-Ahmad Ahmadi. Peri-diagnostic decision support through cost-efficient feature acquisition at test-time. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2020*, pages 572–581. Springer International Publishing, New York, 2020.

[205] Theo Vos, Amanuel Alemu Abajobir, Kalkidan Hassen Abate, and et al. Global, regional, and national incidence, prevalence, and years lived with disability for 328 diseases and injuries for 195 countries, 1990–2016: a systematic analysis for the global burden of disease study 2016. *The Lancet*, 390(10100):1211–1259, September 2017.

[206] Hu Wang, Congbo Ma, Jianpeng Zhang, Yuan Zhang, Jodie Avery, Louise Hull, and Gustavo Carneiro. Learnable cross-modal knowledge distillation for multi-modal learning with missing modality. In *Lecture Notes in Computer Science*, pages 216–226. Springer Nature Switzerland, Basel, 2023.

[207] Wenxuan Wang, Chen Chen, Meng Ding, Hong Yu, Sen Zha, and Jiangyun Li. TransBTS: Multimodal brain tumor segmentation using transformer. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2021*, pages 109–119. Springer International Publishing, New York, 2021.

[208] Zichen Wang, Mu Zhou, and Corey Arnold. Toward heterogeneous information fusion: bipartite graph convolutional networks for in silico drug repurposing. *Bioinformatics*, 36(Supplement_1):i525–i533, July 2020.

[209] Elisa Warner, Najla Al-Turkestani, Jonas Bianchi, Marcela Lima Gurgel, Lucia Cevidanes, and Arvind Rao. Predicting osteoarthritis of the temporomandibular joint using random forest with privileged information. In *Ethical and Philosophical Issues in Medical Imaging, Multimodal Learning and Fusion Across Scales for Clinical Decision Support, and Topological Data Analysis for Biomedical Imaging: 1st International Workshop, EPIMI 2022, 12th International Workshop, ML-CDS 2022, 2nd International Workshop, TDA4BiomedicalImaging, Held in Conjunction with MICCAI 2022, Singapore, September 18–22, 2022, Proceedings*, pages 77–86. Springer Nature Switzerland, Basel, 2022.

[210] Elisa Warner, Joonsang Lee, William Hsu, Tanveer Syeda-Mahmood, Charles Kahn, Olivier Gevaert, and Arvind Rao. Multimodal machine learning in image-based and clinical biomedicine: Survey and prospects, 2023.

[211] Elisa Warner, Joonsang Lee, Santhoshi Krishnan, Nicholas Wang, Shariq Mohammed, Ashok Srinivasan, Jayapalli Bapuraj, and Arvind Rao. Low-parameter supervised learning models can discriminate pseudoprogression and true progression in non-perfusion-based mri. In *2023 45th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*. IEEE, July 2023.

[212] Elisa Warner, Nicholas Wang, Joonsang Lee, and Arvind Rao. *Meaningful incorporation of artificial intelligence for personalized patient management during cancer: Quantitative imaging, risk assessment, and therapeutic outcomes*, page 339–359. Elsevier, 2021.

[213] Hongwei Wen, Yue Liu, Islem Rekik, Shengpei Wang, Zhiqiang Chen, Jishui Zhang, Yue Zhang, Yun Peng, and Huiguang He. Multi-modal multiple kernel learning for accurate identification of tourette syndrome children. *Pattern Recognition*, 63:601–611, March 2017.

[214] World Health Organization. *Ethics and governance of artificial intelligence for health: Guidance on large multi-modal models*. Geneva, 2024.

[215] Le Wu, Peijie Sun, Richang Hong, Yanjie Fu, Xiting Wang, and Meng Wang. Socialgcn: An efficient graph convolutional network based model for social recommendation, 2019.

[216] Xiaohan Xing, Zhen Chen, Meilu Zhu, Yuenan Hou, Zhifan Gao, and Yixuan Yuan. Discrepancy and gradient-guided multi-modal knowledge distillation for pathological glioma grading. In *Lecture Notes in Computer Science*, pages 636–646. Springer Nature Switzerland, Basel, 2022.

[217] Jianhao Xiong, Andre Wang He, Meng Fu, Xinyue Hu, Yifan Zhang, Congxin Liu, Xin Zhao, and Zongyuan Ge. Improve unseen domain generalization via enhanced local color transformation. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2020*, pages 433–443. Springer International Publishing, New York, 2020.

[218] Yingying Xue, Shixiang Feng, Ya Zhang, Xiaoyun Zhang, and Yanfeng Wang. Dual-task self-supervision for cross-modality domain adaptation. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2020*, pages 408–417. Springer International Publishing, New York, 2020.

[219] Rui Yan, Fa Zhang, Xiaosong Rao, Zhilong Lv, Jintao Li, Lingling Zhang, Shuang Liang, Yilin Li, Fei Ren, Chunhou Zheng, and Jun Liang. Richer fusion network for breast cancer classification based on multimodal data. *BMC Medical Informatics and Decision Making*, 21(S1), April 2021.

[220] Jiancheng Yang, Jiajun Chen, Kaiming Kuang, Tiancheng Lin, Junjun He, and Bingbing Ni. MIA-prognosis: A deep learning framework to predict therapy response. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2020*, pages 211–220. Springer International Publishing, New York, 2020.

[221] Xingyi Yang, Xuehai He, Yuxiao Liang, Yue Yang, Shanghang Zhang, and Pengtao Xie. Transfer learning or self-supervised learning? A tale of two pretraining paradigms. *CoRR*, abs/2007.04234, 2020.

[222] Yan Yang, Na Wang, Heran Yang, Jian Sun, and Zongben Xu. Model-driven deep attention network for ultra-fast compressive sensing MRI guided by cross-contrast MR image. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2020*, pages 188–198. Springer International Publishing, New York, 2020.

[223] Ting Yao, Yingwei Pan, Yehao Li, and Tao Mei. Exploring visual relationship for image captioning, 2018.

[224] Rex Ying, Ruining He, Kaifeng Chen, Pong Eksombatchai, William L. Hamilton, and Jure Leskovec. Graph convolutional neural networks for web-scale recommender systems. *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, Jul 2018.

[225] R. J. Young, A. Gupta, A. D. Shah, J. J. Graber, Z. Zhang, W. Shi, A. I. Holodny, and A. M. P. Omuro. Potential utility of conventional MRI signs in diagnosing pseudoprogression in glioblastoma. *Neurology*, 76(22):1918–1924, May 2011.

[226] Daoqiang Zhang, Yaping Wang, Luping Zhou, Hong Yuan, and Dinggang Shen. Multimodal classification of alzheimer's disease and mild cognitive impairment. *NeuroImage*, 55(3):856–867, April 2011.

[227] Fuquan Zhang and Chuansheng Wang. MSGAN: Generative adversarial networks for image seasonal style transfer. *IEEE Access*, 8:104830–104840, 2020.

[228] Lu Zhang, Saiyang Na, Tianming Liu, Dajiang Zhu, and Junzhou Huang. Multimodal deep fusion in hyperbolic space for mild cognitive impairment study. In *Lecture Notes in Computer Science*, pages 674–684. Springer Nature Switzerland, Basel, 2023.

[229] Winston Zhang, Jonas Bianchi, Najla Al Turkestani, and et al. Temporomandibular joint osteoarthritis diagnosis using privileged learning of protein markers. In *2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*. IEEE, November 2021.

[230] Yao Zhang, Nanjun He, Jiawei Yang, Yuexiang Li, Dong Wei, Yawen Huang, Yang Zhang, Zhiqiang He, and Yefeng Zheng. mmFormer: Multimodal medical transformer for incomplete multimodal learning of brain tumor segmentation. In *Lecture Notes in Computer Science*, pages 107–117. Springer Nature Switzerland, New York, 2022.

[231] Yuhao Zhang, Hang Jiang, Yasuhide Miura, Christopher D. Manning, and Curtis P. Langlotz. Contrastive learning of medical visual representations from paired images and text. In *Proceedings of Machine Learning Research*, volume 182 of *Proceedings of Machine Learning Research*, pages 1–24, New York, NY, August 2022. PMLR, Machine Learning for Healthcare.

[232] Tao Zhou, Huazhu Fu, Yu Zhang, Changqing Zhang, Xiankai Lu, Jianbing Shen, and Ling Shao. M2net: Multi-modal multi-channel network for overall survival time prediction of brain tumor patients. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2020*, pages 221–231. Springer International Publishing, New York, 2020.

[233] You Zhou, Gang Yang, Yang Zhou, Dayong Ding, and Jianchun Zhao. Representation, alignment, fusion: A generic transformer-based framework for multi-modal glaucoma recognition. In *Lecture Notes in Computer Science*, pages 704–713. Springer Nature Switzerland, Basel, 2023.

[234] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A. Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *2017 IEEE International Conference on Computer Vision (ICCV)*, New York, oct 2017. IEEE.

[235] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A. Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks, 2017.

[236] Yingying Zhu, Youbao Tang, Yuxing Tang, Daniel C. Elton, Sungwon Lee, Perry J. Pickhardt, and Ronald M. Summers. Cross-domain medical image translation by shared latent gaussian mixture model. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2020*, pages 379–389. Springer International Publishing, New York, 2020.

[237] Marinka Zitnik, Monica Agrawal, and Jure Leskovec. Modeling polypharmacy side effects with graph convolutional networks. *Bioinformatics*, 34(13):i457–i466, Jun 2018.