Envisioning Automated Glaucoma Screening: Domain Generalization for Deep Learning-Based Glaucoma Classification

Hannah Ulman

Advisor: Professor Alex Dytso

Submitted in partial fulfillment

of the requirements for the degree of

Bachelor of Science in Engineering

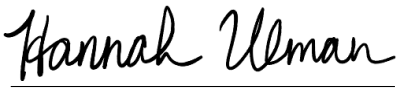
Department of Operations Research and Financial Engineering

Princeton University

April 11, 2024

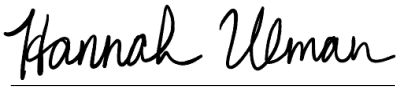
I hereby declare that I am the sole author of this thesis.

I authorize Princeton University to lend this thesis to other institutions or individuals for the purpose of scholarly research.



Hannah Ulman

I further authorize Princeton University to reproduce this thesis by photocopying or by other means, in total or in part, at the request of other institutions or individuals for the purpose of scholarly research.



Hannah Ulman

Abstract

Rapid advancements in deep learning algorithms for computer vision tasks have produced powerful models that can accurately classify diseases from medical images across a variety of specialties. In the field of ophthalmology, these systems have the potential to enable large-scale eye disease screening programs in areas that lack access to vision care or trained specialists. However, deep learning models must be able to generalize to unseen domains, such as retinal images from different healthcare institutions, machines, and patient populations, before they can be deployed for population-wide clinical screening. In this thesis, we analyze domain shift across four public retinal image datasets and investigate its effect on baseline glaucoma classification model performance. We find that domain shift severely degrades classification ability, and, specifically, image features extracted during model training do not generalize to out-of-domain images. We further test three state-of-the-art domain generalization methods and find that one method marginally improves model generalization, though not to an adequate level for clinical use. In the final chapter, we connect the findings of our research to broader themes in global health and health policy. Overall, this thesis begins to fill a noticeable gap in domain generalization research for deep learning-based glaucoma classification.

Acknowledgements

Thank you...

To my advisor, Alex Dytso, for offering invaluable advice and encouragement throughout the research process.

To my parents, for being a constant source of support, wisdom, and unconditional love; and a special shout-out to Dad for proofreading all 39 pages of this thesis.

To my siblings, Abby and Nate, for showing me the ropes and doing their best to keep me humble.

To my friends at Princeton and beyond, for always keeping life fun and exciting throughout the chaos of it all.

Finally, to every professor, TA, staff member, and administrator who has contributed to the success of my Princeton career.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Glaucoma

Glaucoma is a group of progressive eye diseases that cause optic nerve damage and vision loss [43]. A leading cause of irreversible blindness worldwide, glaucoma usually does not present symptoms in its early stages and can only be diagnosed during an eye exam by an ophthalmologist [10]. Although there is currently no cure, treatments for glaucoma can slow or stop its progress [43], so catching the disease in its early stages is crucial.

Unfortunately, many people across the world do not have access to proper vision care due to a lack of comprehensive healthcare coverage and trained ophthalmologists [52, 61]. In the United States, vision care is not typically included in primary care insurance, so people will only visit the eye doctor if they have a vision problem [52], which is often too late to catch glaucoma before it causes irreversible damage. Globally, low- and middle-income countries have the highest burden of vision impairment and blindness due to a lack of access to eye care and prohibitively low ophthalmologist-to-patient ratios [61]. Thus, wide-scale glaucoma screening programs that do not require trained ophthalmologists would provide underserved populations

with access to glaucoma diagnosis, allowing people who have the disease to seek necessary treatment and reducing the global burden of glaucoma.

1.2 Deep Learning in Healthcare

Advances in the field of deep learning, a subset of machine learning focused on deep neural networks, have produced state-of-the-art predictive models that can carry out daily tasks and solve complex problems across a wide variety of industries. In particular, deep learning models for computer vision have successfully classified diseases from medical image data in fields such as radiology, dermatology, and ophthalmology [39, 14, 21]. These disease classification models, which comprise the field of AI-based computer-aided diagnosis (AI-CAD) [19], have the potential to revolutionize the healthcare industry by greatly reducing costs, waiting times, and errors in clinical diagnosis [31]. AI-CAD also enables the creation of large-scale screening programs for common diseases in populations that lack access to affordable healthcare [7], like the glaucoma screening program proposed in Section 1.1. In fact, there have been recent successes in pilot screening programs to diagnose diabetic retinopathy, another common eye disease, but a similar program for glaucoma does not yet exist [61].

Additionally, there are many roadblocks to the successful deployment of an AI-CAD model in a clinical setting. Training these models requires huge amounts of high-quality, annotated data [60], which often come from a handful of sources that are not representative of all potential data that the model will need to classify. This difference in the underlying domains of the source and target data is called *domain shift* [18]. Domain shift is known to diminish model performance, as a model trained on one dataset may not generalize well to unseen images with different styles or features, weakening its predictive power [66]. In practice in a healthcare setting, however, a model must be able to adapt to new data without sacrificing diagnostic accuracy [22],

particularly in a large-scale screening program where, ideally, a single pre-trained model can be used across many clinics to diagnose the disease from their local data sources.

Chapter 2

Literature Review

2.1 Deep Learning for Medical Imaging

Convolutional Neural Networks (CNNs), a class of neural networks that automatically extract feature representations from images through a series of convolutional layers, are the current state-of-the-art in deep learning for computer vision tasks [36]. CNNs have been applied extensively to classification and segmentation problems in the field of medical imaging, even achieving or exceeding physician-level accuracy [17]. Types of medical images to which CNNs have been applied include ultrasounds, X-rays, MRIs, and retinal images— the focus of our research [31]. In the field of ophthalmology, CNNs have been used to classify or segment retinal images to aid diagnosis of diabetic retinopathy, macular degeneration, and glaucoma, among others [21].

2.1.1 Glaucoma Classification

There are numerous CNN models for glaucoma diagnosis in the literature. Many papers use one CNN to precisely segment the optic disc and cup, which is the region of interest for glaucoma diagnosis, and then introduce a second CNN to classify the segmented image [53, 45]. Other papers use a single, often more complex CNN to

classify the images directly [50, 12, 27, 51]. All of these papers focus on achieving good performance in their specific experimental settings rather than in real-world healthcare institutions. There is some research focused on the screening program setting, providing a theoretical framework for implementation [41] and an actual screening program ready for prospective trials [8]. Overall, though, research on CNNs for glaucoma screening is much more scarce than, for example, its radiology counterpart. Our research aims to expand the glaucoma classification literature for the case of multi-institutional glaucoma screening with limited training datasets.

2.2 Domain Generalization

Domain generalization [9] is a theory of learning in the presence of domain shift, or when the target domain differs from the source domain. Machine learning models are trained under the assumption that the source and target data come from the same underlying distribution. In real-world settings, this assumption rarely holds, leading to degradation in performance accuracy [70]. This is especially true in clinical healthcare, where data domains may differ across patient populations, healthcare institutions and practitioners, diagnostic equipment, and time. In fact, the authors that introduced the theory of domain generalization developed their solution for flow cytrometry data, a clinical measurement tool for blood-related pathologies [9].

Domain generalization aims to adjust the data or model to mitigate the detrimental effects of domain shift [58]. It differs from the closely related theory of domain adaptation in that the target data is not assumed to be available at the time of training, which is most realistic to the large-scale clinical scenario.

Domain generalization techniques can largely be divided into two main categories: data-level and model-level techniques. Data-level methods aim to expand or introduce new domains to the training data in a variety of ways, such as through generic

data transformations or adversarial gradients [70]. The basic idea of data-level generalization is that the augmented source data contains such a diverse spread of domains that any unseen target data domain is implicitly represented in the training data. Model-level methods, on the other hand, tweak some aspects of how the model is trained so that it focuses less on learning domain-specific patterns during training; the final parameters used for prediction are, therefore, less dependent on domain-level features. This category includes techniques such as domain alignment, which minimizes some measure of difference between source domains during training, and ensemble and self-supervised learning, which use learning regimes meant to improve model generalization ability but are not specific to the domain generalization scenario [70].

Data-level methods may be simpler to design and implement, but they provide a superficial solution that does not address the identically distributed domain assumption at the root of the domain shift problem. Further, transformations that enable generalization for one type of data will likely not translate to other types of data, meaning that different techniques must be developed and tested for every problem. Conversely, model-level methods target the identical distribution assumption, but this requires changing the fundamental machine learning framework. In addition to being a complex task, domain generalization theory is not well-studied, so removing a central assumption could have unknown and potentially adverse affects on model training.

2.2.1 Application to Glaucoma

Prior work focused on the deployment of machine learning in healthcare cite performance degradation due to domain shift as a key challenge to be addressed before AI diagnostic tools can become readily available for clinical use [46, 65]. This ob-

stacle motivates the need for robust domain generalization methods in AI-CAD [66]. Domain generalization methods have been presented for CNNs for medical imaging in general [35, 34], retinal images [3, 63], and glaucoma classification specifically [25, 69, 28]. The authors of [25] focus on data-level techniques and recommends a specific combination of retinal image pre-processing steps that help mitigate domain shift. The authors of [69] present a framework for learning robust features during training to reduce the effect of domain-specific features on model predictions, and the authors of [28] propose a framework that uses a regression model for the binary classification task to smooth the prediction distribution and reduce false positives in the real-world setting. Overall, domain generalization research for deep learning-based glaucoma classification is limited.

2.3 Problem Statement

This thesis aims to develop a domain-generalized glaucoma classification framework to enable multi-institutional glaucoma screening using a single trained model and local retinal images. We focus on understanding how domain shift manifests across retinal image datasets and how different domain generalization methods affect how a model learns image features during training.

Chapter 3

Data

Since domain generalization focuses on the setting in which the source data comes from a different distribution than the target data, it is crucial to have multiple distinct datasets with different underlying distributions on which to conduct our experiments. To that end, we choose four of the most popular public retinal image datasets for our research. The rest of this section introduces the datasets and describes image preprocessing steps.

3.1 Datasets

The four datasets used in this paper are:

1. ACRIMA Project[1] (ACRIMA) [16]
2. Online Retinal fundus Image database for Glaucoma Analysis (ORIGA) [68]
3. REtinal FUndus Glaucoma ChallengE (REFUGE) [44]
4. Retinal IMage database for Optic Nerve Evaluation for Deep Learning (RIM-ONE DL) [5]

[1] https://www.cvblab.webs.upv.es/en/project/acrima_en/

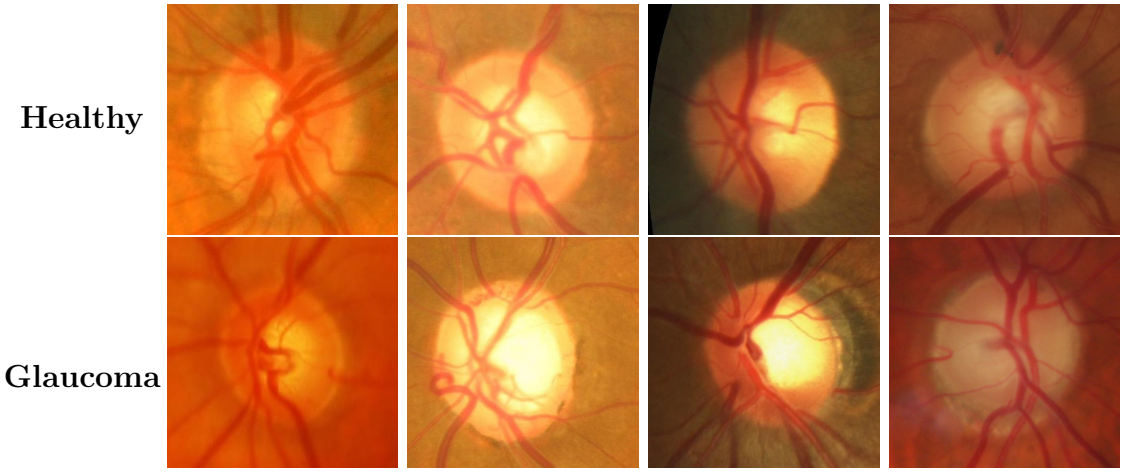
Name	Glaucoma	Total	% Glaucoma	Source
ACRIMA	396	705	56%	Ministerio de Economía y Competitividad of Spain
ORIGA	168	650	26%	Singapore Eye Research Institute
REFUGE	80	805	10%	unknown
RIM-ONE	177	490	36%	3 Spanish hospitals

Table 3.1: Retinal Image Datasets.

The details of each dataset are listed in Table 3.1. The columns correspond to the name of the dataset, the number of images in the positive class (glaucoma), the total number of images, the class balance ($100 \cdot \frac{glaucoma}{total}$), and the institution that sponsored or created the retinal image database, if known. REFUGE and ORIGA were downloaded from Kaggle[2], ACRIMA from Figshare[3], and RIM-ONE DL from GitHub[4]. Each dataset consists of hundreds of color fundus retinal images labeled as either "healthy" or "glaucoma", making them well-suited for the binary classification task. They were chosen because they are publicly available, frequently used in the literature, and among the largest labeled retinal image datasets to date [4].

A typical retinal image from each dataset is presented in Figure 3.1, shown post-processing procedure described in Section 3.2. The images' brightnesses and saturations differ between datasets, but the differences between the "healthy" and "glaucoma" classes in each dataset are less obvious to the untrained eye. A bright center, or enlarged optic cup, which is a common characteristic of glaucoma [16], can be observed in the "glaucoma" images from ORIGA and REFUGE. However, this feature is not present across all glaucoma cases, and it cannot be used as a sole indicator of the disease.

[2]https://www.kaggle.com/datasets/arnavjain1/glaucoma-datasets
[3]https://figshare.com/s/c2d31f850af14c5b5232
[4]https://github.com/miag-ull/rim-one-dl



(a) ACRIMA (b) ORIGA (c) REFUGE (d) RIM-ONE

Figure 3.1: Retinal Images from Four Datasets.

3.2 Region of Interest Extraction

We begin by extracting the region of interest (ROI) from each retinal image. In practice, glaucoma diagnosis is based on inspection of the optic cup-to-disc ratio [67], so cropping retinal images to the optic disc region reduces image size and noise while preserving relevant information for classification. ROI extraction is a common step in retinal image pre-processing, and recent literature use either standard machine learning [12, 68] or deep learning segmentation approaches [53, 16]. We adapt the machine learning algorithm described in [67], which builds on the algorithm from [37], to extract the optic disc ROIs from the ORIGA and REFUGE images; ACRIMA and RIM-ONE images are already cropped around the optic disc.

The generalized ROI extraction algorithm from [37] works as follows: First, the retinal image is converted to grayscale, and the pixels in the top 0.5% of intensities are selected. The image is then divided into an 8×8 grid, and the grid tile that contains the highest amount of top 0.5% pixel intensities is identified as the ROI. Since the optic disc is usually the brightest part of a retinal image, the region with the highest concentration of the brightest pixels has the greatest probability of containing the disc.

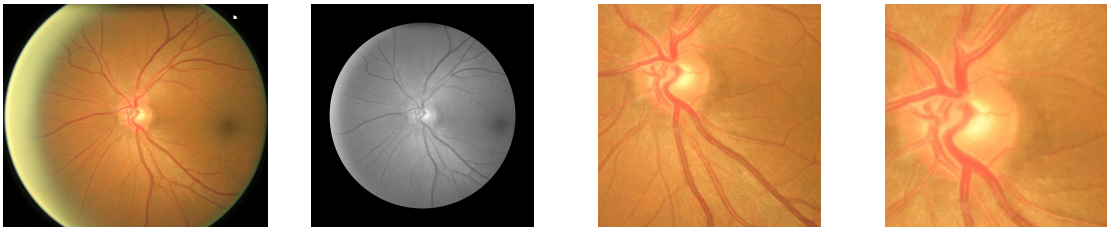
The authors of [67] notice that retinal images often have bright fringes along their circumferences, likely due to camera setup or practitioner error. During ROI extraction, this fringe is found to have the highest concentration of bright pixels, so the algorithm fails to identify the optic disc and cup. The authors address this issue by masking out the outer rim of each image and performing ROI extraction on the new, masked image.

We make minor changes to the algorithms of [37] and [67]. Since the original images are mostly rectangular, we pad the shorter edge with black pixels to create a square array; thus, the brightest pixels do not change and the final ROI will also be square. We implement the masking technique from [67] by creating a circle mask centered on the original image with $radius = \frac{5}{6} * edge$. Pixels inside the circle are assigned a value of 1 and outside a value of 0. This circle mask is then multiplied pixel-wise with the original grayscale image, and the resulting masked image is used as input for the first iteration of the ROI algorithm. Once the optic disc tile is identified, we add one tile to each side to ensure the entire disc is contained in the image. Finally, we repeat the algorithm using the output of the first iteration as the input of the second, this time padding the new brightest tile with two tiles on each side.

A radius scaling of $\frac{5}{6}$ in the masking step was chosen so that bright fringes were masked out while keeping optic discs near the image border visible. The two iterations of the algorithm ensure that the optic disc is centered in the final image; simply changing the pixel brightness threshold or grid tile size did not accomplish the same result.

Images from each step in the ROI extraction process are shown in Figure 3.2. Figure 3.2b shows the bright rim successfully masked out of the full image. Figure 3.2c shows the output of the first iteration of the algorithm with the optic disc off-center, while the second iteration re-centers and tightly crops the optic disc in Figure 3.2d. The algorithm parameters were empirically chosen so that the ROI extracted images

matched the size of the pre-cropped ACRIMA and RIM-ONE images to allow for direct comparison during training and testing.



(a) Original Image (b) Masked Image (c) Intermediate ROI (d) Final ROI

Figure 3.2: ROI Extraction Process.

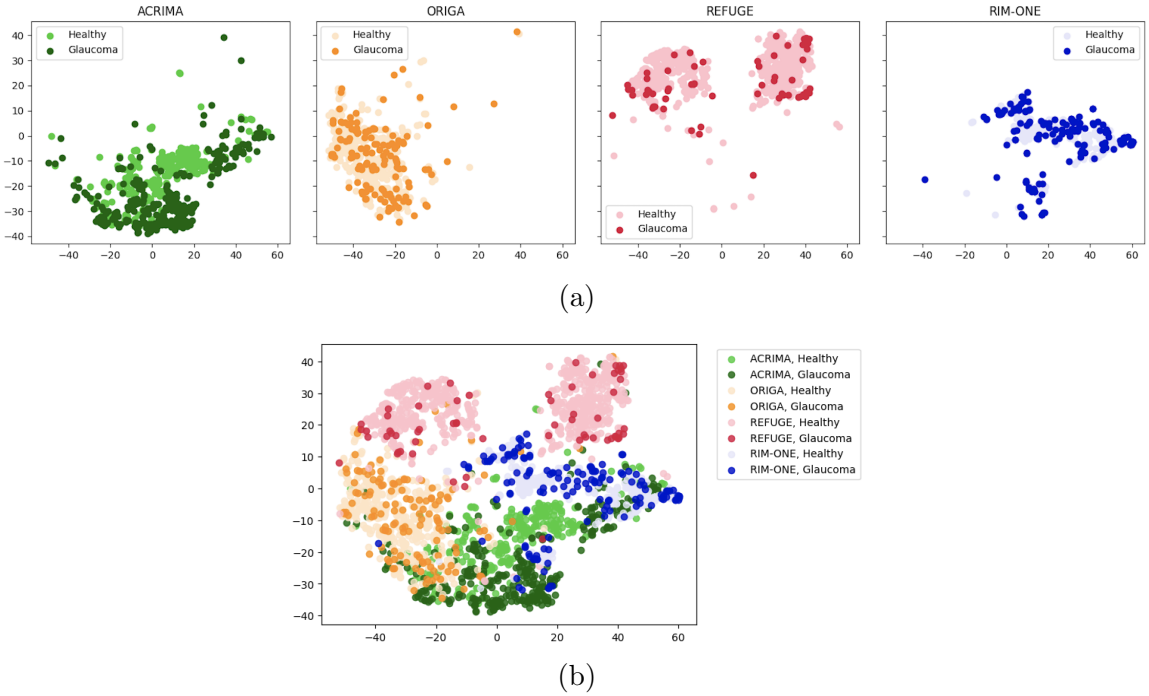
3.3 Domain Comparison

We compare the domain distributions of the four datasets to better understand domain shift in the feature space. A color (RGB) image is mathematically represented by a three-dimensional array $\in \mathbb{R}^{H \cdot W \cdot 3}$, where the dimensions correspond to image height, width, and color channel, respectively. For even comparison, we resize each ROI image to $224 \times 224 \times 3$ pixels. We can then represent a set of n retinal ROI images as a $150,528 \times n$ array by flattening each $224 \times 224 \times 3$ image into a single vector.

We use t-Distributed Stochastic Neighbor Embedding (t-SNE), introduced by Van der Maaten and Hinton in 2008 [57], to project this high-dimensional dataset representation onto the 2-D plane for easy visualization. t-SNE is an extremely popular dimensionality reduction technique that preserves the essential structure of high-dimensional data in low-dimensional space by optimizing the distances between pairs of data points, meaning similar points will be closer together and dissimilar points farther apart in the 2-D projection. We concatenate all four datasets and apply t-SNE

using scikit-learn's t-SNE package[5] in Python to obtain a global comparison of the datasets' domains.

Figure 3.3 shows the results of the t-SNE projection, with each dataset separated in 3.3a for better visualization and overlaid on a single plot in 3.3b. Each dataset forms its own distinct cluster in the projection space with relatively little overlap between clusters, demonstrating domain shift between the four retinal image datasets. REFUGE forms two distinct clusters in the upper half of the plot, likely due to the differences in images from the test and validation sets in the original database that were randomly reassigned by us to mitigate intra-dataset domain shift. Within the datasets, there is no clear separation between the two classes, indicating that the domain-specific stylistic features are more prominent than the features containing information related to glaucoma status in the raw image representations.



(a)

(b)

Figure 3.3: t-SNE Projections of ROI Image Embeddings.

[5]https://scikit-learn.org/stable/modules/generated/sklearn.manifold.TSNE.html

As described in Section 2.2, there are many possible reasons for this domain shift. The images in each dataset were taken by different cameras, equipment settings, and clinical practitioners. Further, they were collected as part of separate retinal image database projects. The researchers who created ACRIMA describe experts' selecting patients, annotating the images, and discarding inadequate data points [16]; all of these variables affect the final database, and each project had different experts and criteria.

In addition to providing evidence of domain shift, these graphs give a sense of the difficulty of the domain generalization task at hand. A domain generalized glaucoma classification model must not only identify the subtle, high-level features containing information about the presence of glaucoma, but it must do so while ignoring or discarding the much louder information about dataset-specific image characteristics.

Chapter 4

Methodology

First, we use a transfer learning approach with two popular pre-trained convolutional neural network (CNN) architectures to investigate the effect of out-of-domain training data on test performance for the glaucoma classification task. Section 4.1 describes the transfer learning framework, and Section 4.2 provides more details about the specific models. Then, in Section 4.3, we introduce two domain generalization methods and describe our implementations for this problem. All experiments were performed using an NVIDIA T4 or A100 GPU on Google Colaboratory[1] or the Princeton University Adroit cluster, respectively.

4.1 Transfer Learning

Transfer learning is the theory that the knowledge gained from learning a simple task can be leveraged to learn related, more complicated tasks. In deep learning for computer vision, the final parameters or features at convergence of a classification CNN on a huge general image database can be used as initialization parameters or input features, respectively, for a classification model on a smaller, more specific dataset [32]. Since the pre-trained model has already learned how to identify generic low-

[1] https://colab.google/

level image features like edges and shapes, the second, fine-tuned model can focus on learning higher-level features from a specific domain during training.

Transfer learning has a natural application to medical image datasets, which often do not satisfy the massive size requirements for training deep learning models from scratch [32]. Most recent transfer learning approaches use popular CNN architectures pre-trained on the ImageNet database [15], an open-source database containing over 14 million images and 20,000 categories [59]. Papers on glaucoma classification commonly use this regime [53, 45, 27, 51] following a landmark paper showing the success of transfer learning for predicting diabetic retinopathy from retinal images [24].

4.2 Baseline Model Training

Two CNN architectures, ResNet101 and YOLOv8, are fine-tuned on the retinal image datasets to obtain baseline measures of in-domain and out-of-domain glaucoma classification performance. To prepare for training, each dataset of ROI extracted and resized retinal images is randomly divided into a training, testing, and validation set using an 80/10/10% split. We load the pre-trained versions of each model architecture, which have both been trained on the ImageNet-1K database, and re-train only the final fully-connected layer on the retinal ROI images for the binary glaucoma classification task; this is known as *fine-tuning*. The specific configurations for ResNet101 and YOLOv8 are given in Sections 4.2.1 and 4.2.2, respectively. We save the model parameters with the best validation accuracy and use that state dictionary for testing. Both architectures are fine-tuned on each dataset and tested on all four datasets for a total of eight trained models and thirty-two sets of results. The results are presented in Section 5.1.

4.2.1 ResNet101

ResNet101 is a 101-layer version of the ResNet architecture [26] that is ubiquitous in computer vision tasks and has been successfully used for glaucoma classification in the literature [11, 50, 16, 49, 54]. We use the deep learning framework PyTorch[2] for all experiments. The pre-trained ResNet101 model with the most recent ImageNet-1K weights is downloaded through PyTorch's torchvision.models subpackage[3].

In the pre-processing step, images are normalized by the average mean and standard deviation of each color channel of the *training* set. For each experiment, we train the model for 20 epochs using a batch size of 32, the Adam optimizer [33] with default PyTorch settings (learning rate = 0.001, betas = $[0.9, 0.999]$) and standard cross-entropy loss.

4.2.2 YOLOv8

You Only Look Once (YOLO) is a state-of-the-art vision AI algorithm best known for its object detection capabilities [47]. The initial and subsequent iterations of the YOLO system were developed by researchers at Ultralytics[4], a Los Angeles-based software company. The 8th version of the YOLO algorithm, YOLOv8, was released in January 2023 and is impressively accurate and efficient for many vision tasks, including object detection, segmentation, and classification.

We use the pre-trained YOLOv8n-cls model[5], built on PyTorch and available through the ultralytics package in Python, for model fine-tuning. Ultralytics provides an extremely easy-to-use "train" function for fine-tuning on new datasets, which we use to train the model for 100 epochs using a batch size of 32, the 'auto' optimizer (AdamW; learning rate = 7.14×10^{-4}, momentum = 0.9) and cross-entropy loss.

[2]https://pytorch.org/
[3]https://pytorch.org/vision/main/models/generated/torchvision.models.resnet101.html
[4]https://www.ultralytics.com/yolo
[5]https://docs.ultralytics.com/tasks/classify/

4.3 Domain Generalization Methods

We train the following three domain generalization methods for the glaucoma classification task and report the results in Section 5.2.

4.3.1 Fourier-based Augmented Co-teacher

Fourier Augmented Co-Teacher (FACT) is a recent and widely-cited domain generalization framework inspired by properties of the Fourier transform [64]. Based on the observation that the phase component of a Fourier transform contains high-level semantic information while the amplitude component captures low-level noise, the authors propose a Fourier-based data augmentation strategy during training that encourages the model to learn features primarily from the image's phase information. Additionally, they introduce a co-teacher regularization component to ensure agreement between predictions from the original and augmented data.

FACT is model-agnostic and can be used on top of any CNN architecture, including for fine-tuning of pre-trained models. We use the authors' code implementation[6] and adapt it to our four retinal image datasets. We follow the authors of [64] and train FACT in leave-one-domain-out style, resulting in four models each trained on three out of the four datasets. For training, we fine-tune FACT using the pre-trained ResNet101 model for 20 epochs with a batch size of 16 and the optimizer and loss function specified by the method. For direct comparison with the baseline CNNs, we also train ResNet101 and YOLOv8 again to obtain the four leave-one-domain-out models for each architecture.

[6] https://github.com/MediaBrain-SJTU/FACT

4.3.2 Causality Inspired Representation Learning

Causality Inspired Representation Learning (CIRL) similarly attempts to separate high-level signals from low-level image noise and only learns features from the former during training [38]. The authors assume that image data can be theoretically separated into class-related information, or the features which have a causal link to their corresponding class label, and domain-specific stylistic information, which hinders generalization ability. While these differences are unobservable in practice, the authors use statistical properties of causal factors, as well as the data augmentation strategy for extracting phase information proposed by the authors of FACT [64], to generate jointly independent causal factor representations from the image data. Finally, they use an adversarial mask model to ensure that each jointly independent feature representation contributes sufficient information to the class prediction.

We use the authors' code implementation[7] for training and testing, which itself is heavily borrowed from the FACT code. Like FACT, CIRL can be implemented in conjunction with any pre-trained CNN. Thus, we fine-tune CIRL using the same configuration settings as FACT to get the four leave-one-domain-out models.

4.3.3 Variational Autoencoder for Domain Generalization

The authors of [13] introduce the Variational Autoencoder for Domain Generalization (VAE-DG). The method uses a conventional variational autoencoder to obtain an estimate for the posterior distribution of latent variables from the retinal images. This distribution represents features in an "optimally disentangled latent space," which the authors theorize are naturally domain-agnostic and can therefore be used to accurately classify retinal images across different domains. VAE-DG uses the pre-trained ResNet50, the 50-layer version of the ResNet architecture [26], as the backbone of the encoder. The method also uses a tri-objective weighted loss function to simulta-

[7] https://github.com/BIT-DA/CIRL

neously minimize the standard empirical error, the KL divergence between the prior and posterior encoder distributions, and the difference between the original image and its reconstruction from the decoder [13]. We use the authors' code implementation,[8] which adheres to the DomainBed structure for standardization of domain generalization research proposed in [23].

VAE-DG is the only method explored in this paper that was specifically developed for and tested on retinal image data. However, the authors developed their method for the task of multi-class diabetic retinopathy classification, not glaucoma diagnosis. Diabetic retinopathy is another progressive eye disease that can be classified from retinal images on a scale of increasing severity [2], and, to our knowledge, VAE-DG is only the second domain generalization technique created for diabetic retinopathy diagnosis in the deep learning literature. The authors of [13] report that VAE-DG outperforms the first method, DRGen, which is introduced in [3]. To the best of our knowledge, the only deep learning-based domain generalization method explicitly designed for glaucoma classification is Data Augmentation-based Feature Alignment (DAFA) [69], which has no public code implementation to date.

[8] https://github.com/sharonchokuwa/VAE-DG

Chapter 5

Results

In this chapter, we present the results for the baseline ResNet101 and YOLOv8 models in Section 5.1 and the three domain generalization methods in Section 5.2.

5.1 Baseline Results

We report two performance measures for each test set: Accuracy (ACC) (5.1) and Area under the Receiver Operating Characteristic (ROC) Curve (AUC). AUC score represents the probability that a model assigns a randomly chosen positive sample a higher probability of belonging to the positive class than a randomly chosen negative sample [29]; scores closer to 1 indicate better performance, while a completely random classifier would report an AUC score around 0.5. In the literature, the ROC curve and corresponding AUC score are considered better metrics for determining the quality of a binary classifier than raw accuracy [29].

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \tag{5.1}$$

**TP = True Positives, FN = False Negatives*

As a baseline for model performance, we first evaluate the performance of each training set on the same dataset's test set, i.e., without domain shift. Both models achieve fairly good results on in-domain test sets, and YOLOv8 either matches or outperforms ResNet101 for all datasets. The fine-tuned YOLOv8 model reaches its highest accuracy and AUC of 100% and 100% , respectively, for the ACRIMA dataset and its lowest of 86.2% and 90.9% , respectively, for ORIGA. This indicates that our transfer learning approach is well-suited for glaucoma classification and provides a reasonable baseline from which to study domain shift.

Next, we turn to the domain shift case by evaluating the performance of the trained models on every test set. Table 5.1 shows these results; the best test performance for each trained model is in bold. For every trained model, there is a drop in both performance metrics when tested on an out-of-domain (OOD) dataset, or a dataset different from the dataset on which the model was trained[1]. This supports the claim that predictive performance degrades in the presence of domain shift, specifically for color fundus retinal images from different sources.

		Test Set							
		ACRIMA		ORIGA		REFUGE		RIM-ONE	
Model	**Train Set**	**ACC**	**AUC**	**ACC**	**AUC**	**ACC**	**AUC**	**ACC**	**AUC**
YOLOv8	ACRIMA	**1.000**	**1.000**	0.462	0.553	0.778	0.811	0.327	0.625
	ORIGA	0.493	0.693	0.862	**0.909**	**0.914**	0.780	0.796	0.895
	REFUGE	0.592	0.623	0.769	0.717	**0.975**	**0.983**	0.755	0.897
	RIM-ONE	0.592	0.636	0.585	0.741	0.901	0.838	**0.918**	**0.987**
ResNet101	ACRIMA	**1.000**	**1.000**	0.631	0.684	0.309	0.580	0.388	0.629
	ORIGA	0.648	0.691	**0.831**	**0.861**	0.765	0.488	0.694	0.818
	REFUGE	0.324	0.350	0.708	0.616	**0.963**	**0.986**	0.694	0.643
	RIM-ONE	0.620	0.553	0.600	0.721	0.605	0.741	**0.898**	**0.919**

Table 5.1: Baseline CNN Results.

[1]This is untrue for the ORIGA-trained YOLOv8 model, which has higher accuracy for REFUGE than ORIGA. However, the REFUGE accuracy may be misleadingly high due to class imbalance; AUC is a better performance measure.

Additionally, YOLOv8 outperforms ResNet101 in AUC score on all but two OOD test sets. YOLOv8 is a new, general-purpose vision model that is best known for its object detection abilities, and, to the best of our knowledge, it has not been used in any retinal image research to date. Its dominance over ResNet101 for both in- and out-of-domain performance indicates that researchers in the field of AI for medical imaging should investigate YOLOv8 as an alternative to more established CNNs. Future research should also explore why general-purpose vision models may offer advantages over classification- or segmentation-specific models, even in single task settings.

It is also important to acknowledge the perfect scores for both in-domain ACRIMA models. A potential explanation for this phenomenon can be found in the t-SNE projection of the dataset's image embeddings when performed locally, or on each dataset separately, rather than globally, as in Section 3.3. Figure 5.1 shows that even before any normalization or model training, the two classes in the ACRIMA dataset are already roughly distinct[2]. It is unknown whether this is a result of the data collection or validation process or an inherent property of the images, but it makes sense that both CNNs are able to learn the glaucoma-related features from the ACRIMA images more easily than the other datasets.

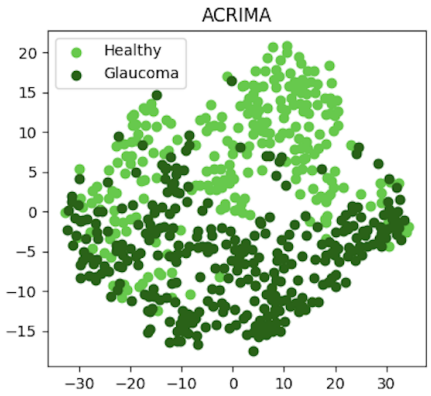
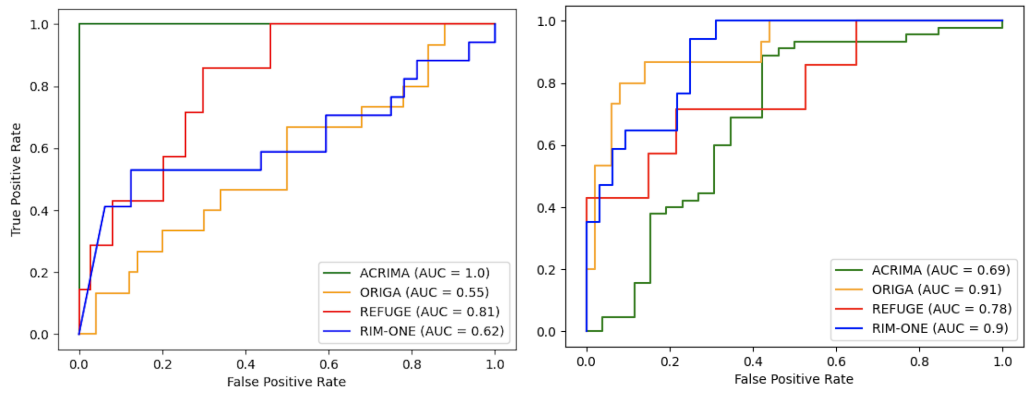


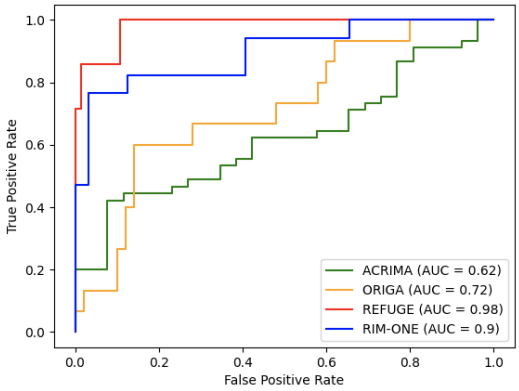
Figure 5.1: t-SNE Projection of ACRIMA ROI Image Embeddings.

[2]The local projections for all four datasets can be found in the Appendix. This pattern is only observed with ACRIMA.

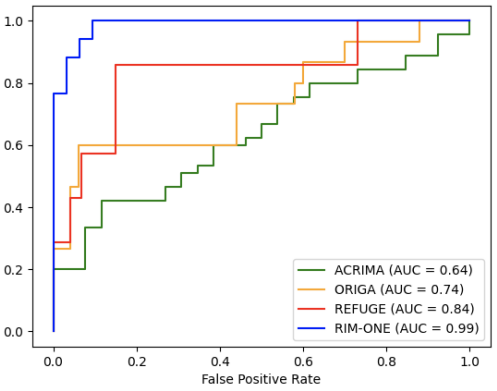


(a) ACRIMA

(b) ORIGA



(c) REFUGE



(d) RIM-ONE

Figure 5.2: ROC Curves For Finetuned YOLOv8 Models.

Given the difference in magnitude of domain shift-driven performance drop across different test-train combinations, we present the ROC curves for all four finetuned YOLOv8 models in Figure 5.2. Each graph represents a trained model, and each line is the ROC curve for a test dataset. The ROC curve plots the true and false positive rates for a given binary classifier at different classification thresholds, or the softmax probability above which a test sample is classified as positive. A perfect ROC curve hugs the top left corner of the graph, as seen in the ACRIMA-ACRIMA training-testing line; a completely random classifier would form the $y = x$ line.

There are clear differences in the patterns of the OOD ROC curves. For example, as an OOD test set, ACRIMA has consistently low performance, while REFUGE

and RIM-ONE seem to perform fairly well when OOD for the other's model. The ORIGA-trained model shows the least evidence of performance drop due to domain shift, especially in its test performance on RIM-ONE, though this could be due to its low accuracy in baseline performance. Further, while the ORIGA-trained model tests well on RIM-ONE, the opposite case does not hold. To further explore this diversity in domain shift-driven performance drop, Section 5.1.1 will investigate how the CNN represents OOD images through features it learns during model training.

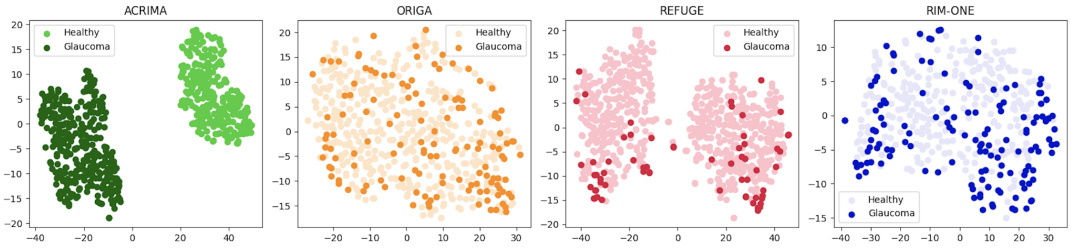
5.1.1 CNN Learned Features

Any classification CNN can be conceptually split into two components: a feature extractor, or the series of convolutional layers which derive complex patterns from the training images to effectively characterize their respective classes; and a classifier, typically consisting of a simple feedforward neural network, which assigns each image to its most probable class based on the patterns that the feature extractor has identified. When a trained model makes a prediction on a new image, the image is fed through the convolutional layers to extract its relevant features; immediately before the classification step, then, each image is represented as a high-dimensional vector of CNN learned features. Like the raw image embeddings from Section 3.3, we use t-SNE to visualize these features in the two-dimensional space.

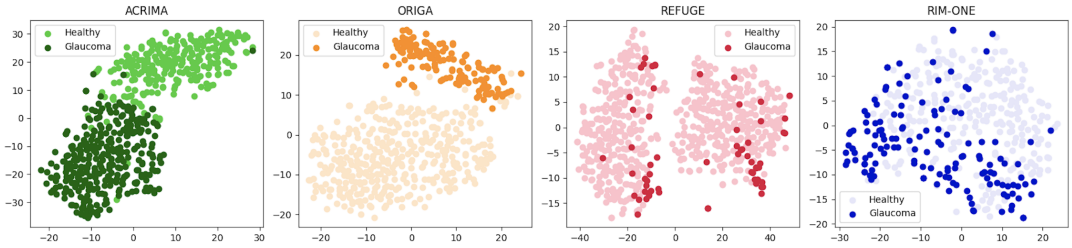
These projections are shown in Figure 5.3, where each row is a trained model, and each column is a dataset of image features[3]. t-SNE has been applied to each dataset separately for a local comparison of features. The subplots along the main diagonal, which correspond to the in-domain test sets, show the ideal result: perfectly separated clusters for each class, which the classifier can easily distinguish in the next step. If the model were able to generalize to OOD images, one would expect similar

[3]We use the training rather than testing split for this visualization so that there is enough data to easily identify patterns. Thus, the accuracies of the in-domain class separations are amplified, but the overall patterns remain the same as the testing data.

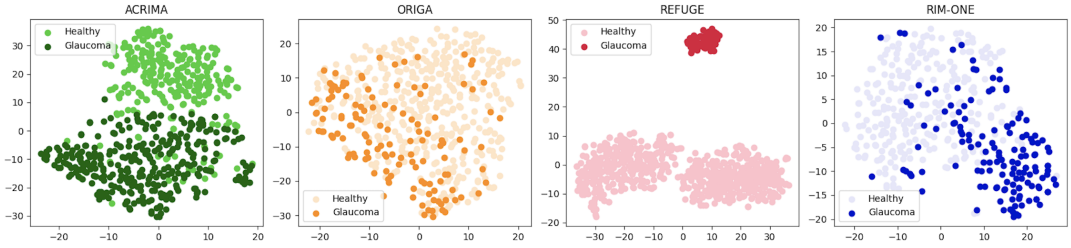
separations in the OOD figures. However, it appears that the CNN learned patterns for each in-domain dataset largely fail to translate to the three other datasets.



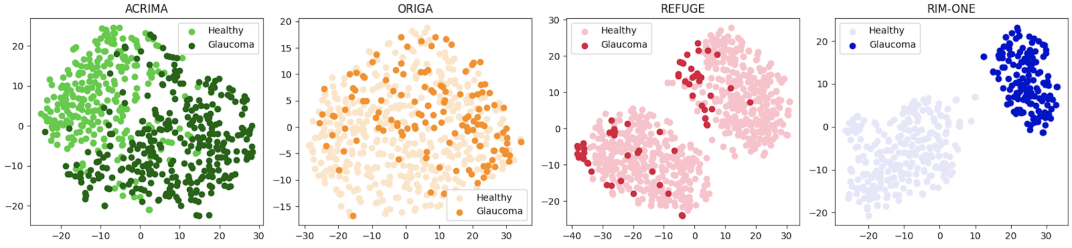
(a) ACRIMA



(b) ORIGA



(c) REFUGE



(d) RIM-ONE

Figure 5.3: t-SNE Projections of CNN Learned Features from Final Layer of YOLOv8.

Although this result reflects the OOD performance drop observed in Section 5.1, the degree of failure is surprising. Despite their domain differences, the four datasets all contain the same type and view of retinal ROI image, so one would expect that at least some of the glaucoma-related features were characteristic of all retinal images, rather than entirely source-specific. This is clearly not the case; besides perhaps the REFUGE-trained results for RIM-ONE, where the glaucoma class mostly occupies the right side of the figure, the classes for the OOD images seem randomly dispersed in their respective feature spaces[4]. In other words, it is not that the model is not generalizing *enough* to OOD retinal images; it is simply not generalizing *at all*.

This result motivates the need for domain generalization training frameworks that focus on forcing the model to learn domain-agnostic features. Further, the t-SNE projections can be used to compare features learned by the vanilla CNNs to those learned in a domain generalization setting.

5.2 Domain Generalization Results

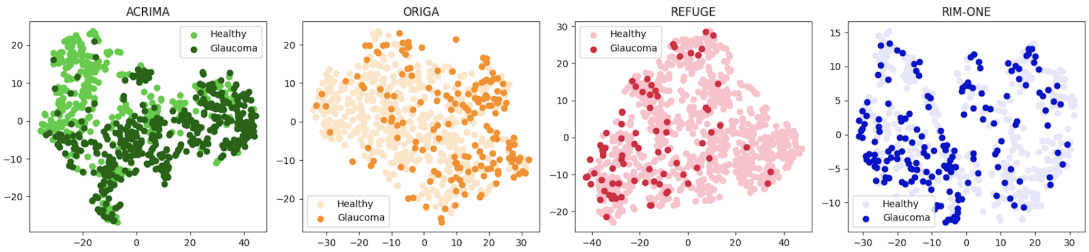
The leave-one-domain-out AUC scores for the two baseline and three domain generalization methods are presented in Table 5.2, with the highest and second highest AUC scores for each left out dataset in bold and underlined, respectively. Each column corresponds to a model trained on the three other datasets and tested on the full left out dataset. The CIRL method has the best performance on average, achieving the first or second highest AUC score on all four OOD datasets. All three domain generalization methods outperform the baseline ResNet model on which they are built, but only CIRL manages to outperform the baseline YOLOv8, by 2.5%. Overall, domain generalization test performance of around 75-80% on OOD data is consistent with reportedly successful methods in the domain generalization literature, but it still fails to meet the standards for practical use.

[4]Again, besides ACRIMA, which was already roughly separated before feature extraction.

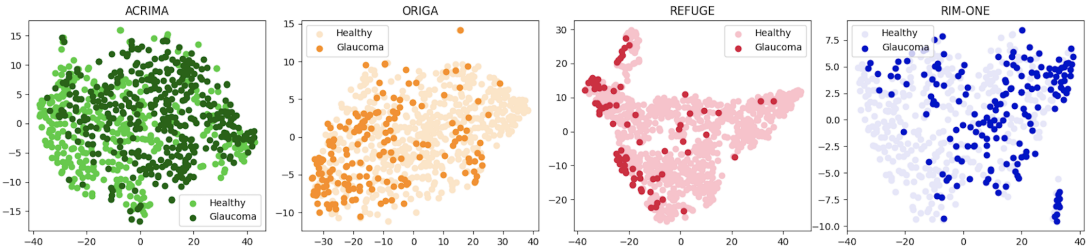
	Left Out Dataset				
Method	ACRIMA	ORIGA	REFUGE	RIM-ONE	Average
YOLOv8 (Baseline)	**0.783**	0.720	0.815	<u>0.794</u>	<u>0.778</u>
ResNet101 (Baseline)	0.748	0.712	0.743	0.623	0.707
FACT	0.701	<u>0.758</u>	0.823	0.761	0.761
CIRL	<u>0.751</u>	**0.789**	**0.869**	**0.802**	**0.803**
VAE-DG	0.716	0.656	<u>0.860</u>	0.762	0.748

Table 5.2: AUC Scores from Leave-One-Out Domain Generalization Methods.

We can examine the features learned by the CNN in the domain generalization setting to determine whether these methods are successful in extracting domain-agnostic features for classification on OOD datasets. Figure 5.4 shows the t-SNE projections of the features learned by the baseline ResNet101 and CIRL models when predicting on each left out dataset. The features learned through the CIRL method are able to roughly group the two classes in the REFUGE dataset, as most glaucomatous images are clustered along the left side of the plot. The glaucoma classes for ORIGA and RIM-ONE are roughly grouped in the left and right sides of their respective plots as well, though they still greatly overlap with the healthy classes, while CIRL seemingly fails to separate the ACRIMA classes in the projected feature space. The empirical observations also mirror the AUC score results for the CIRL method on each left out dataset. Although the t-SNE projections of CNN learned features are not a rigorous measure of model performance, they show that CIRL is somewhat successful in identifying glaucoma-related features during training that can be generalized to unseen retinal image domains.



(a) ResNet101 (Baseline)



(b) CIRL

Figure 5.4: t-SNE Projections of CNN Learned Features from Final Model Layer for OOD Datasets.

Chapter 6

Conclusion

In this thesis, we have studied the impact of domain shift across retinal image datasets on deep learning models for glaucoma classification and the effectiveness of three domain generalization methods. In Section 5.1, we find that the differences in domain distributions across four color fundus retinal image datasets are prominent enough to degrade the glaucoma classification abilities of two conventional convolutional neural networks when tested in unseen retinal image domains. We visualize the high-dimensional features learned during model training in Section 5.1.1, and we demonstrate that features which successfully divide the healthy and glaucoma classes for in-domain images overwhelmingly fail to translate to out-of-domain data. These results indicate that strongly domain-specific feature extraction is at least one cause of domain shift-driven performance degradation, although other factors, such as differences in class imbalances across datasets, could also contribute to the problem.

Ultimately, the findings in Section 5.1 motivate the need for domain generalization techniques that focus on forcing the backbone feature extractor to identify purely domain-agnostic features. To this end, we train and test three such domain generalization methods in Section 5.2. We find that Causality Inspired Representation Learning [38] achieves the highest average test performance on out-of-domain datasets, and we

verify that these scores correspond to improved class clustering in the CNN learned feature space. However, there remains a pressing need for out-of-domain performance improvement before trained models can be safely deployed in unseen domains in real-world clinical scenarios.

Our research begins to fill the gap in the scientific literature surrounding domain generalization techniques for deep learning-based glaucoma classification. To our knowledge, we are the first to apply FACT, CIRL, and VAE-DG to the glaucoma classification task, and we are among the first to visualize domain generalization performance from the perspective of the models' learned feature representations. A secondary contribution of this research is the finding that the baseline YOLOv8 algorithm outperforms ResNet101 for in-domain glaucoma classification. YOLOv8 has not yet been thoroughly explored in the field of medical imaging, and, to date, it has never been used for retinal image classification. In Section 6.1, we propose future research directions to build upon the discoveries in this thesis.

6.1 Future Work

There is a notable lack of research on the application of domain generalization techniques to the problem of glaucoma diagnosis from retinal images. This thesis has tested three methods— two developed for several standard classification datasets, and one for diabetic retinopathy grading of retinal images. However, all three methods are derived from fundamental machine learning principles; none integrate information specific to the type of data they evaluate. To this end, in addition to testing a wider variety of existing techniques, future research should focus on developing domain generalization methods specifically for glaucoma classification. For example, expert knowledge about glaucoma-related retinal image features could be incorporated into the feature representation constraints proposed in [64] and [38] to obtain more prob-

able representations of glaucoma-causing features for classification.

Additionally, the favorable performance of the YOLOv8 model for baseline glaucoma classification suggests the need for further research on the application of YOLOv8 to medical image tasks. Unfortunately, the latest version of the YOLO architecture will not be added to PyTorch's torchvision package[1], and its implementation details are private. This opacity makes it difficult to integrate YOLOv8 into existing PyTorch-based machine learning training frameworks, including the domain generalization methods used in this thesis. In order for YOLOv8's potential in clinical diagnostics to be fully realized, the algorithm's architecture must be made available for use in public research.

The importance of domain generalization research for medical AI cannot be understated. As deep learning models are increasingly deployed for use in large-scale clinical screening programs, they must be extensively validated on unseen domains, including diverse medical institutions and patient populations. There is huge potential for the use of AI diagnostics in settings where patients have limited access to quality healthcare or trained specialists, but this is only possible if pre-trained models can be trusted to make accurate predictions on new medical image data. For the future of medical AI to be truly accessible, we must invest in the research and development of robust, domain generalizable systems.

[1]https://github.com/ultralytics/ultralytics/issues/286#issuecomment-1687128150

Chapter 7

Application to Global Health Policy

This thesis has investigated the effect of domain shift and subsequent domain generalization techniques on the ability of deep learning models to accurately identify the presence of glaucoma from retinal images. Domain shift refers to the case in which the data sources that a machine learning model encounters in practice fall outside of the distribution of data sources used to train the model. Domain shift can arise from a variety of sources, such as across retinal images taken at different healthcare institutions, by different practitioners or machines, of different patient populations, or a combination of these factors. An image is out-of-domain when it comes from a source that was not used for training the model. Domain generalization techniques encompass methods which alter the training data or the model framework so that the trained deep learning model can better generalize to unseen testing domains.

We find that even in the presence of qualitatively minimal domain shift, i.e., color fundus retinal images from different public databases, the trained deep learning model is unable to identify glaucoma-related features from out-of-domain images and thus suffers from performance degradation in the glaucoma classification task. Further,

we implement three widely-cited domain generalization methods and find that only one is able to improve out-of-domain performance over the baseline, although still not enough to be approved for clinical use. This chapter will outline the potential of AI diagnostic systems— particularly glaucoma classification models— to achieve current global vision goals while emphasizing barriers that must be overcome prior to the widespread adoption of these technologies.

7.1 Current State of Global Vision Care

In 1999, the World Health Organization (WHO) and the International Agency for the Prevention of Blindness announced their joint "VISION 2020: The Right to Sight" campaign, which aimed to eliminate preventable blindness globally by 2020 through a series of policy changes and healthcare initiatives [1]. Although an estimated 1 billion people were still affected by avoidable vision impairment in 2020 [62], the campaign was successful in bringing international awareness to the issue of preventable blindness, spurring the adoption of national policy plans in more than 100 countries and mobilizing vast amounts of funding and resources from governments, private companies, and non-governmental organizations [30].

In 2020, WHO passed an updated resolution to continue addressing preventable vision loss. Recognizing the disproportionately high burden of avoidable blindness on low- and middle-income countries (LMICs) and rural areas, the resolution frames the future of vision care through an accessibility lens and urges member states to integrate vision care into universal health coverage policies and primary care systems [62]. A significant reduction in global rates of preventable vision impairment is only achievable if people can access quality eye care, including regular checkups and affordable treatments, through their normal care pathways.

AI-powered medical diagnostics offer a promising vehicle for the integration of vision care into existing health systems. Further, as the top cause of irreversible vision loss worldwide, with LMICs bearing the brunt of the global disease burden, glaucoma is a central focus of global vision care campaigns [48]. The authors of [20] explain that glaucoma is a prime candidate for AI-driven screening programs, as the progressive eye disease is sufficiently prevalent and can be diagnosed by deep learning models, and patients benefit greatly from early diagnosis in regards to both vision retention and future cost. Traditional diagnostic methods are prohibitively expensive for large-scale screening, however. Rather than needing a dedicated eye clinic and trained ophthalmologist, resources which areas with the highest eye disease burdens systematically lack, to diagnose glaucoma, the screening framework proposed in this thesis would only require access to a primary care clinic with a retinal imaging machine and a functioning computer on which to deploy the preprogrammed AI model. The next section will provide an overview of the current status of AI-driven screening programs for common eye diseases in LMICs.

7.2 Artificial Intelligence-Driven Vision Screening Programs

There are several pilot programs exploring the feasibility of AI screening for diabetic retinopathy (DR), another high-burden progressive eye disease found in people with diabetes, in LMICs. Prospective studies conducted on patients with diabetes in Zambia, India, Thailand, and China report promising results for automated detection of DR from retinal images [61, 40]. Notably, the screening program in Zambia used a model trained on retinal images from Singaporean patients of Chinese, Malay, and Indian descent with different camera and image settings [7], indicating that these demographic and technical changes from training to testing did not degrade model

performance in this case. This could mean that this shift was not strong enough to cause significant differences between the training and testing image domains, or domain shift did occur but did not noticeably affect model accuracy. Similarly, researchers in Singapore found that a deep learning model for DR classification trained on retinal images from Singaporean patients could successfully generalize to image data from several different multiethnic populations [55]. In contrast, our research finds prominent performance drop when the glaucoma classification model is tested on an outside retinal image dataset.

Possible explanations for this discrepancy in results could come from the differences between visual indicators of DR and glaucoma in retinal images; glaucoma features may be harder for a deep learning model to identify from a retinal image, leading it to rely more on domain-specific characteristics that do not generalize to other datasets. Additionally, all validation datasets in the Singaporean study were still cleaned by the same team of researchers, while the four retinal image databases that we use come from four distinct studies and research groups. It is possible that the domain shift we observe is driven primarily by the cleaning and validation process rather than factors associated with the camera setup or image subjects. Overall, there is a pressing need for an external validation study similar to [55] for glaucoma classification models, as well as more extensive research on all potential sources of domain shift across retinal image datasets.

The prospective DR study conducted in Mumbai, India showed strong potential for a large-scale screening program using smartphone images and an offline AI system. Healthcare workers, untrained in retinal photography, took photos of diabetes patients' retinas at their local healthcare clinics. A first offline AI model validated the quality of each retinal image, prompting the user to retake the photo if unsatisfactory, and a second offline AI model classified the image as positive or negative for referable DR [42]. The relatively high performance of this method in the prospective analysis

indicates that smartphone cameras and offline AI models may be feasible alternatives to professional retinal imaging equipment and cloud-based software. These alternatives would greatly decrease screening costs and improve the likelihood of successful implementation in resource-scarce healthcare settings. As before, these methods should also be tested for glaucoma screening.

In general, however, the majority of prospective vision AI screening programs are conducted in high-income countries, which have stronger healthcare infrastructures and can afford to develop and use their own specialized technology [6]. Further, although there is a strong case for AI-powered glaucoma screening programs in LMICs, none had been introduced as of 2020 [61]. While there are certainly many challenges to address before a fully automated glaucoma screening program can be deployed at the general population level [11], it is unclear why so many more deep learning methods and prospective studies exist for DR than for glaucoma.

7.3 Policy Recommendations

This chapter has demonstrated how AI-powered vision screening programs in LMICs can be used to integrate accessible and affordable eye care into primary care systems, thus advancing modern global vision targets. Before generic AI models are feasible for multi-institutional and largely unsupervised clinical use, however, many challenges must be addressed. First, critics argue that AI vision systems still require expensive technology and a reliable Internet connection, which is not guaranteed in rural regions of LMICs [48]. This highlights the importance of the Indian program [42], which appears to be the first of its kind. Smartphone-based imaging and offline AI models should be a central focus of future AI screening studies, as the low-resource setting is both the most likely in LMICs and the hardest for which to design robust AI systems. This research area should include the creation of labeled smartphone

retinal image datasets that can be used for public research.

Second, while screening programs increase early diagnosis and public knowledge of common eye diseases, there is no guarantee that people who test positive for a disease will have access to proper treatment. Still, treatment is never possible without a diagnosis, and automated diagnostic tools would free up time for the limited population of ophthalmologists to focus on treatment for sick patients. To be clear, AI technology is not a comprehensive fix for the entire medical pipeline. Instead, it should be used in areas where it can realistically lower costs and improve accessibility; eye disease screening in LMICs strongly fits this criteria.

Finally, there is a major disconnect between medical imaging research in the deep learning field and reports from public health experts on the real-world application of current AI technologies. For example, dozens of glaucoma classification models have been published in the deep learning literature, but no prospective screening studies have been conducted to date; this could be due to the models' failing to translate from idealized lab conditions to clinical use or simply a lack of funding for such programs internationally. The authors of [11] blame the former issue and recommend stricter standardization and validation procedures for glaucoma classification studies, which could be accomplished with greater resources and attention from public health organizations to the research field.

This disconnect is especially strong with respect to AI model generalization, the central focus of this thesis. Our research has shown that state-of-the-art glaucoma classification models utterly fail to generalize to out-of-domain retinal image datasets. This failure effectively prohibits the practical deployment of AI-powered glaucoma screening systems at the population level, but there is a noticeable lack of research on solutions to the problem. On the other hand, while reviews of AI systems for retinal image classification constantly cite concerns of inadequate model generalization to unseen populations [61, 56, 48, 11, 6], none mention the field of domain general-

ization, which, in theory, is the natural starting point for improving generalization ability.

From the technical point of view, domain generalization is a relatively new and extremely difficult problem to solve [70]. The field was only formally introduced in 2011 [9], and it challenges a fundamental assumption of machine learning theory. While incremental progress in the performance of domain generalization methods has been made throughout the last decade [70], there is no universal solution that can guarantee adequate generalization ability in deep learning models, and the field is significantly underresearched compared to other branches of machine learning theory. In short, domain generalization is not the most successful nor profitable branch of machine learning, and it has therefore largely been ignored in favor of its flashier cousins.

Governments and public health organizations are uniquely capable of funding medical AI programs based on global health needs rather than projected revenue. Retinal image classification models need better generalization capabilities, so investment in domain generalization research should be a central focus of vision screening program development. Further, these national and international bodies must mobilize to unite deep learning research and current public health goals, ensuring that AI diagnostics are deployed at a pace and scale that reflects the known limits of machine learning models.

Medical AI is an extremely powerful tool that has the potential to revolutionize healthcare in underserved communities. Before that vision can be realized, however, we must adopt an interdisciplinary approach to AI in the global health field, which can only be achieved through the serendipitous collision of machine learning experts, medical professionals, and public health officials.

Appendix A

Tables and Figures

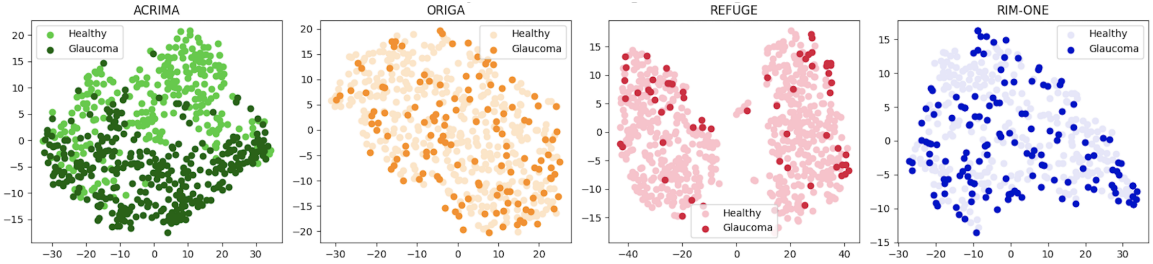
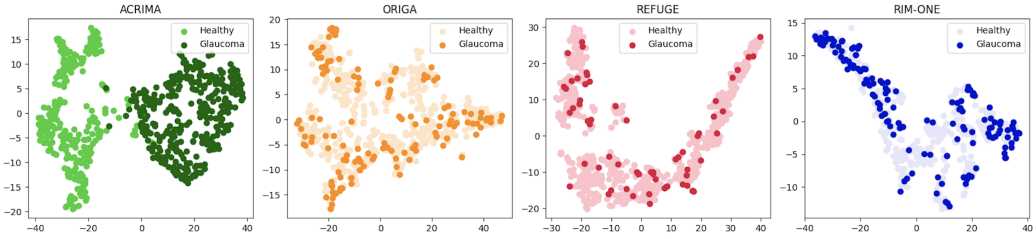
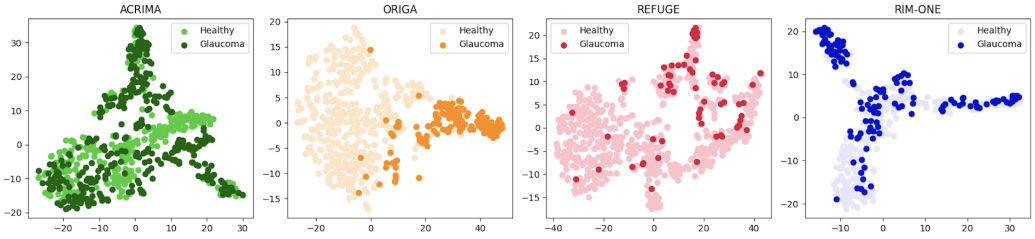


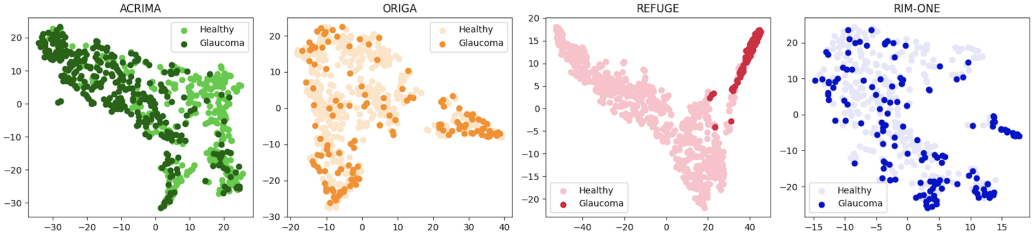
Figure A.1: t-SNE Projections of Local ROI Image Embeddings.



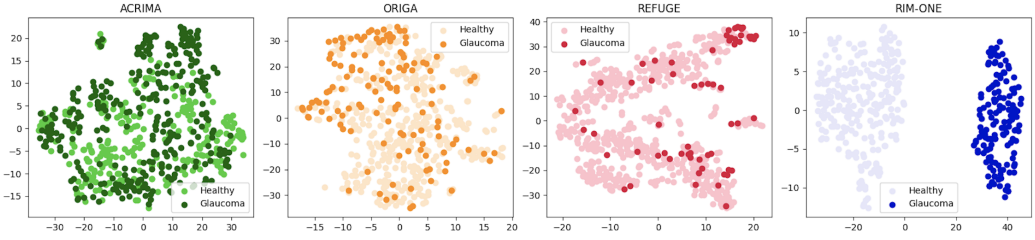
(a) ACRIMA



(b) ORIGA



(c) REFUGE



(d) RIM-ONE

Figure A.2: t-SNE Projections of CNN Learned Features from Final Layer of ResNet101.

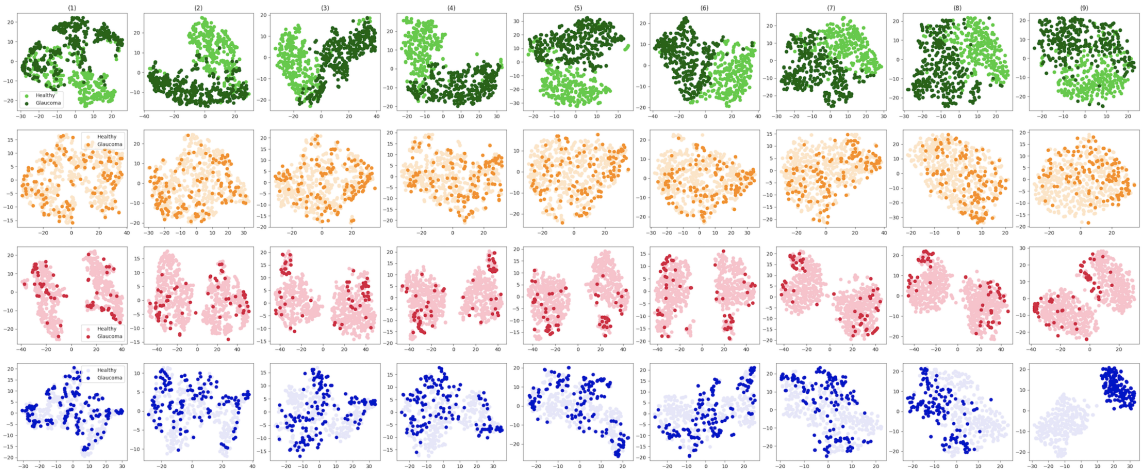


Figure A.3: t-SNE Projections of CNN Learned Features from Each Layer of RIM-ONE-Trained YOLOv8.

Each row is a dataset, and each column is a successive CNN layer, from left to right.

Bibliography

[1] P. Ackland. The accomplishments of the global initiative VISION 2020: The Right to Sight and the focus for the next 8 years of the campaign. *Indian Journal of Ophthalmology*, 60(5):380–386, 2012.

[2] M. Atwany, A. Sahyoun, and M. Yaqub. Deep Learning Techniques for Diabetic Retinopathy Classification: A Survey. *IEEE Access*, 10:28642–28655, 2022.

[3] M. Atwany and M. Yaqub. DRGen: Domain Generalization in Diabetic Retinopathy Classification. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 635–644. Springer, 2022.

[4] M. N. Bajwa, G. A. P. Singh, W. Neumeier, M. I. Malik, A. Dengel, and S. Ahmed. G1020: A Benchmark Retinal Fundus Image Dataset for Computer-Aided Glaucoma Detection. In *2020 International Joint Conference on Neural Networks*, pages 1–7. IEEE, 2020.

[5] F. J. F. Batista, T. Diaz-Aleman, J. Sigut, S. Alayon, R. Arnay, and D. Angel-Pereira. RIM-ONE DL: A Unified Retinal Image Database for Assessing Glaucoma Using Deep Learning. *Image Analysis & Stereology*, 39(3):161–167, 2020.

[6] V. Bellemo, G. Lim, T. H. Rim, G. S. Tan, C. Y. Cheung, S. Sadda, M.-g. He, A. Tufail, M. L. Lee, W. Hsu, and D. S. W. Ting. Artificial Intelligence Screening for Diabetic Retinopathy: the Real-World Emerging Application. *Current Diabetes Reports*, 19:1–12, 2019.

[7] V. Bellemo, Z. W. Lim, G. Lim, Q. D. Nguyen, Y. Xie, M. Y. Yip, H. Hamzah, J. Ho, X. Q. Lee, W. Hsu, M. Lee, L. Musonda, M. Chandran, G. Chipalo-Mutati, M. Muma, G. S. W. Tan, S. Sivaprasad, G. Menon, T. Wong, and D. S. W. Ting. Artificial intelligence using deep learning to screen for referable and vision-threatening diabetic retinopathy in Africa: a clinical validation study. *The Lancet Digital Health*, 1(1):e35–e44, 2019.

[8] A. Bhuiyan, A. Govindaiah, and R. T. Smith. An Artificial-Intelligence- and Telemedicine-Based Screening Tool to Identify Glaucoma Suspects from Color Fundus Imaging. *Journal of Ophthalmology*, 2021.

[9] G. Blanchard, G. Lee, and C. Scott. Generalizing from Several Related Classification Tasks to a New Unlabeled Sample. *Advances in Neural Information Processing Systems*, 24, 2011.

[10] K. Boyd. What Is Glaucoma? Symptoms, Causes, Diagnosis, Treatment. https://www.aao.org/eye-health/diseases/what-is-glaucoma, 2023.

[11] A. K. Chaurasia, C. J. Greatbatch, and A. W. Hewitt. Diagnostic Accuracy of Artificial Intelligence in Glaucoma Screening and Clinical Practice. *Journal of Glaucoma*, 31(5):285–299, 2022.

[12] X. Chen, Y. Xu, D. W. K. Wong, T. Y. Wong, and J. Liu. Glaucoma detection based on deep convolutional neural network. In *2015 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, pages 715–718. IEEE, 2015.

[13] S. Chokuwa and M. H. Khan. Generalizing Across Domains in Diabetic Retinopathy via Variational Autoencoders. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 265–274. Springer, 2023.

[14] M. Cullell-Dalmau, M. Otero-Viñas, and C. Manzo. Research Techniques Made Simple: Deep Learning for the Classification of Dermatological Images. *Journal of Investigative Dermatology*, 140(3):507–514, 2020.

[15] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. ImageNet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255. IEEE, 2009.

[16] A. Diaz-Pinto, S. Morales, V. Naranjo, T. Köhler, J. M. Mossi, and A. Navea. CNNs for automatic glaucoma assessment using fundus images: an extensive validation. *Biomedical Engineering Online*, 18:1–19, 2019.

[17] A. Esteva, A. Robicquet, B. Ramsundar, V. Kuleshov, M. DePristo, K. Chou, C. Cui, G. Corrado, S. Thrun, and J. Dean. A guide to deep learning in healthcare. *Nature Medicine*, 25(1):24–29, 2019.

[18] A. Farahani, S. Voghoei, K. Rasheed, and H. R. Arabnia. A Brief Review of Domain Adaptation. *Advances in Data Science and Information Engineering: Proceedings from ICDATA 2020 and IKE 2020*, pages 877–894, 2021.

[19] H. Fujita. AI-based computer-aided diagnosis (AI-CAD): the latest review to read first. *Radiological Physics and Technology*, 13(1):6–19, 2020.

[20] M. J. Girard and L. Schmetterer. Chapter 3 - Artificial intelligence and deep learning in glaucoma: Current state and future prospects. In G. Bagetta and C. Nucci, editors, *Glaucoma: A Neurodegenerative Disease of the Retina and Beyond - Part B*, volume 257 of *Progress in Brain Research*, pages 37–64. Elsevier, 2020.

[21] P. S. Grewal, F. Oloumi, U. Rubin, and M. T. Tennant. Deep learning in ophthalmology: a review. *Canadian Journal of Ophthalmology*, 53(4):309–313, 2018.

[22] H. Guan and M. Liu. Domain Adaptation for Medical Image Analysis: A Survey. *IEEE Transactions on Biomedical Engineering*, 69(3):1173–1185, 2021.

[23] I. Gulrajani and D. Lopez-Paz. In Search of Lost Domain Generalization. *ArXiv Preprint arXiv:2007.01434*, 2020.

[24] V. Gulshan, L. Peng, M. Coram, M. C. Stumpe, D. Wu, A. Narayanaswamy, S. Venugopalan, K. Widner, T. Madams, J. Cuadros, K. Ramasamy, R. Raman, P. Nelson, J. Mega, and D. Webster. Development and Validation of a Deep Learning Algorithm for Detection of Diabetic Retinopathy in Retinal Fundus Photographs. *JAMA*, 316(22):2402–2410, 2016.

[25] H. Gunasinghe, J. McKelvie, A. Koay, and M. Mayo. Domain Generalisation for Glaucoma Detection in Retinal Images from Unseen Fundus Cameras. In *Asian Conference on Intelligent Information and Database Systems*, pages 421–433. Springer, 2022.

[26] K. He, X. Zhang, S. Ren, and J. Sun. Deep Residual Learning for Image Recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016.

[27] R. Hemelings, B. Elen, J. Barbosa-Breda, S. Lemmens, M. Meire, S. Pourjavan, E. Vandewalle, S. Van de Veire, M. B. Blaschko, P. De Boever, and I. Stalmans. Accurate prediction of glaucoma from colour fundus images with a convolutional neural network that relies on active and transfer learning. *Acta Ophthalmologica*, 98(1):e94–e100, 2020.

[28] R. Hemelings, B. Elen, A. K. Schuster, M. B. Blaschko, J. Barbosa-Breda, P. Hujanen, A. Junglas, S. Nickels, A. White, N. Pfeiffer, P. Mitchell, P. De Boever, A. Tuulonen, and I. Stalmans. A generalizable deep learning regression model

for automated glaucoma screening from fundus images. *NPJ Digital Medicine*, 6(1):112, 2023.

[29] J. Huang and C. X. Ling. Using AUC and accuracy in evaluating learning algorithms. *IEEE Transactions on Knowledge and Data Engineering*, 17(3):299–310, 2005.

[30] International Agency for the Prevention of Blindness. Vision 2020. https://www.iapb.org/about/history/vision-2020/.

[31] J. Ker, L. Wang, J. Rao, and T. Lim. Deep Learning Applications in Medical Image Analysis. *IEEE Access*, 6:9375–9389, 2017.

[32] H. E. Kim, A. Cosa-Linan, N. Santhanam, M. Jannesari, M. E. Maros, and T. Ganslandt. Transfer learning for medical image classification: a literature review. *BMC Medical Imaging*, 22(1):69, 2022.

[33] D. P. Kingma and J. Ba. Adam: A Method for Stochastic Optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[34] C. Li, X. Lin, Y. Mao, W. Lin, Q. Qi, X. Ding, Y. Huang, D. Liang, and Y. Yu. Domain generalization on medical imaging classification using episodic training with task augmentation. *Computers in Biology and Medicine*, 141:105144, 2022.

[35] H. Li, Y. Wang, R. Wan, S. Wang, T.-Q. Li, and A. Kot. Domain Generalization for Medical Imaging Classification with Linear-Dependency Regularization. *Advances in Neural Information Processing Systems*, 33:3118–3129, 2020.

[36] Z. Li, F. Liu, W. Yang, S. Peng, and J. Zhou. A Survey of Convolutional Neural Networks: Analysis, Applications, and Prospects. *IEEE Transactions on Neural Networks and Learning Systems*, 2021.

[37] J. Liu, D. Wong, J. Lim, H. Li, N. Tan, Z. Zhang, T. Wong, and R. Lavanya. ARGALI: An Automatic Cup-to-Disc Ratio Measurement System for Glaucoma Analysis Using Level-set Image Processing. In *13th International Conference on Biomedical Engineering: ICBME 2008 3–6 December 2008 Singapore*, pages 559–562. Springer, 2009.

[38] F. Lv, J. Liang, S. Li, B. Zang, C. H. Liu, Z. Wang, and D. Liu. Causality Inspired Representation Learning for Domain Generalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8046–8056, 2022.

[39] M. A. Mazurowski, M. Buda, A. Saha, and M. R. Bashir. Deep learning in radiology: An overview of the concepts and a survey of the state of the art with focus on MRI. *Journal of Magnetic Resonance Imaging*, 49(4):939–954, 2019.

[40] S. Ming, K. Xie, X. Lei, Y. Yang, Z. Zhao, S. Li, X. Jin, and B. Lei. Evaluation of a novel artificial intelligence-based screening system for diabetic retinopathy in community of China: a real-world study. *International Ophthalmology*, 41:1291–1299, 2021.

[41] A. S. Mursch-Edlmayr, W. S. Ng, A. Diniz-Filho, D. C. Sousa, L. Arnould, M. B. Schlenker, K. Duenas-Angeles, P. A. Keane, J. G. Crowston, and H. Jayaram. Artificial Intelligence Algorithms to Diagnose Glaucoma and Detect Glaucoma Progression: Translation to Clinical Practice. *Translational Vision Science and Technology*, 9(2):55–55, 2020.

[42] S. Natarajan, A. Jain, R. Krishnan, A. Rogye, and S. Sivaprasad. Diagnostic Accuracy of Community-Based Diabetic Retinopathy Screening With an Offline Artificial Intelligence System on a Smartphone. *JAMA Ophthalmology*, 137(10):1182–1188, 2019.

[43] National Glaucoma Research. Glaucoma: Facts and Figures. https://www.brightfocus.org/glaucoma/article/glaucoma-facts-figures, 2022.

[44] J. I. Orlando, H. Fu, J. B. Breda, K. Van Keer, D. R. Bathula, A. Diaz-Pinto, R. Fang, P.-A. Heng, J. Kim, J. Lee, J. Lee, X. Li, P. Liu, S. Lu, B. Murugesan, V. Naranjo, S. S. Phay, S. Shankaranarayana, A. Sikka, J. Son, and H. Bogunović. REFUGE Challenge: A unified framework for evaluating automated methods for glaucoma assessment from fundus photographs. *Medical Image Analysis*, 59:101570, 2020.

[45] S. Phasuk, P. Poopresert, A. Yaemsuk, P. Suvannachart, R. Itthipanichpong, S. Chansangpetch, A. Manassakorn, V. Tantisevi, P. Rojanapongpun, and C. Tantibundhit. Automated glaucoma screening from retinal fundus image using deep learning. In *2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, pages 904–907. IEEE, 2019.

[46] M. P. Recht, M. Dewey, K. Dreyer, C. Langlotz, W. Niessen, B. Prainsack, and J. J. Smith. Integrating artificial intelligence into the clinical practice of radiology: challenges and recommendations. *European Radiology*, 30:3576–3584, 2020.

[47] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi. You Only Look Once: Unified, Real-Time Object Detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 779–788, 2016.

[48] D. M. Rooney, G. Kalra, G. Tabin, S. Kavitha, R. Venkatesh, A. A. Aref, I. Conner, F. Shakarchi, and M. Slabaugh. Glaucoma in the Developing World. https://eyewiki.aao.org/Glaucoma_in_the_Developing_World, 2019.

[49] A. Sallam, A. S. Gaid, W. Q. Saif, A. Hana'a, R. A. Abdulkareem, K. J. Ahmed, A. Y. Saeed, and A. Radman. Early Detection of Glaucoma using Transfer

Learning from Pre-trained CNN Models. In *2021 International Conference of Technology, Science and Administration*, pages 1–5. IEEE, 2021.

[50] S. Serte and A. Serener. A Generalized Deep Learning Model for Glaucoma Detection. In *2019 3rd International Symposium on Multidisciplinary Studies and Innovative Technologies*, pages 1–5. IEEE, 2019.

[51] A. Shoukat, S. Akbar, S. A. Hassan, S. Iqbal, A. Mehmood, and Q. M. Ilyas. Automatic Diagnosis of Glaucoma from Retinal Images Using Deep Learning Approach. *Diagnostics*, 13(10):1738, 2023.

[52] S. D. Solomon, R. Y. Shoge, A. M. Ervin, M. Contreras, J. Harewood, U. T. Aguwa, and M. M. Olivier. Improving Access to Eye Care: A Systematic Review of the Literature. *Ophthalmology*, 2022.

[53] S. Sreng, N. Maneerat, K. Hamamoto, and K. Y. Win. Deep Learning for Optic Disc Segmentation and Glaucoma Diagnosis on Retinal Images. *Applied Sciences*, 10(14):4916, 2020.

[54] Y. Sun, G. Yang, D. Ding, G. Cheng, J. Xu, and X. Li. A GAN-based Domain Adaptation Method for Glaucoma Diagnosis. In *2020 International Joint Conference on Neural Networks*, pages 1–8. IEEE, 2020.

[55] D. S. W. Ting, C. Y.-L. Cheung, G. Lim, G. S. W. Tan, N. D. Quang, A. Gan, H. Hamzah, R. Garcia-Franco, I. Y. San Yeo, S. Y. Lee, E. Y. M. Wong, C. Sabanayagam, M. Baskaran, F. Ibrahim, N. C. Tan, E. Finkelstein, E. Lamoureux, I. Wong, N. Bressler, S. Sivaprasad, R. Varma, J. Jonas, M. G. He, C.-Y. Cheng, G. C. M. Cheung, T. Aung, W. Hsu, M. L. Lee, and T. Y. Wong. Development and Validation of a Deep Learning System for Diabetic Retinopathy and Related Eye Diseases Using Retinal Images From Multiethnic Populations With Diabetes. *JAMA*, 318(22):2211–2223, 2017.

[56] D. S. W. Ting, L. R. Pasquale, L. Peng, J. P. Campbell, A. Y. Lee, R. Raman, G. S. W. Tan, L. Schmetterer, P. A. Keane, and T. Y. Wong. Artificial intelligence and deep learning in ophthalmology. *British Journal of Ophthalmology*, 103(2):167–175, 2019.

[57] L. Van der Maaten and G. Hinton. Visualizing Data using t-SNE. *Journal of Machine Learning Research*, 9(11), 2008.

[58] J. Wang, C. Lan, C. Liu, Y. Ouyang, T. Qin, W. Lu, Y. Chen, W. Zeng, and P. Yu. Generalizing to Unseen Domains: A Survey on Domain Generalization. *IEEE Transactions on Knowledge and Data Engineering*, 2022.

[59] Wikipedia. ImageNet. https://en.wikipedia.org/wiki/ImageNet, 2023.

[60] M. J. Willemink, W. A. Koszek, C. Hardell, J. Wu, D. Fleischmann, H. Harvey, L. R. Folio, R. M. Summers, D. L. Rubin, and M. P. Lungren. Preparing Medical Imaging Data for Machine Learning. *Radiology*, 295(1):4–15, 2020.

[61] L. B. Williams, S. G. Prakalapakorn, Z. Ansari, and R. Goldhardt. Impact and Trends in Global Ophthalmology. *Current Ophthalmology Reports*, 8:136–143, 2020.

[62] World Health Organization. Resolution WHA 73.4: Integrated people-centred eye care, including preventable vision impairment and blindness. 2020.

[63] J. Xiong, A. W. He, M. Fu, X. Hu, Y. Zhang, C. Liu, X. Zhao, and Z. Ge. Improve Unseen Domain Generalization via Enhanced Local Color Transformation. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2020: 23rd International Conference, Lima, Peru, October 4–8, 2020, Proceedings, Part II 23*, pages 433–443. Springer, 2020.

[64] Q. Xu, R. Zhang, Y. Zhang, Y. Wang, and Q. Tian. A Fourier-Based Framework for Domain Generalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14383–14392, 2021.

[65] A. Zhang, L. Xing, J. Zou, and J. C. Wu. Shifting machine learning for healthcare from development to deployment and from models to data. *Nature Biomedical Engineering*, 6(12):1330–1345, 2022.

[66] H. Zhang, N. Dullerud, L. Seyyed-Kalantari, Q. Morris, S. Joshi, and M. Ghassemi. An empirical framework for domain generalization in clinical settings. In *Proceedings of the Conference on Health, Inference, and Learning*, pages 279–290, 2021.

[67] Z. Zhang, B. H. Lee, J. Liu, D. W. K. Wong, N. M. Tan, J. H. Lim, F. Yin, W. Huang, H. Li, and T. Y. Wong. Optic disc region of interest localization in fundus image for Glaucoma detection in ARGALI. In *2010 5th IEEE Conference on Industrial Electronics and Applications*, pages 1686–1689. IEEE, 2010.

[68] Z. Zhang, F. S. Yin, J. Liu, W. K. Wong, N. M. Tan, B. H. Lee, J. Cheng, and T. Y. Wong. ORIGA-light: An online retinal fundus image database for glaucoma analysis and research. In *2010 Annual International Conference of the IEEE Engineering in Medicine and Biology*, pages 3065–3068. IEEE, 2010.

[69] C. Zhou, J. Ye, J. Wang, Z. Zhou, L. Wang, K. Jin, Y. Wen, C. Zhang, and D. Qian. Improving the generalization of glaucoma detection on fundus images via feature alignment between augmented views. *Biomedical Optics Express*, 13(4):2018–2034, 2022.

[70] K. Zhou, Z. Liu, Y. Qiao, T. Xiang, and C. C. Loy. Domain Generalization: A Survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022.