Independent University **Bangladesh (IUB)**

IUB Academic Repository

Computer Science and Engineering | Undergraduate Thesis

2025-12

Ensemble Deep Learning for Retinal Disease Detection from Fundus Images: A Bilateral Ensemble Approach for Retinal Disease Classification

Shanto, Abdullah Al Alam

IUB

https://ar.iub.edu.bd/handle/11348/1068

Downloaded from IUB Academic Repository



Ensemble Deep Learning for Retinal Disease Detection from Fundus Images: A Bilateral Ensemble Approach for Retinal Disease Classification

Abdullah Al Alam Shanto (2110897)
Mumtahina Esha (2231514)
Sadia Alam (2110819)

Dissertation submitted in partial fulfillment for the Degree of Bachelor of Science in Computer Science and Engineering

Department of Computer Science & Engineering
School Of Engineering, Technology & Sciences (SETS)
Independent University, Bangladesh

December 2025

Attestation

We understand the nature of plagiarism, and we are aware of the university's strict policy on the matter. We certify that this is the original work done by us. However, others' written work or software used in any part of this project is properly cited following internationally accepted academic guidelines.

Signature:

I. Abdullah Al Alam Shanto:

II. Mumtahina Esha:

III. Sadia Alam:

Evaluation Committee

Supervision Panel

……	……
Supervisor	Co-supervisor

External Examiners

……	……
External Examiner 1	External Examiner 2

Office Use

……	……
Program Coordinator	Head of the Department

……

Director Graduate Studies, Research and Industry Relations

……

Library

Acknowledgment

We are extremely grateful and feel deeply indebted to our supervisor, Dr. Md. Rashedur Rahman. Without his guidance and care, it would not be possible for us to complete this thesis. We are also grateful for getting the expert supervision from Dr. Saadia Binte Alam in the process of conducting our thesis. We express our sincere gratitude to Mr. Siam Tahsin Bhuiyan for providing us with his constant support and technical guidance during our study. We would also like to extend our heartiest gratitude to Ms. Halima Khatun for her cooperation together with Mr. Riyadul Islam for his valuable contributions made throughout the duration of our project.

We would also like to acknowledge that we used Perplexity and Gemini to enhance the understandability and readability of the manuscript. The tools were solely used for refinement and not used for data analysis, result interpretation or image generation.

Letter Of Transmittal

December 14, 2025

To,
Md. Rashedur Rahman
Assistant Professor
Department of Computer Science and Engineering
Independent University, Bangladesh

Subject: Submission of Dissertation on “Ensemble Deep Learning for Retinal Disease Detection from Fundus Images: A Bilateral Ensemble Approach for Retinal Disease Classification.”

Dear Sir,

It is an honor for us to submit our dissertation titled “Ensemble Deep Learning for Retinal Disease Detection from Fundus Images: A Bilateral Ensemble Approach for Retinal Disease Classification.” This report records the work completed during our project, completed under your consistent guidance.

During this project we experimented with various deep learning models with an ocular disease dataset and gained knowledge and experience in constructing a fundus image-based classification method for ocular diseases. The time spent on this project has helped us to get an understanding of computer vision, deep learning model development and the challenges in the ophthalmology domain. This report presents the project's objectives, methods and outcomes as well as the limitations with clarity.

We are grateful to have been able to work with you and your team. We hope the report is adequate and meets your expectations as much as possible. Thank you for your time, guidance, patience and encouragement throughout the entirety of this project.

Yours Sincerely,

Abdullah Shanto
2110897

Mumtahina Esha
2231514

Sadia Alam
2110819

Abstract

Retinal diseases including Age-related Macular Degeneration (AMD), Diabetic Retinopathy (DR), Cataract, and Myopia, are among the most prevalent causes of irreversible vision loss worldwide. Early and accurate diagnosis is vital for effective treatment and management, yet manual assessment of retinal fundus images is time-consuming, subject to inter-observer variability, and often challenging due to subtle pathological features. Motivated by these clinical demands, we experimented with the classification of retinal diseases using deep learning models, starting from conventional deep learning methods and advancing towards a bilateral ensemble solution.

The first phase of our experimentations involved the use of convolutional neural networks and vision transformers for multi-class classification using individual ocular fundus images. Conventional approaches showed promise however they are limited in their capacity to extract complex features specific to each disease and limited to analysis of single eye which can make the diagnostic accuracy of these approaches to be unreliable as single eye diagnosis might not capture the potential correlations of both eyes of the same individual. Such limitations were suggestive in the experimental scenarios where both eyes presented disease cues, and the method was restricted in capturing complementary information.

It is in view of such limitations that the next stage of experimentations has made use of various ensemble learning strategies. Satisfying the need for more potentials, ensemble methods tried to incorporate a better set of features by combining the predictions of several state-of-the-art deep learning network architectures. The comparative analysis examined different decision ensemble techniques as well as feature ensemble techniques, finding out that though the ensemble methods have improved the overall classification task over conventional models, there were still certain problems. Most of all, ensemble strategies were still considering images from each eye as separate entities, ignoring valuable information that are present in both eyes of a subject affected by the same eye disease.

Our major contribution, based on these insights, is the design and validation of a custom bilateral ensemble framework for the classification of retinal diseases. This strategy is found to be unique in combining corresponding fundus images from both eyes of a subject from a ConvNeXt-XLarge backbone with preprocessing and bilateral feature fusion technique. Using the disease features in both eyes together, the proposed framework showed a promising performance for the multiclass disease detection. Comprehensive experiments on the publicly available OIA-ODIR dataset demonstrated that the bilateral ensemble approach overcomes conventional method limitations, delivering improved and more reliable results.

The research establishes a foundation for advanced deep learning in detection of eye diseases. Our work presents that systematic improvements to data presentation and system architecture can lead to better diagnostic performance and improved potential of artificial intelligence based systems in the medical field. These findings support future computational ophthalmology research and practical diagnostic tool development for global eye health.

List Of Publications

1. A. A. Alam Shanto, M. Esha, S. A. Kotha, R. Islam, S. Wasi, S. T. Bhuiyan, R. Rahman and S. B. Alam "Assessing the Impact of Architectural Design in Deep Learning Models for Fundus-Based Ocular Disease Detection" in *Proc. IEEE 2nd International Conference on Computing, Applications and Systems (COMPAS 2025),* Kushtia, Bangladesh, Oct. 23-24, 2025. (Presented)

2. S. A. Kotha, A. A. Alam Shanto, M. Esha, S. T. Bhuiyan, H. Khatun, R. Rahman, A. Islam and S. B. Alam "Learning from Both Eyes: A Bilateral Ensemble Approach for Retinal Disease Classification" in *Proc. 2025 IEEE International Women in Engineering (WIE) Conference on Electrical and Computer Engineering (WIECON-ECE),* Cox's Bazar, Bangladesh, Dec. 21-22, 2025. (Presented)

3. M. Esha, S. A. Kotha, A. A. Alam Shanto, H. Khatun, R. Islam, S. T. Bhuiyan, A. Islam, R. Rahman and S. B. Alam "A Comprehensive Evaluation of Feature-Level and Decision-Level Ensembles for Retinal Disease Detection Using Fundus Images" in *Proc. IEEE International Conference on Engineering and Frontier Technologies (ICEFronT 2026),* Tangail, Bangladesh, Jun. 25-26, 2026. (Submitted)

Table Of Contents

CHAPTER 1: INTRODUCTION 1
1.1 Introduction 1
1.2 Background 2
1.2.1 Retinal Diseases: Clinical Significance 2
1.2.2 Conventional Diagnostic Methods and Clinical Challenges 4
1.3 Motivation 4
1.4 Scope of the Study & Objectives 5
1.5 Contribution 6
1.6 Outline of Thesis 7

CHAPTER 2: LITERATURE REVIEW 8
2.1 Traditional Approaches and Handcrafted Features 8
2.2 Deep Learning in Medical Domain 9
2.3 Deep Learning in Retinal Diseases 9
2.4 Advanced Architectures: Vision Transformers and Ensembles 10
2.5 Research Gaps and Thesis Motivation 11

CHAPTER 3: FUNDAMENTAL OF CONVOLUTIONAL NEURAL NETWORK 12
3.1 Input Layer 12
3.2 Convolutional Layers 12
3.3 Activation Functions: Non-Linearity Introduction via ReLU 12
3.4 Pooling Layers: Dimensionality Reduction and Spatial Invariance 13
3.5 Fully Connected Layers 13
3.6 SoftMax Activation and Probabilistic Output 13

CHAPTER 4: DATASET 14
4.1 Dataset Overview 14
4.2 Dataset Categories 15
4.3 Dataset Format and Structure 16
4.4 Data Splitting and Class Distribution 16
4.5 Data set Strengths and Characteristics 17

CHAPTER 5: METHODOLOGY 18
5.1 Preprocessing 18
5.2 Methodology 18
5.2.1 Experimentation of Conventional Architectures 18
5.2.1.1 Training Parameters of Conventional Architecture 21

5.2.2 Exploration and Analysis of Ensemble Methods 22
5.2.2.1 Training Parameters of Decision and Feature Ensembles 25
5.2.3 Bilateral Ensemble (Proposed Method) 25
5.2.3.1 Proposed Method Preprocessing 25
5.2.3.3 Training parameters of Proposed Method 27
5.3 Evaluation Metrics 28
5.3.1 Confusion Matrix 28
5.3.2 Accuracy 28
5.3.3 Precision 29
5.3.4 Recall 29
5.3.5 Specificity 29
5.3.6 F1 Score 29
5.3.7 AUROC 29
5.3.8 Grad-CAM 30
CHAPTER 6: RESULTS AND DISCUSSION 31
6.1 Results and Discussion of Conventional Architectures 31
6.2 Results and Discussion of Decision and Feature Ensembles 36
6.2.1 Results and Discussion of Feature Ensembles 36
6.2.2 Results and Discussion of Decision Ensembles 38
6.3 Results and Discussion of Proposed Bilateral Ensemble Method 42
CHAPTER 7: CONCLUSION 45
7.1 Summary 45
7.2. Limitations 46
7.3. Future Work 46
References **47**

List of Figures

Figure 1.1: Normal Eye VS Diabetic Retinopathy 2
Figure 1.2: Healthy Lens VS Lens with Cataract 3
Figure 1.3: Healthy Eye VS Eye with AMD 3
Figure 1.4: Normal Vision VS Vision with Myopia 4
Figure 4.1: Sample Images of Eight Classes of OIA-ODIR: (a)A: Age related Macular Degeneration (AMD); (b) C: Cataract; (c)M: Myopia; (d) DR: Diabetic Retinopathy; (e) N:Normal; (f) O:Other; (g) G:Glaucoma; (h) H:Hypertension. 14
Figure 5.1: Conventional Method (Single Eye Classification) 19
Figure 5.2: ConvNeXt Base Architecture 20
Figure 5.3: EfficientNet Base Architecture 20
Figure 5.4: ViT Base Architecture 21
Figure 5.5: MobileNetV3 Base Architecture 22
Figure 5.6: Overview of Feature Ensemble Method 23
Figure 5.7: Overview of Decision Ensemble Method 24
Figure 5.8: Sample Preprocessed Images. (a) Raw; (b) Color Normalized 25
Figure 5.9: Overview of the Proposed Bilateral Ensemble Method 27
Figure 5.10: Confusion Matrix Representation 28
Figure 6.1: ROC Curve of ConvNeXt-XLarge 32
Figure 6.2: ROC Curve of EfficientNetB3 33
Figure 6.3: ROC Curve of ViT-B-16 34
Figure 6.4: Grad-CAM Visualization. (a) EfficientNetB3; (b) ConvNeXt-XLarge; (c) ViT-B-16 35
Figure 6.5: ROC Curves of Feature Ensemble. (a) Multiplication; (b) Summation; (c) Concatenation 38
Figure 6.6: ROC Curves of Decision Ensemble (a) Probability Average (b) Max Rule (c) Weighted Probability Average (d) Majority Voting 41
Figure 6.7: ROC Curve of Bilateral Method Using Color Corrected Images 43

List of Tables

Table 4.1: Description of Dataset Classes 15
Table 4.2: Images Per Class 16
Table 6.1: Images Per Class After Data Split 31
Table 6.2: Class Wise Evaluation Scores of ConvNeXt-XLarge 31
Table 6.3: Class Wise Evaluation Scores of EfficientNetB3 32
Table 6.4: Class Wise Evaluation Scores of ViT-B-16 33
Table 6.5: Images Per Class After Data Split 36
Table 6.6: Class Wise Evaluation Scores of Multiplication Merging 36
Table 6.7: Class Wise Evaluation Scores of Concatenation Merging 37
Table 6.8: Class Wise Evaluation Scores of Summation Merging 37
Table 6.9: Class Wise Evaluation Scores of ConvNeXt-XLarge 38
Table 6.10: Class Wise Evaluation Scores of EfficientNetB3 38
Table 6.11: Class Wise Evaluation Scores of MobileNetV3Large 39
Table 6.12: Class Wise Evaluation Scores of Majority Voting 39
Table 6.13: Class Wise Evaluation Scores of Max Rule 40
Table 6.14: Class Wise Evaluation Scores of Weighed Probability Average 40
Table 6.15: Class Wise Evaluation Scores of Probability Average 40
Table 6.16: Images Per Class After Data Split 42
Table 6.17: Class Wise Evaluation Scores with Color Normalizing Bilateral Ensemble Method 42
Table 6.18: Class Wise Evaluation Scores Without Color Normalizing Bilateral Ensemble Method 44

CHAPTER 1: INTRODUCTION

1.1 Introduction

Retinal diseases are a significant global health problem and are among the primary causes of visual impairment and permanent blindness around the world. According to the World Health Organization (WHO), at least 2.2 billion people globally have some form of vision impairment, and a large portion of this is due to retinal conditions that are either preventable or treatable [1]. The most common causes of this vision loss include diabetic retinopathy (DR), age-related macular degeneration (AMD), cataract, and myopia [1]. These conditions are more likely to affect older individuals and those with chronic metabolic diseases. These diseases often progress silently in their early stages without any symptoms. Timely detection and intervention for such diseases is an urgent need, necessary for prevention of permanent damage to the visual system.

Traditional methods, albeit dependable, come across major difficulties in keeping up with the increasing demand for widespread eye screening services. The manual review of color fundus photographs requires a significant amount of time and specialized clinical skills. This process is also prone to noticeable differences in judgment, both between different experts (inter-observer) and by the same expert at different times (intra-observer) [2]. The demand for expert ophthalmologists is ever growing, especially in low and middle income nations as the population is growing [3]. Additionally, standard methods of analyzing only one eye at a time often miss bilateral asymmetries (differences between the two eyes) and other related inter-ocular patterns. These are clinically important indicators that eye doctors regularly use when making a diagnosis [4].

The emergence of deep learning, particularly convolutional neural networks (CNNs), has significantly transformed automated medical image analysis and shown remarkable promise for diagnosis in ophthalmology. A study by Elhokly et al. [5] used CNNs and OCT images for detection of healthy, Diabetic Macular Edema (DME), Choroidal Neovascular Membranes (CNM), and Age-related Macular Degeneration (AMD) reported a 97% accuracy while another study by Shoukat et al. [6] reported 98% accuracy on four different glaucoma datasets. Recent state-of-the-art (SOTA) architectures, including ConvNeXt [7], EfficientNet [8], and Vision Transformers (ViTs) [9], have achieved diagnostic accuracy rates above 90% in various retinal disease classification tasks [10]. However, most current methods analyze each eye in isolation and use inconsistent image preparation strategies. This approach, as a result, limits their ability to use information from both eyes and to detect the subtle differences between the eyes that ophthalmologists regularly consider during a clinical diagnosis [4], [10].

This thesis addresses these specific limitations by providing a comprehensive investigation of bilateral ensemble learning approaches for classifying multiple retinal diseases. The created bilateral ensemble system establishes a framework which enables researchers to detect distinct visual characteristics that exist between the two human eyes. The system processes both eyes

through two separate network paths which combine their extracted features. This method is intended to emulate the complete diagnostic reasoning used by experienced clinicians.

1.2 Background

1.2.1 Retinal Diseases: Clinical Significance

There is a heterogenous range of eye diseases that affect the light-sensitive neural tissue of the posterior pole of the eye as a whole, ending in a progressive impairment of vision that occurs if untreated. This section gives an overview of the 4 major pathologies affecting the retina that are studied in our research.

The first, diabetic retinopathy (DR) that can be seen in Figure 1.1, is a microvascular complication of diabetes mellitus and it is one of the most common causes of blindness in working age adults in the developed nations [11]. Its pathophysiology consists of chronic hyperglycemia induced damage in the blood vessels in the retina leading to the precipitation of increase in the vascular permeability, capillary occlusion, ischemia and neovascularization in the advanced proliferative stages [12]. In the United States, it is estimated that of the 26.4% of patients with diabetes, some form of DR is present [13]. The disease occurs in progression through clinically defined stages ranging from mild non-proliferative (NPDR) to proliferative (PDR). Early detection is extremely important as urgent intervention using panretinal photocoagulation [14] or anti-vascular endothelial growth factor (anti-VEGF) treatment [15] can reduce the potential for severe vision loss.

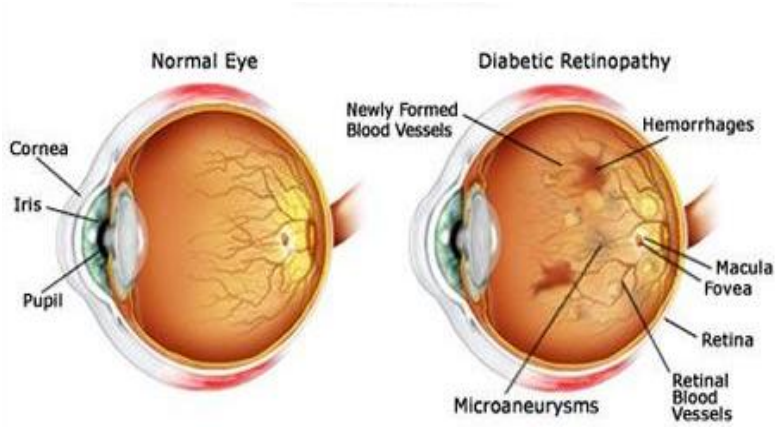

Figure 1.1: Normal Eye VS Diabetic Retinopathy [16]

The second pathology, cataract, is characterized by the progressive opacification of crystalline lens, causing progressive vision deterioration which can be seen in Figure 1.2, representing the leading cause of reversible blindness globally. The Global Burden of Disease (GBD) study estimated that in 2020 17.0 million people were blind from cataract, and this represented 39.6% of the total cases of blindness worldwide [17]. Despite the amenability to treatment through surgical lens extraction and intraocular lens implantation, there was a 29.7% increase in the absolute numbers of people who are blind from cataract solely from 1990 to 2020, and this has been mainly due to the ageing population [17].

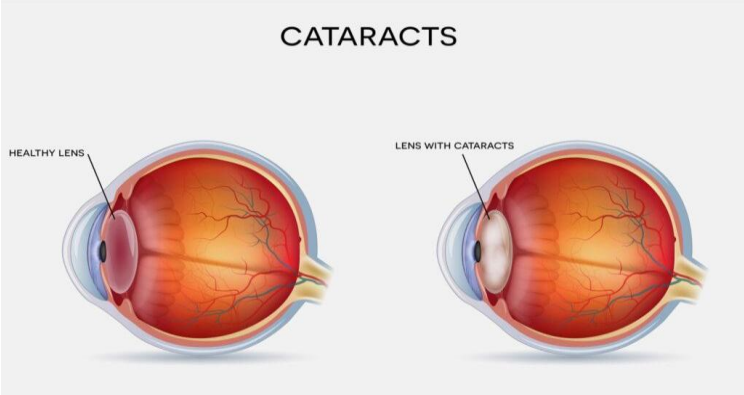

Figure 1.2: Healthy Lens VS Lens with Cataract [18]

A third significant condition is age related macular degeneration (AMD), which is a progressive degenerative disorder of the macula which can be seen in Figure 1.3 which is a leading cause of severe irreversible vision impairment in those aged 50 years and older [19]. AMD affects an estimated 196 million people worldwide with epidemiological projections showing that this global burden will increase to around 288 million by 2040 [20]. The disease is clinically classified into two main phenomenology forms: dry (atrophic) AMD, corresponding to 85-90% of cases and wet (neovascular) AMD, corresponding to most of the severe vision loss [19].

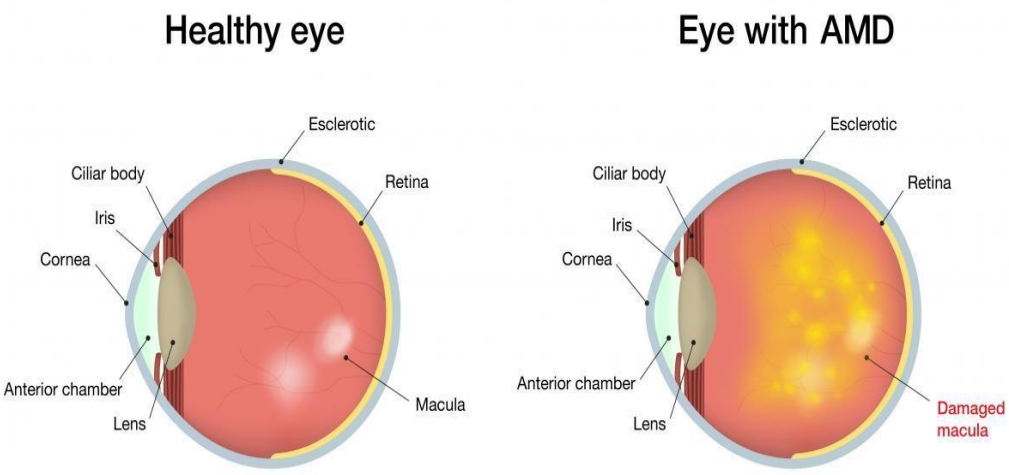

Figure 1.3: Healthy Eye VS Eye with AMD [21]

Finally, myopia or nearsightedness is a refractive error in which the light rays are converging in front of the retina because of too much axial elongation of the globe that is visualized in Figure 1.4, therefore, distant objects are blurred. While pathological or high myopia is related to axial elongation, some cases in which myopia is related to corneal curvature abnormalities [22]. The prevalence of myopia is now world-wide at epidemic levels and epidemiological projections suggest that by 2050 about 50% of the world's population (almost 5 billion individuals) will be myopic [23]. High myopia is of particular concern because of its strong link to pathological complications of the eye, such as rhegmatogenous retinal detachment, myopia maculopathy and open angle glaucoma [24].

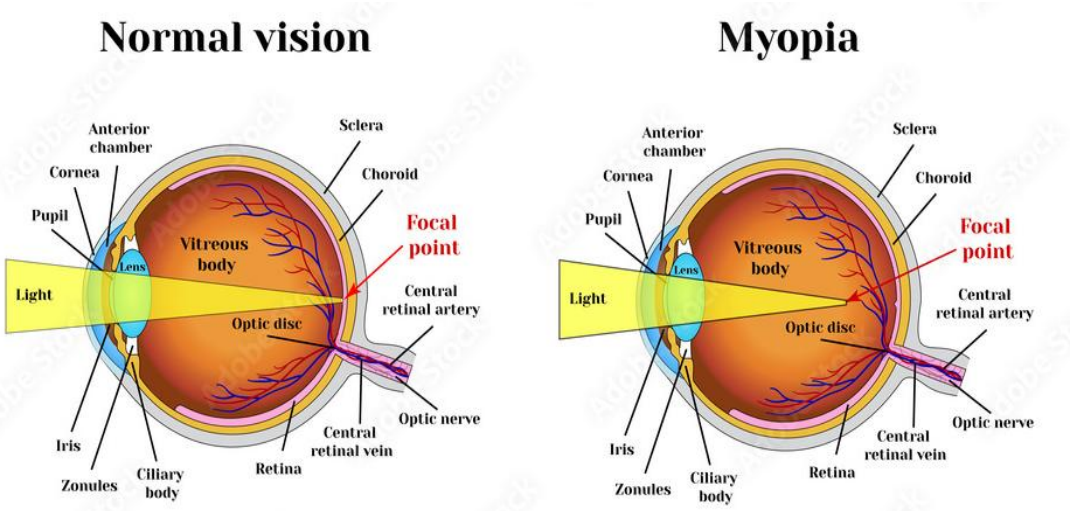

Figure 1.4: Normal Vision VS Vision with Myopia [25]

1.2.2 Conventional Diagnostic Methods and Clinical Challenges

Color fundus photography remains the cornerstone of retinal disease screening and longitudinal documentation. Expert ophthalmologists manually inspect these images to identify pathological features including microaneurysms, hemorrhages, exudates, drusen, and neovascularization. However, this manual interpretation process is inherently time-intensive, with some analyses suggesting much time for patients for a comprehensive bilateral examination [26] and remains vulnerable to significant inter-observer variability. Studies investigating inter-rater reliability in DR grading have reported Cohen's kappa coefficients ranging from 0.60 to 0.88, indicating only moderate to substantial agreement even among experienced retinal specialists [2], [27].

The global burden of retinal diseases is compounded by a critical shortage of trained ophthalmologists, which is particularly acute in low- and middle-income countries [3]. This workforce crisis is further exacerbated by the increasing prevalence of chronic metabolic diseases, such as the rising global incidence of diabetes mellitus [28], which drives exponential growth in screening demands. The convergence of escalating disease burden, workforce shortages, and diagnostic variability has created a compelling imperative for automated, objective, and scalable diagnostic solutions.

1.3 Motivation

Early and accurate diagnosis is extremely important for effective clinical management, but it's often hindered by the subtlety and complexity of pathological changes within the images of the retina. Early computational approaches, which include traditional image processing and feature based machine learning, greatly increased the reproducibility of the screening but was highly dependent on handcrafted features and manual input hours, which made them less reliable and non-generalizable for real-world settings. Retinal diseases, especially AMD and DR, take the form of progressive disease however early diagnosis can significantly reduce the effects of such diseases and in some cases, it can also avoid irreversible loss of vision [14], [19].

The arrival of deep learning has dramatically advanced the field by enabling models to autonomously extract and learn intricate patterns directly from retinal images. Convolutional neural networks (CNNs) and related architectures now match or surpass expert-level performance for certain detection and classification tasks. However, prevailing models typically process each eye's image in isolation, omitting context that is often crucial for ophthalmologists, who routinely compare and integrate information from both eyes during diagnosis. This single-eye analysis has a limitation as it does not account for the asymmetric or bilateral manifestations of many retinal diseases, potentially unable to capture inter-eye relationships that can provide critical diagnostic clues [4], [29]. Most automated systems developed to date have disregarded this bilateral reasoning process, treating each eye image as an independent entity [10].

The Bilateral ConvNeXt Ensemble framework used in our paper takes inspiration from both clinical practice and computational. The ConvNeXt family of models consists of large kernels which allow them to pick up wide areas in an image while their large number of parameters help them to identify and extract the features of an image. Clinically, ophthalmologists systematically compare left and right eye to look for symmetry, identify asymmetrical progression patterns, and know systemic manifestations [29]. By processing left and right images as two parallel ConvNeXt branches [7] and fusing their deep feature representations, the proposed model is a computational analogy of this human diagnostic behavior.

Furthermore, the motivation stems from clinical and practical challenges confronting ophthalmic practice. An AI-assisted bilateral diagnostic system capable of delivering high-accuracy screening directly from fundus photographs could drastically improve early detection rates [30], reduce clinical workload, ensure consistent diagnostic quality irrespective of geographic location, and enhance telemedicine-based screening in remote, low-resource regions [3].

1.4 Scope of the Study & Objectives

Considering the limitations of conventional single-eye diagnostics the main goal with this research is to design, develop, validate and interpret an explainable, and clinically relevant Bilateral Ensemble Deep Learning Framework. Our approach is devoted to the automated categorization of fundus images into five clinically major diagnostic categories Normal, DR, AMD, Cataract, and Myopia with the publicly available OIA-ODIR dataset [31].

This study involves building a data preprocessing pipeline to ensure the integrity of the bilateral pairs, the use of advanced normalization techniques. The research focuses on the development and training of a novel architecture, called the Bilateral ConvNeXt Ensemble Model using a dual branch architecture [10] incorporating a mechanism called bilateral feature fusion inspired by clinical diagnostic reasoning. The performance of the model is carefully evaluated in terms of various metrics such as accuracy, precision, recall, F1-scores and confusion matrix analysis and the area under the curve (AUC). Extensive testing with strict splits per patient to avoid data leakage from different eyes.

An important part of this study is the use of Explainable AI (XAI) to gain clinical transparency with the help of Grad-CAM visualization [32] to validate attention maps of the models and check if they align with known clinical features. While this work explicitly excludes three-dimensional imaging modalities and temporal modeling of disease progression, the work creates a serious validation of the usefulness of the bilateral model and establishes the basis for more robust and clinically relevant automated screening systems.

1.5 Contribution

Our study makes a systematic contribution to the field of retinal disease classification by introducing a custom bilateral ensemble for fundus image Classification which is inspired by existing bilateral ensemble techniques. Through the use of a staged approach, our study first confirms the limitations of traditional techniques and single eye deep learning methods, that they are unable to fully capture the nature of disease when the analysis is limited to one image of the patient's fundus. It then moves on to more complex ensemble techniques, discussing the prospects for realizing combination strategies of neural networks, but showing that something more than the simple combination of the outputs from different models must be done in order to achieve any meaningful improvement.

The fundamental novelty of our work is the proposal and realization of a paired fundus image processing model for both left and right eyes in parallel, based on a high-capacity convolutional backbone, which is coupled with data pre-processing. This novel analytical paradigm mimics the method applied by clinicians to compare both eyes to identify asymmetric and cross-ocular patterns, as opposed to considering each eye separately [28]. A major architectural contribution is the adaptation and modification of the ConvNeXt backbone [10] to work with fundus-based disease recognition. This makes use of its modernized CNN design to offer its greatest contextual learning and performance that is subjectively benchmarked against other state-of-the-art models. To ensure a validation of clinical relevance and accountability, a systematic application of Grad-CAM [32] is used. Grad-CAM is a method that directly addresses the most crucial shortcoming of previous work by modeling the across-eye relationships and context that is central to correct clinical interpretation. It is an important step to make the model’s decisions more transparent and trustworthy. Experimental validation using the OIA-ODIR dataset that includes paired images and multiclass disease labels gives good arguments for the feasibility and potential impact of bilateral deep learning that demonstrate that bilateral models can provide richer and more clinically aligned disease assessment.

The findings of this study contribute to development of AIs assisted systems especially for the eye domain, by providing a way that effectively leverage convolution neural networks for proper feature extraction and effective utilization for a reliable classification decision which is needed for diagnosis. Furthermore, this study serves as a methodological foundation for advancement of deep learning based eye disease diagnosis systems by incorporating binocular visual information. By taking complementary disease cues from both eyes, the proposed framework enhances diagnostic reliability and reduces ambiguity present in single-eye analysis.

1.6 Outline of Thesis

The thesis is organized into seven chapters as follows:

Chapter 1: The first chapter presents the background information and clinical significance of retinal diseases which include Age-related Macular Degeneration and Diabetic Retinopathy and Cataract and Myopia. Additionally, the chapter talks about the obstacles of manual diagnosis as well as the challenges faced by methods of single-eye classification which are used in computer-based retinal disease assessment from fundus images. The chapter also discusses the objectives behind the development of the bilateral deep learning framework and its potential contribution to a more comprehensive and reliable diagnostic system.

Chapter 2: The chapter demonstrates the progress of computer aided methods used for diagnosis in the ophthalmology domain. It presents the advancement from single-eye deep learning systems and conventional image processing to hybrid and ensemble methods. It also compares the limitations of single-eye analysis with the potential benefits of integrating bilateral information and identifies the specific research limitations addressed by this thesis.

Chapter 3: The background information on the foundational concepts necessary to understand deep learning have been discussed in this chapter. It explains convolutional neural networks (CNNs), their layer types and operations (convolution, pooling, activations).

Chapter 4: This chapter introduces the OIA-ODIR dataset which was used for the experimentations of this thesis. The chapter entails the details of its structure, the availability of paired left and right fundus images, and multiclass disease labels. The relevance of this dataset to bilateral analysis is explained, and the chapter clarifies how the dataset supports modeling of both unilateral and bilateral ocular conditions.

Chapter 5: This chapter reports the preprocessing steps taken to prepare the images for our proposed method such as the preprocessing steps used to prepare images such as normalization and data splitting. It then describes the design of the bilateral deep learning framework, with a focus on how paired-eye inputs are processed using ConvNeXt-XLarge architecture as the feature extractor and classifier.

Chapter 6: This chapter presents the results obtained from the bilateral framework on the OIA-ODIR dataset, using appropriate evaluation metrics. It analyzes how bilateral modeling addresses the limitations of single-eye approaches, discusses the interpretation of findings, and reflects on the ability of the approach to improve the detection of retinal diseases in a clinically relevant context.

Chapter 7: The concluding chapter gives a summary of the research outcomes and contributions of the proposed method. This chapter discusses the current limitations of the study and the advances made through bilateral disease modeling, also suggests future research directions for improving automated retinal disease diagnosis with deep learning.

CHAPTER 2: LITERATURE REVIEW

The diagnosis of retinal diseases from fundus images has evolved considerably, moving from handcrafted image processing methods to sophisticated deep learning architectures. A structured review tracing the progression of these methods have been discussed systematically in this chapter. The review starts with traditional feature-based techniques and their limitations, then discusses the application of deep learning in general medical imaging and specifically in retinal disease detection. The review then examines ensemble and hybrid approaches that combine multiple models to improve performance. Finally, it identifies the specific gap that remains unaddressed in existing methods which is the lack of explicit bilateral modeling in automated retinal disease classification and positions this gap as the primary motivation for the bilateral ensemble framework developed in our study.

2.1 Traditional Approaches and Handcrafted Features

Traditional diagnostic approaches such as expert inspection of color fundus photographs and optical coherence tomography (OCT) are highly reliable but remain time consuming, resource-intensive, and prone to inter-observer variability [33]. Traditional diagnostic approaches rely heavily on ophthalmologists manually interpreting imaging from three primary modalities: fundus photography, optical coherence tomography (OCT), and fundus fluorescein angiography (FFA). These methods demand significant time and expertise, but a growing imbalance between the number of specialists and the increasing patient population restricts access to timely diagnosis [34]. Such methods require time and attention to detail; however, the patient numbers are ever growing while there is a lack of specialists to tackle them all in time [35]. Classical machine learning with handcrafted features such as texture descriptors, vessel patterns, or histogram-based features have also been investigated [33], yet these methods often fail to generalize across diverse imaging conditions and disease types.

Early signs of diseases like AMD, cataract, DR etc. can be difficult to distinguish and diagnose, as they all show signs in the retinal region [36]. However, such diseases need to be diagnosed early to prevent further damage to the retina and preserve vision. Traditional methods of diagnosis rely on ophthalmologists to manually check fundus images, FFA, and OCT.

However, these traditional methods suffered from significant limitations. These traditional and handcrafted approaches demonstrate clear limitations in scalability, generalization, and timely diagnosis, which motivates automated image-based methods. Their primary weakness was a high dependency on parameters, requiring meticulous tuning for different imaging devices or datasets. These handcrafted rules could not generalize images with diverse illumination or pathological presentations, leading to inconsistent performance. They particularly struggled with complex textures and lacked the ability to learn abstract, hierarchical relationships between features especially for diseases that show overlapping disease cues which hampered their distinguishing ability between diseases, hence showing poor performance. The heavy burden of manual feature engineering and poor robustness to noise created a clear imperative for more adaptive, data-driven paradigms.

2.2 Deep Learning in Medical Domain

In recent years, researchers have delved into automated image-based diagnostic methods to help ophthalmologists and make their tasks more efficient. The development of convolutional neural networks has led to the creation of various CNN models that can extract local features– edges, textures, shapes, and patterns from images and make decisions based on those features [37]. These CNNs have been tested on various modalities, especially in the medical domain, such as the diagnosis of COVID-19 from chest X-ray images [38], kidney abnormalities [39], as well as the detection of intracerebral hemorrhages [40] from computed tomography scans.

The progress of deep learning, particularly convolutional neural networks (CNNs), has significantly advanced automated retinal image analysis. Numerous studies have reported high diagnostic performance across specific disease tasks. Beyond CNNs, explainable architectures such as ResNet-101 and hybrid models such as HOG-CNN have also demonstrated competitive performance, 94% accuracy for ResNet-101 [41] and 98.5% binary DR accuracy for HOG-CNN [42].

However, despite these strong results, deep learning models in medical imaging can still suffer from limitations such as dependence on large, well-annotated datasets, possible domain shift across scanners or institutions, and limited explicit modeling of relationships between multiple related views.

This manual process was labor-intensive and influenced by human intuition, which often resulted in a failure to capture the full complexity of disease indicators. Furthermore, these models became less effective and difficult to scale as datasets grew larger in size. Finally, they lacked the capacity to learn hierarchical representations and the ability to build complex features from simple ones. This capability is critical for distinguishing subtle signs of pathology, and its absence necessitated the transition toward end-to-end deep learning

2.3 Deep Learning in Retinal Diseases

In the ophthalmology domain, CNNs have demonstrated remarkable performance. Cataract classification achieved more than 90% accuracy [43] and diabetic retinopathy classification with an accuracy over 95% [44]. CNNs' ability to capture fine-grained image details makes them particularly well-suited for medical image analysis [45]. Recently, CNNs have been applied in ophthalmology to tasks. In ophthalmology, CNNs have demonstrated strong performance in classifying ocular diseases from fundus images.

An optimized CNN for DR and diabetic macular edema achieved 97.91% accuracy, 97.82% sensitivity, and 98.64% specificity [46]. In multi-class classification, Wang et al. used fusion features and a customized 17-layer CNN on 17,766 fundus images across eight disease types, reporting 93.0% accuracy and a Kappa of 0.92 [47]. On ultra-wide-field fundus images, ResNet-152 achieved an AUC of 96.47% (95% CI: 0.931–0.974) [48]. Similarly, a ConvNeXt model trained for macular degeneration classification reached 96.89% accuracy and a quadratic Kappa of 94.99% [49].

Despite high reported performance, many of these CNN-based ocular disease studies still treat each image independently, commonly focusing on single-eye or single-modality inputs, without explicitly modeling bilateral relationships or complementary information between left and right eyes.

2.4 Advanced Architectures: Vision Transformers and Ensembles

To address the limitations of standard CNNs, researchers explored more advanced architecture. Vision Transformers (ViTs) [7] marked a major shift, using self-attention mechanisms to model global dependencies across an image. This allows ViTs to capture long-range spatial interactions, such as the diffuse, scattered patterns seen in AMD, which can be missed by the localized receptive fields of CNNs. Architectures like the Swin Transformer [50] further optimized this by introducing hierarchical window-based attention. Recognizing the complementary strengths of both paradigms, hybrid CNN-ViT models were developed to fuse the local feature precision of CNNs with the global contextual awareness of transformers.

Alongside hybrid models, ensemble learning emerged as a powerful strategy for enhancing robustness. For a more reliable result, researchers took the ensemble approach. An ensemble uses a group-thinking approach where more than one model is involved in the decision-making process [51]. A prior ensemble study has reported an accuracy of 98.8% in the classification of pancreatic cancer using CT scans [52]. Additionally, ensembles have also shown promising results in the eye disease domain of 80.5% accuracy in 8-class classification from fundus images [53]. Prior studies have concentrated on feature-level ensembles and have received remarkable results.

In the multi-label setting, the bilateral feature enhancement network (BFENet) leveraged both eyes to classify ophthalmic diseases, reporting an F1 score of 0.892 and AUC of 0.912 on the off-site test set [54]. Recently, methods that process both eyes simultaneously have demonstrated improved performance in multi-class retinal disease classification. G. Huo et al. introduced a single custom architecture that inputs both left and right eye images, using a dual-modal multi-scale Siamese design to jointly learn bilateral features [53]. DMS-Net achieved 92.6% accuracy and an average F1 score of 0.89 across multiple retinal disease categories, showing the benefits of leveraging bilateral information directly within a single network [53]. Similarly, another approach employs a two-backbone architecture with shared weights to process both eyes concurrently, extracting complementary features while maintaining high generalization. This approach reported an overall accuracy of 91.8% and an AUC of 0.93 on a multi-disease dataset, demonstrating that dual-eye processing can improve robustness and diagnostic reliability [34].

Despite these advances, most ensemble and hybrid approach still either operate on single-eye inputs or fuse both eyes within a single tightly coupled architecture. This leaves open questions about how to best balance shared information and eye-specific representation and motivates new designs that can explicitly model complementary and asymmetric patterns between the two eyes. Such designs could further enhance diagnostic robustness by selectively emphasizing inter-eye correlations while preserving unique disease markers present in individual eyes.

2.5 Research Gaps and Thesis Motivation

Despite improvement from the more traditional approaches to deep ensemble architectures, one glaring and consistent omission is the inability to model the optical bilaterality of ophthalmic diagnosis. In the clinical setting, ophthalmologists never assess an eye in complete isolation, and they compare the two eyes systematically in terms of symmetry, asymmetric progression, and systemic manifestations of disease [2]. Existing models, in which each eye is counted as a separate sample, throw away this rich source of information for diagnosis. To fix this flaw, we propose a new framework of bilateral ensemble.

In comparison to the currently existing models that use dual eyes, our proposed method differs from both existing ones. Rather than treating both eyes in a unified architecture, or sharing a set of weights, we explicitly exploit the natural feature similarity of the two eyes by obtaining features separately by using two instances of the backbone network. These features are then integrated using a new ensemble strategy which maintains independent representations, while making full use of complementary and asymmetric patterns.

This design makes it possible for the model to retain eye-specific detail while still benefiting from the context bilaterally, overcoming the limitations of single architecture and shared-weight dual-eye models. By framing the problem as thus, our strategy directly addresses the shortcoming that is still identified in the literature: the fact that there are methods that exploit information from both eyes and keep the flexibility to model the differences between them and improve the reliability of the process for automated multi-class retinal disease classification.

CHAPTER 3: FUNDAMENTAL OF CONVOLUTIONAL NEURAL NETWORK

A Convolutional Neural Network (CNN) is a specialized type of neural network designed for analyzing visual data. It is widely used in image processing tasks such as classification, recognition, and object detection. The deep learning CNN models' architecture consists of different layers that facilitate to the whole process needed for their tasks, each layer serves a distinct function to help the network understand and interpret image features effectively. This hierarchical arrangement of all the layers enables CNNs to progressively learn intricate visual patterns from the features present in the data.

3.1 Input Layer

The input layer of a CNN takes the values of the raw image as an input, consisting of pixels. It keeps the same dimensions as the original image. 224 by 224 pixels with 3 color channels for an RGB image. Before being fed into the network the pixel values are typically normalized to a range from 0 to 1 in order to assist the model to train itself better. Image is often processed in batches, thus the input shape also must cater for batch size. Additionally, for the sake of making the model more robust, techniques similar to data augmentation such as rotating or flipping the images are usually applied during training to increase the variety of the input data. This layer sort of prepares the image data for the feature extraction steps followed in CNN. Appropriate preparation of inputs during this stage has a major impact on the overall network performance and convergence speed.

3.2 Convolutional Layers

The Convolutional Layer is the heart of a CNN. It utilizes filters or kernels that slide across the input image in order to obtain important image features such as edges, textures, and shapes. The convolution operation involves multiplication of the filter values with the pixel values of the input and adding them up to get a feature map. This layer is used to reduce the spatial dimensions of the input based on the stride (which controls filter movement steps), and padding (which can add pixels on the boundaries of images to maintain the same size of the output). The convolutional layer is used to detect patterns and preserve spatial information important in order to understand the image. Multiple convolutional layers are stacked sequentially for the network to extract increasingly abstract and complex visual representations.

3.3 Activation Functions: Non-Linearity Introduction via ReLU

The Activation Layer applies a non-linear function to the feature maps to introduce non-linearity into the network. Without this, CNNs would be limited to learning only linear patterns. The most common activation function is the Rectified Linear Unit (ReLU), which outputs the input value if it is positive, and zero otherwise. This activation speeds up training and helps prevent the vanishing gradient problem, allowing deeper networks to learn effectively. Other activation functions include sigmoid and hyperbolic tangent (tanh), though ReLU remains the

dominant choice. The choice of suitable activation functions has a direct effect on the model's ability to approximate complex decision boundaries.

3.4 Pooling Layers: Dimensionality Reduction and Spatial Invariance

After the convolution and activation layers, next comes the Pooling layer. This layer decreases the spatial size of feature maps, which reduces the amount of computation and helps prevent overfitting. There are different kinds of pooling, these include max pooling, average pooling, and L2 pooling. Max pooling selects the maximum value within a small region of the feature map, effectively keeping the most prominent features while discarding less important information. Pooling also provides translation invariance which ensures that the network recognizes features irrespective of their position changes within the image.

3.5 Fully Connected Layers

Dense Layers or fully connected layers take the high-level features extracted by previous layers, flatten them into a one-dimensional vector, and perform linear combinations followed by activation functions. This process enables the network to make a comprehensive feature combination for the final predictions. The fully connected layers function similarly to traditional neural networks and are essential for classification tasks. The general linear function is:

$$y = W \times X + b \tag{1}$$

where W represents the weight matrix, X is the input vector (flattened features), and b is the bias vector. The output y is then often passed through an activation function.

3.6 SoftMax Activation and Probabilistic Output

The Softmax Layer often serves as the output layer in CNNs used for multi-class classification. It transforms the dense layer's raw score outputs into normalized class probabilities that sum to one. The class with the highest probability becomes the network's prediction. Mathematically, SoftMax for an output x_i is computed as:

$$softmax(x_i) = \frac{e^{x_i}}{\sum_j \quad e^{x_j}} \tag{2}$$

where the denominator sums over all classes j. This function ensures output values are interpretable as probabilities.

CHAPTER 4: DATASET

This chapter describes the main data set used in this research, Ocular Disease Intelligent Recognition (OIA-ODIR) benchmark. OIA-ODIR is a massive dataset which is publicly available and was created with the purpose of promoting research in the detection of ocular diseases in an automated way [31]. Released by the Nankai University in cooperation with several medical institutions, the dataset is especially intended to reflect what occurs in clinical diagnoses, where ophthalmologists analyze images of both eyes (left and right) and often assign multiple disease labels to a patient [31].

4.1 Dataset Overview

The OIA-ODIR dataset was curated to address the shortcomings of previous datasets for ophthalmological tasks in that they focused on single disease classification or only used monocular images. It comprises 10,000 high-quality-fundus images from 5,000 patients with each patient record comprising a pair of left and right eye images. This structure is basic to the current research because it provides the capability of simultaneous multi-disease detection and bilateral feature correlation analysis.

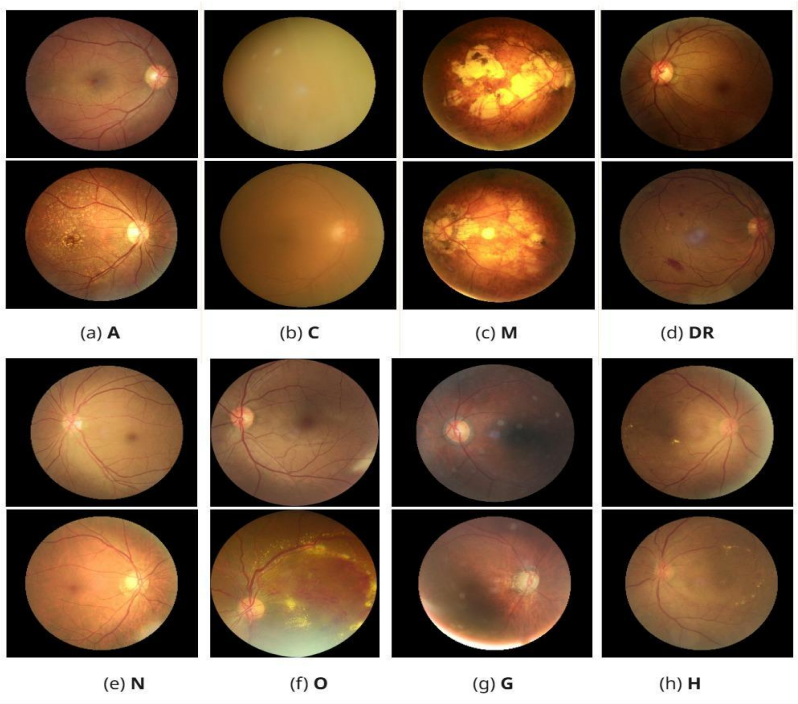

Figure 4.1: Sample Images of Eight Classes of OIA-ODIR: (a)A: Age related Macular Degeneration (AMD); (b) C: Cataract; (c)M: Myopia; (d) DR: Diabetic Retinopathy; (e) N:Normal; (f) O:Other; (g) G:Glaucoma; (h) H:Hypertension.

The dataset is based on a large clinical database of more than 1.6 million fundus images which were collected in 487 hospitals from 26 different provinces in China [31]. Images were acquired with different fundus cameras that have different resolutions, illumination and settings. This process introduces a large amount of variance of color tone, brightness and image

quality, and reflects the heterogeneity of the actual situation of screening. Each image was annotated by a team of expert ophthalmologists with any disagreements resolved by a consensus-based adjudication process. The careful validation process makes this dataset clinically acceptable and diagnostically reliable.

The key characteristics of the dataset have been given below:

- **Total Images:** 10,000 (5,000 left–right eye pairs)
- **Total Patients:** 5,000
- **Annotations:** Eight disease categories, allowing multi-label cases per image
- **Data Type:** RGB color fundus photographs
- **Labeling Process:** Validated by a panel of junior and senior ophthalmologists with consensus validation.
- **Geographic Coverage:** 487 hospitals across 26 provinces in China [31].

4.2 Dataset Categories

A binocular image pair from each patient is annotated under one or more of the following eight diagnostic categories given in Table 4.1, which represent a spectrum of common eye disease seen in clinical practice [31]

Table 4.1: Description of Disease Classes

Label	Disease Name	Diagnostic Description
N	Normal	No visible pathological features; healthy fundus.
D	Diabetic Retinopathy (DR)	Microaneurysms, hemorrhages, hard/soft exudate, or neovascularization.
G	Glaucoma	Increased optic cup-to-disc ratio, neuro retinal rim thinning, or nerve fiber layer defects.
C	Cataract	Lens opacity affects image clarity and fundus structure visibility.
A	Age-related Macular Degeneration (AMD)	Drusen accumulation and pigmentary changes in the macular area.
H	Hypertension	Arteriovenous nicking, vessel narrowing, and hemorrhages associated with hypertensive retinopathy.
M	Myopia	Pathological myopia presenting leopard-spot patterns, choroidal thinning, or optic disc tilt.
O	Other Diseases	Rare or mixed ocular conditions are not classified above.

These categories cover the most common and clinically significant pathological conditions of the retina, enabling the training of generalized and robust diagnostic models. Each disease type displays its own distinct morphological and textural patterns across fundus regions, thereby supporting deep learning models in learning differentiated visual features.

4.3 Dataset Format and Structure

The dataset is arranged into a structured layout which contains metadata and corresponding image folders, containing both single-eye and paired eye input formats.

- **Folder Structure:** The dataset includes RGB fundus photographs (in .jpg format) and a comprehensive metadata file (in .csv format).

- **Metadata File:** Here, patient IDs are linked to their respective left and right eye images, along with demographic data (age, gender) and multi-label diagnostic annotations.

- **Diagnostic Keywords:** Text-based notes provided by ophthalmologists are also included for each image, offering qualitative clinical context.

4.4 Data Splitting and Class Distribution

The OIA-ODIR dataset is pre-divided into three subsets which are organized subjectwise to prevent data leakage between training and testing.

The category-wise image distribution, as detailed in the source publication [31], highlights an inherent class imbalance, a common characteristic in real-world clinical datasets.

Table 4.2: Images Per Class

Category	Normal	DR	Glaucoma	Cataract	AMD	Hypertension	Myopia
Training	1138	1130	215	212	164	103	174
Off-site Test	162	163	32	31	25	16	23
On-site Test	324	327	58	65	49	30	46
Total	**1624**	**1620**	**305**	**308**	**238**	**149**	**243**

A significant observation from the class distribution is the pronounced imbalance [31]. The Hypertension (H) and AMD (A) categories possess the fewest samples, whereas Normal (N) and Other (O) categories are the most represented. This imbalance presents a significant challenge for model generalization and necessitates the use of advanced data balancing, augmentation, or loss-weighting strategies during training. However, it also shows the real

world scenario where some disease cases are rare and collecting sufficient, well-annotated samples remains challenging.

4.5 Data set Strengths and Characteristics

The OIA-ODIR dataset ,used in our study, had the most suitable features which were necessary for the goals of this thesis:

- **Multi-Disease, Multi-Label Annotation:** Each image may contain multiple disease labels, supporting research into complex multi-label classification and the analysis of feature interaction potential for further study.

- **Binocular Data Representation:** This dataset includes paired eye images for every patient, which enabled the bilateral feature correlation model that is proposed in this research, necessary to mimic the clinical diagnostic.

- **Clinical Realism and Diversity:** The dataset is diverse as it contains images aquired from 487 hospitals with varying imaging systems, device type, illumination, and patient demographics. This captures the natural variability and noise present in real-world screenings, promoting the development of robust AI models.

- **Rich Metadata:** The inclusion of metadata such as age, gender, and diagnostic keywords facilitates demographic and clinical cross-analysis.

For our proposed method, we kept the images of subjects that had the same disease in both eyes to leverage the signs of the disease in different stages of progression. The subjects that had more than one disease in both eyes were not taken. As the study evaluates the method's feature extraction and feature differentiating ability, only single labeled images were utilized. Additionally, only five disease classes Normal, Myopia, Age-related Macular Degeneration, Cataract, Diabetic retinopathy, were used from the OIA-ODIR dataset as the disease class glaucoma requires more modalities to understand their disease markers [55] and the hypertension class had a limited sample size in the dataset [56]. The other class was also excluded for this study as it can introduce class heterogeneity that can hamper the model's learning stage [57].

CHAPTER 5: METHODOLOGY

First, we compared well-known pre-trained CNN models and ViT to determine how well they extract information from fundus images and their classification performance for multi-class classification. We then applied Grad-CAM to get a deeper understanding of their decision-making process. From the CNNs, we then picked the backbones that were suitable for our task.

The conventional backbones had their limitations as they missed disease markers and misclassified or confused the classes, revealed by their Grad-CAM and their evaluation metrics.

To overcome such limitations and reach a more reliable decision, we analyzed the CNNs in ensemble settings to understand how each ensemble affects the decisions and feature extraction ability of the models.

Building on the insights, we developed our proposed method which involves the combination of two backbones (ConvNeXt-XLarge) and combination of feature maps using concatenation from a subject's pair-wise left and right eye image containing the same disease in both eyes.

5.1 Preprocessing

The OIA-ODIR [31] dataset contains images of varying sizes ranging from 1444x1444 pixels to 3,456 x 2,304 pixels. The images were resized to 224x224 size to keep it consistent with the size that was given during its pre-training using the ImageNet [58] dataset.

5.2 Methodology

5.2.1 Experimentation of Conventional Architectures

Convolutional Neural Networks and Vision transformers are well known and have shown promising performance on various types of datasets. Each model family of CNNs have different architectures which contribute to their ability for feature extraction from an image.

To establish a performance baseline and thoroughly investigate the capabilities of various modern architectures, we conducted a systematic set of experiments. This study focused on using model variants as standalone feature extractors. The purpose of this experiment was to understand their fundamental feature extraction and their class distinguishing ability.

For this study, we selected model variants from several prominent deep learning families. From the convolutional neural network (CNN) domain, we experimented with architecture from the ConvNeXt family, which represents a modern evolution of CNNs incorporating design principles from transformers. We also included models from the EfficientNet family, which is renowned for its powerful compound scaling and high efficiency. To contrast with these convolutional approaches, we also experimented with models from the Vision Transformer

(ViT) family, which relies on global self-attention mechanisms rather than local convolutions to interpret image data.

In this phase, we utilized the left and right eye images as individual images without any subject-wise relation. The dataset was then split to train, test and validation folders only the single labeled images of each subject were considered for input. A subset of OIA-ODIR dataset was used for this phase of experimentations.

The models were kept frozen to apply effective transfer learning, this was done for this phase of experimentation, as well as in the subsequent phases that followed. This reduced training time and mitigated the risk of overfitting, as unfreezing the models would introduce a large number of additional parameters and potentially result in divergent learning dynamics across architectures.

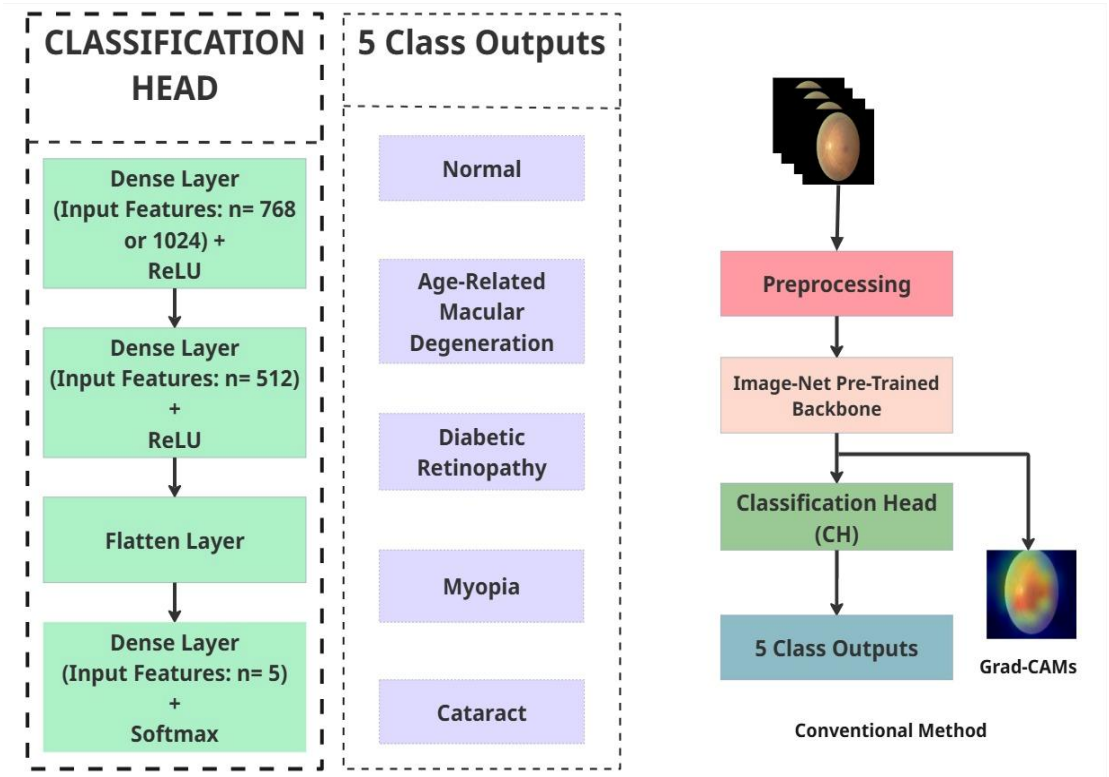

Figure 5.1: Conventional Method (Single Eye Classification)

Among the tested base models, ConvNeXt-XLarge had the best performance, belonging from the family of ConvNeXt [7]. This family of models is inspired by Vision transformer design elements to capture both local and global features. Its architectural layout uses convolutional features to mimic a Vision transformer.

The ConvNeXt-XLarge model contains approximately 350 million parameters. The model consists of ConvNeXt blocks that contain depth wise and pointwise convolutions with 7x7 kernel size, along with layer normalization and GELU (Gaussian Error Linear Unit) activation. The base architecture is shown in figure 5.2 [59].

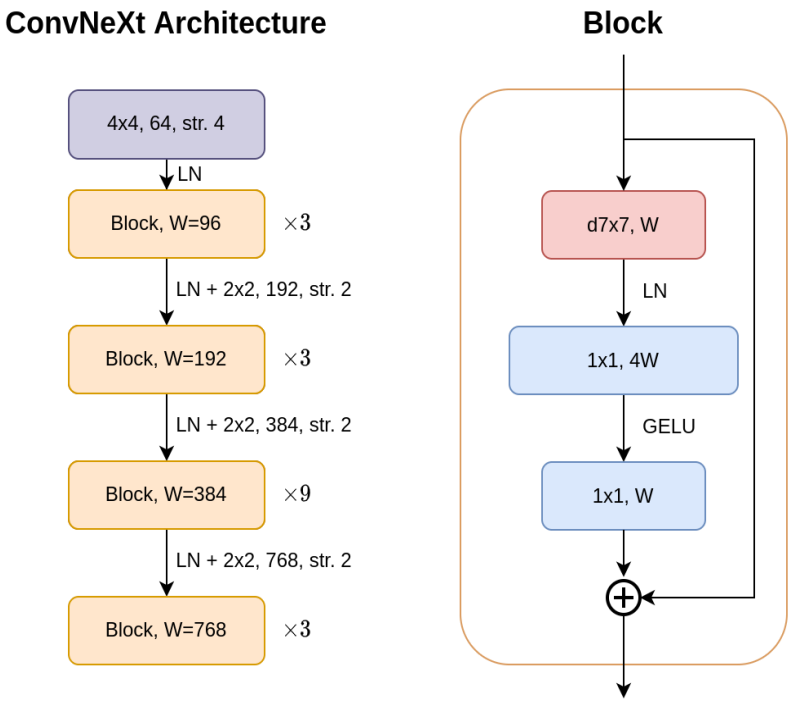

Figure 5.2: ConvNeXt Base Architecture [59]

EfficientNetB3 performed best from the EfficientNet family [6] which is known for its compound scaling method to optimize width, depth, and input resolution for balanced efficiency. The model contains mobile inverted convolutional blocks with squeeze-and-excitation modules to improve its feature extraction, and its global average pooling layer reduces dimension but retains all information. The base architecture is shown in the figure 5.3 [60].

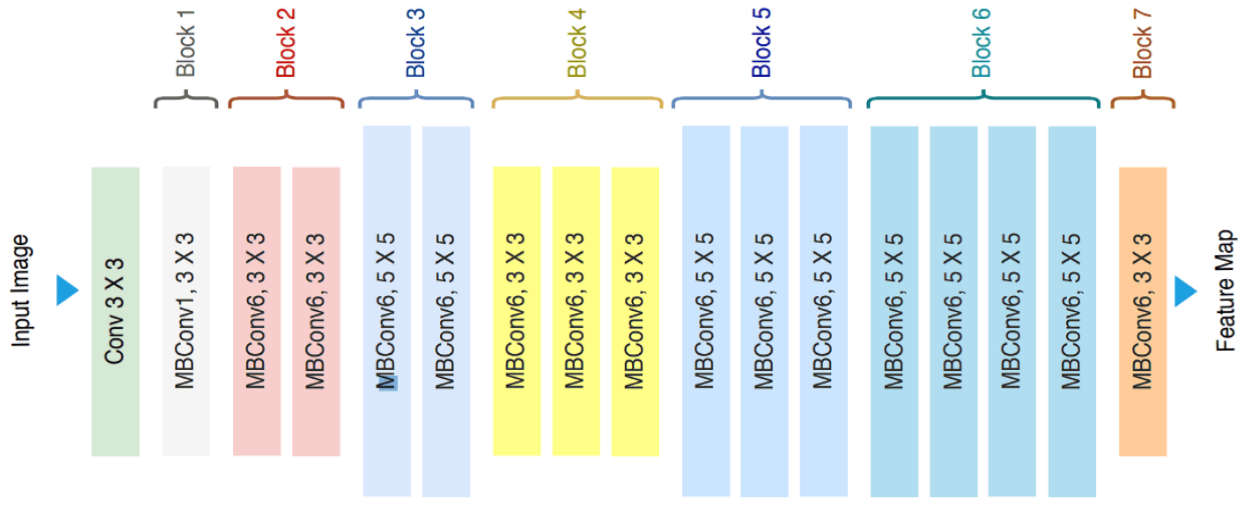

Figure 5.3: EfficientNet Base Architecture [60]

ViT-B-16 performed best among the ViTs. ViTs (Vision transformers) are deep learning models that use transformer architecture for image-based tasks [9]. In ViTs, the input images are divided into patches, the patches are tokenized with positional encodings, and a self-attention mechanism is used to capture information from the image and make predictions. The base architecture is shown in the figure 5.4 [61].

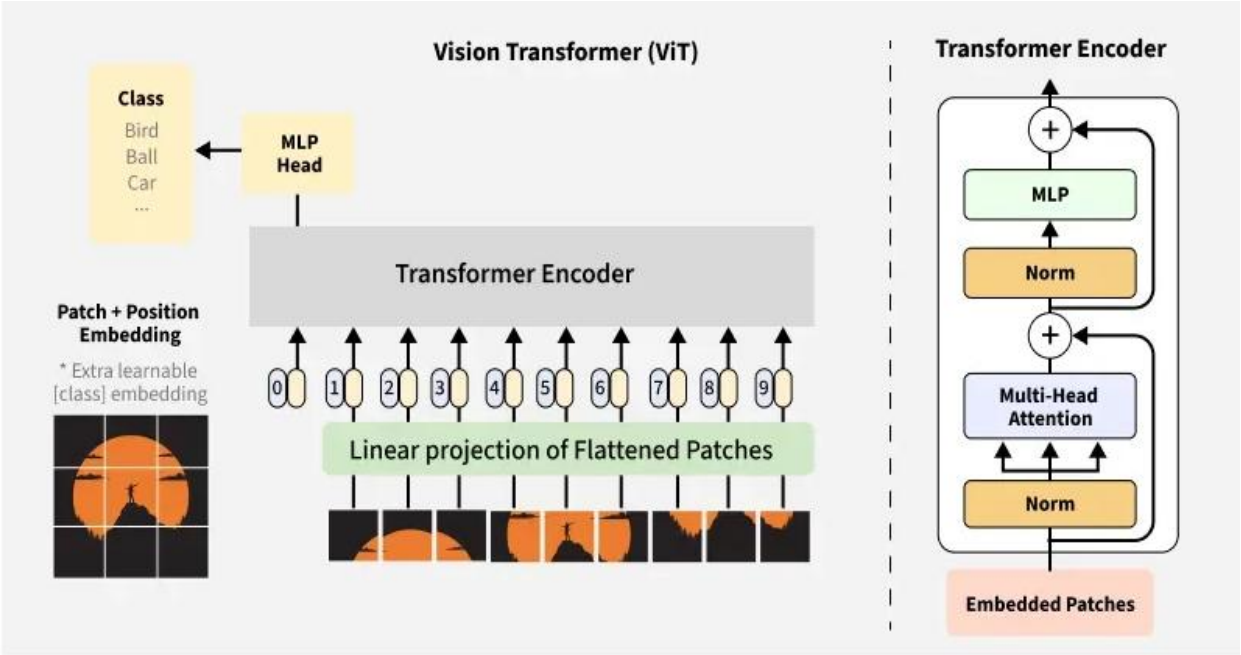

Figure 5.4: ViT Base Architecture [61]

We then applied Grad-CAM to understand the decision-making process of each model and to understand how the architectural differences of the models aid to their feature extraction and classification performance.

Each model was experimented on the ODIR-5K dataset in similar conditions with its backbone kept frozen and same training parameters which may have constricted their potential but also showed their limitations. ViTs use the self-attention mechanism that captures global dependencies, but they miss out on small local features while CNNs capture small local patterns, the overlap of disease markers may have caused its misclassifications. Their Grad-CAMs revealed that ConvNeXt-XLarge was able to focus on broad regions across the image which makes it suitable for extracting a wide range of features. We then experimented with ensemble settings for a more reliable approach as well as to find the best method that can increase feature representation needed for classification of the diseases.

5.2.1.1 Training Parameters of Conventional Architecture

To train the conventional architectures, the models were trained for 40 epochs with a learning rate of 0.0001 and a patience of 10 epochs and the ReduceLROnPlateu was used as the learning rate scheduler to reduce the learning rate. The Adam optimizer algorithm was used. Only three dense layers were added for feature learning and sorting of the features, and a flattening layer

was incorporated in the classification head to make the 2D feature maps to 1D vectors. This classification head was added to the pre-trained frozen backbones. A SoftMax activation layer was used to get multi-class output for the final classification results.

5.2.2 Exploration and Analysis of Ensemble Methods

Different ensembles were approached to combine the results of more than one model to increase reliability as well as classification performance. Three types of well-known feature-level ensemble and four types of decision ensembles were experimented to understand how each of them works and influence the performance and if they address the limitations faced by standalone feature extractors.

For diversity of prediction decisions, MobileNetV3Large was used along with ConvNeXt-XLarge and EfficientNetB3. Each CNN contains different convolutional architectures.

MobileNetV3Large [62] is a CNN model which is known for its low latency and high accuracy on mobile devices. Its architecture comprises depth-wise convolutions, squeeze-and-excitation modules, and optimized activation functions along with platform-aware Neural Architecture Search (NAS) and NetAdapt that gives it its low latency and high accuracy performance. The base architecture is shown in the figure 5.5 [63].

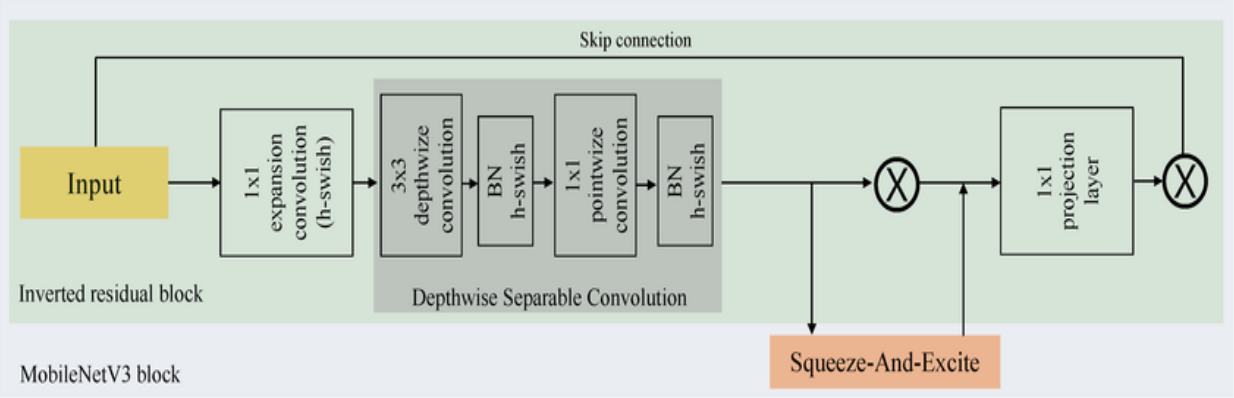

Figure 5.5: MobileNetV3 Base Architecture [63]

For feature ensembles, we tried three ways of merging the extracted feature maps:

Concatenation: A common merging technique where the feature maps from more than one model are merged into a single vector by combining them side by side to a chosen dimension which increases the feature dimensions [64].

Multiplication: The feature maps are merged by multiplying them elementwise to produce a vector the same size as the input feature maps [64].

Summation: Here the maps are merged by adding them elementwise which gives a vector that is the same size as the input feature maps [64].

The feature maps taken from the three CNNs were merged to get a better representation of the features present in each image. This was done to understand how the combination of more than one model's feature map affects the performance of disease classification as better feature representation is a necessary step for the model to understand discriminative features.

It helped to evaluate the effect of multi-model feature fusion on disease classification performance and assess whether combining complementary features from different CNNs leads to more discriminative representations than single-model approaches**.**

Concatenation ensemble increases feature representation which allows more features from the backbones to be included in the feature map without loss of information; however, in case of summation and multiplication merging feature ensemble, the feature map remains the same size hence this may have led to loss of features.

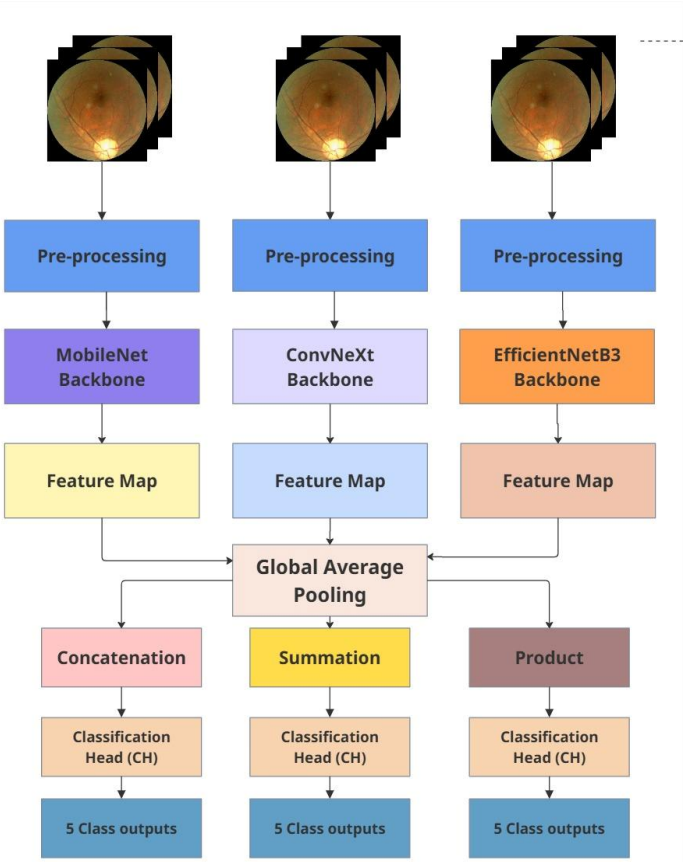

Figure 5.6: Overview of Feature Ensemble Method

For decision ensembles, we tried four types of decision ensemble techniques:

Majority Vote: In this decision ensemble technique, each classifier casts a vote for the predicted class, and the ensemble assigns the final class label based on the majority vote [65]. This helps to reduce errors that might arise from any single model. By combining multiple decisions, this strategy aimed to produce more reliable predictions than standalone models.

Max Rule: For this technique, the three models predict probabilities or confidence scores for each class on the test input images. The model with the highest score for a class is chosen as the final prediction for the corresponding image. The predictions are then compared with the true labels of the image, and accuracy is calculated from there [65].

Weighted Probability Average: For this decision ensemble technique, the models' decisions were given weights according to their accuracy; therefore, the model with the highest accuracy had more influence in the classification prediction [66].

Probability Average: In this setting, the models predict SoftMax probabilities or confidence scores on the test images of each class. The confidence scores of each model for a given class are averaged to give the final prediction, so every model contributes equally to the final decision [67].

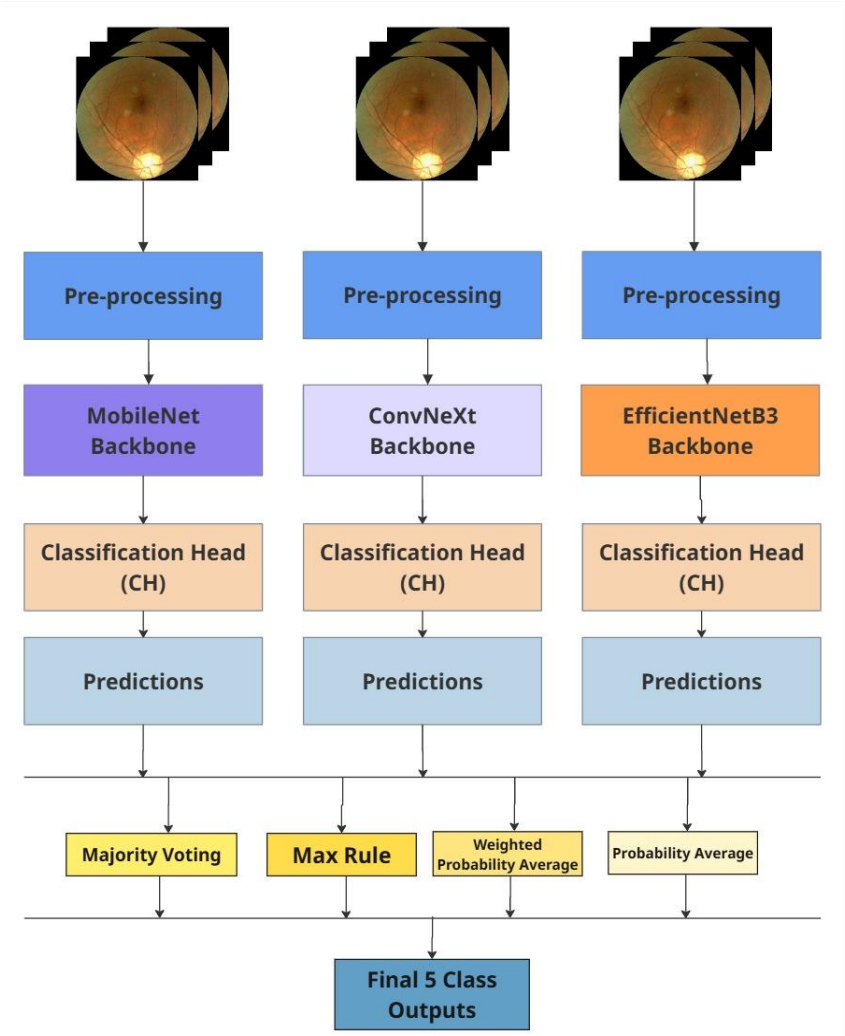

Figure 5.7: Overview of Decision Ensemble Method

These decision ensembles were experimented with to understand how the prediction of each model affects the final decision in classifying the diseases correctly.

While more than one model's decision was taken for the decision ensemble methods, the models may not have given diverse decisions for each class, this ultimately led to a biased decision for the classes.

From this study, feature concatenation ensemble showed the most reliable approach as more than one CNNs feature map was combined, which also ensured diverse feature inclusion and prevention of information loss. Such insights led us to choose feature concatenation ensemble as the ensemble strategy for our method.

5.2.2.1 Training Parameters of Decision and Feature Ensembles

For the ensembles, the training was done for 40 epochs with a learning rate of 0.0001 and the ReduceLROnPlateu was used as the learning rate scheduler to reduce the learning rate with a patience of 10 epochs. The Adam optimizer algorithm was used. Three dropout layers and three Global Average Pooling layers were used during the training to prevent the models from memorizing the dataset and not being fixated on a specific pattern. Global Average Pooling was also used so that an average of the features was taken and reduced dimensions.

5.2.3 Bilateral Ensemble (Proposed Method)

5.2.3.1 Proposed Method Preprocessing

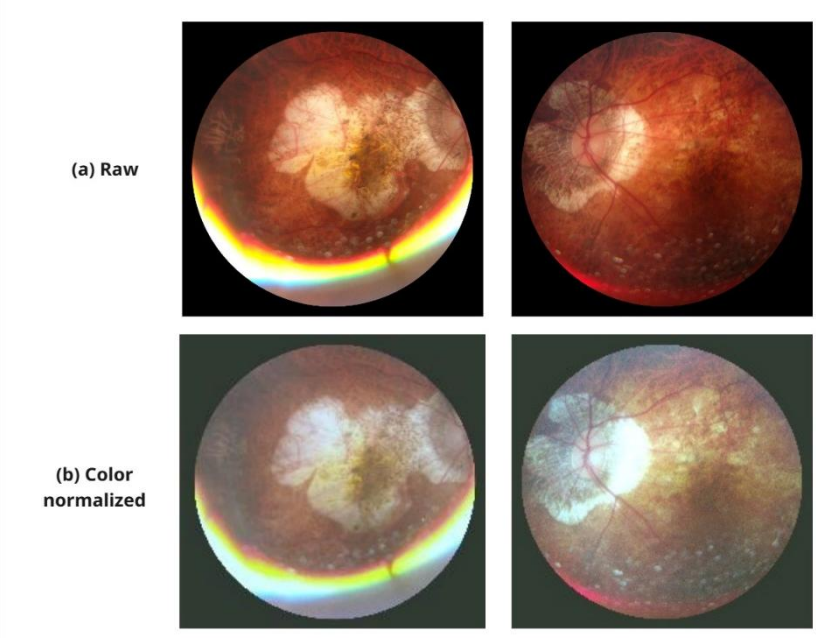

Figure 5.8: Sample Preprocessed Images. (a) Raw; (b) Color Normalized

For the purpose of our method, the images of the dataset were color normalized. The images of the dataset have varying contrasts and brightness as they were sourced from different hospitals and medical centers. To make the images consistent and within a valid intensity range of 0 to 255 pixel values color normalization is applied. This removes unwanted brightness and adjusts the images to be similar to ImageNet. This was done so that the model focuses more on the relevant features and not be fixated on areas with brightness and contrast shifts. Since the left and right eye images of a subject are processed together, they are naturally correlated. To

prevent differences in illumination or color distribution from causing one eye to dominate the learned features of another, color normalization was applied as a preprocessing step. This ensures balanced feature extraction from both eyes and allows the model to focus on disease-related patterns rather than being fixated on regions due to varying resolution of images caused by the capturing of images using different acquisition conditions that was inherent in the dataset.

5.2.3.2 Bilateral Method

Retinal diseases can affect both eyes, but each eye may be at a different stage of progression. This can lead to one eye showing one type of sign and the other showing signs of an advanced stage of the same disease. To leverage these complimentary signs from both eyes of a subject, our method proposes a bilateral ensemble approach.

To fully exploit the complementary information present in the left and right eye and keeping binocular vision in mind, a bilateral ensemble strategy was implemented to enhance the accuracy of retinal disease classification. For this strategy, two identical ConvNeXt-XLarge backbone networks were used, each independently processing images from one eye specifically, one network analyzed the left eye images while the other handled the right eye images of the same subject, only subjects that had the same disease in both eyes were used. ConvNeXt-XLarge was chosen as the feature extractor due to its large kernels covering wide areas and its large number of parameters that give it enough capability to extract and learn the features. The two identical backbones extracted the deep feature maps from its respective input images, to capture detailed representations of retinal structures and potential pathological indicators. After feature extraction, the resulting feature maps from both eyes were concatenated to create a unified, comprehensive portrayal of the features that contain bilateral retinal information. This unified feature map was then fed into a classification head composed of fully connected dense layers, connected to a SoftMax activation function that gave probabilistic predictions across the five retinal disease classes. The design of this classification head was optimized to interpret the complex interactions between bilateral features, enabling the model to discern subtle inter eye variations that might be critical for accurate diagnosis. To further investigate the impact of image preprocessing on bilateral analysis, the same two-backbone ensemble was applied to color-corrected fundus images as well. The color correction step was aimed at standardizing image appearance and enhancing retinal feature visibility, thereby potentially increasing the quality of feature extraction in each network branch. This color correction was done by the application of color normalization. This setup allowed for a direct comparison of model performance with and without the benefit of preprocessing under the bilateral framework. The OIA-ODIR dataset was used for this method.

In previous studies, each method analyzed used the conventional single eye classification approach. For the single eye classification approach, the CNN only takes one image as input hence the complimentary information present in both eyes of a subject is not utilized. This potentially limits relevancy and trustworthiness of the results.

The proposed method takes information from both eyes using two ConvNeXt-XLarge backbones. This CNNs large kernel allows it to extract both local and global features and the concatenation merging technique combines the extracted features from both backbones. Concatenation retains all information of both feature maps which is then used for final classification decisions.

As clinicians take both eyes into consideration when scanning a patient's eyes this method tries to employ bilateral diagnostic reasoning by jointly learning correlated disease features of both eyes, making the method more clinically aligned and reliable.

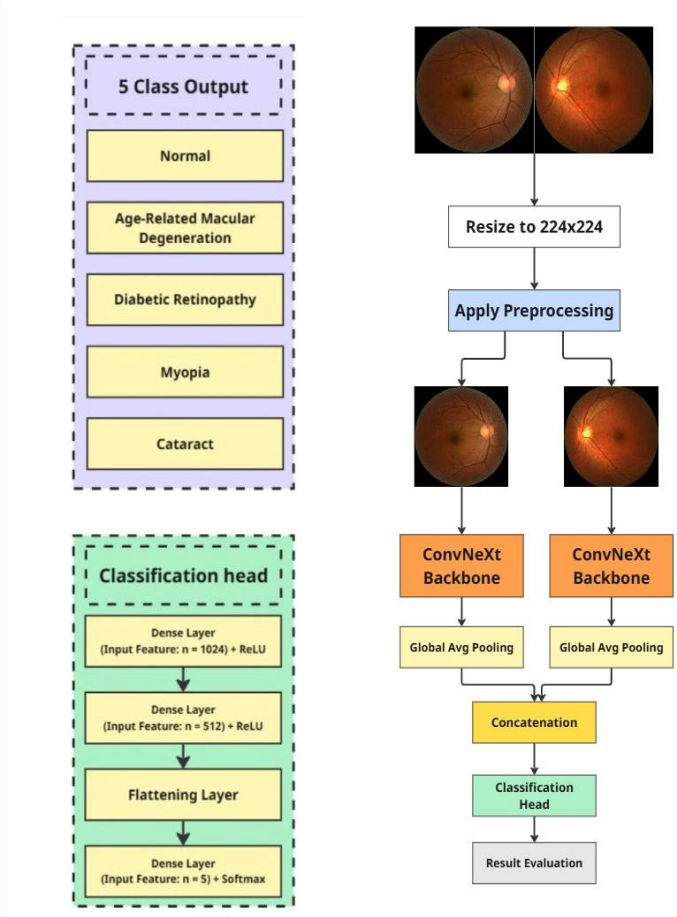

Figure 5.9: Overview of the Proposed Bilateral Ensemble Method

5.2.3.3 Training parameters of Proposed Method

The models were trained for 40 epochs with a batch size of 32. Adam optimizer was used as the loss function optimizer. With an initial learning rate of 0.0001, the ReduceLROnPlateu was used as the learning rate scheduler to reduce the learning rate with a patience of 10 epochs.

Two Global Average Pooling layers were incorporated to capture the most prominent features as well as reduce overfitting during training and three dense layers were incorporated in the classification head of the models with a SoftMax for multi-class classification.

Training was done using Google Colab's T4 GPU with 16 GB RAM.

5.3 Evaluation Metrics

To evaluate each method's performance, we used a comprehensive set of metrics, which include Accuracy, Precision, Recall, Specificity, F1 Score, AUROC (Area Under the Receiver Operating Characteristic curve).

Grad-CAM or Gradient-weighted Class Activation Mapping was also used to understand the underlying work of the conventional architectures and their focal point for making a prediction.

5.3.1 Confusion Matrix

Confusion matrix is a table that is used to measure how well a model is performing by comparing the model's predictions with the actual true labels of the image. It breaks the results into four categories: True positive, False positive, False negatives, True negatives.

False positive value indicates the times the model says there is a disease when it's a healthy eye image while False negative is the times the model says the eye is healthy when there is a disease.

True positive indicates the number of times the model correctly detects the disease, and True Negative indicates the number of times the model correctly identifies no disease cases.

These values were used to calculate the evaluation metrics.

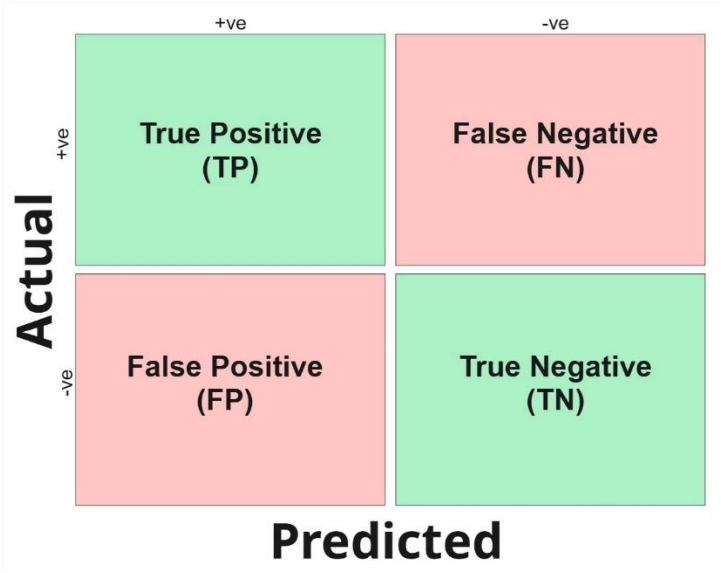

Figure 5.10: Confusion Matrix Representation

5.3.2 Accuracy

Accuracy is one of the most used evaluation metrics in classification tasks. It measures the proportion of correct predictions out of all predictions made by the model. Accuracy is

calculated by using equation (3) is a measure of correctly classified diseases which includes both the presence and absence of the target diseases.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{3}$$

5.3.3 Precision

Precision calculated using equation (4) is an indication of the number of cases predicted as the target disease that were truly correct, so higher precision means fewer false positive cases. From a medical perspective a false alarm can lead to unnecessary treatments.

$$Precision = \frac{TP}{TP + FP} \tag{4}$$

5.3.4 Recall

Recall is calculated by using the equation (5), it gives a measure of the actual target disease cases that were correctly identified which leads to fewer missed disease cases. Higher recall means a model misses fewer true disease cases. This is essential as a missing disease case can have consequences.

$$Recall = \frac{TP}{TP + FN} \tag{5}$$

5.3.5 Specificity

Specificity is calculated by using equation (6), is a measure of the negative cases that did not have the target disease which were correctly identified as negative. A high specificity value means the model is effective at correctly rejecting non-target images, which is important for avoiding false positive predictions of disease in healthy eyes.

$$Specificity = \frac{TN}{TN + FP} \tag{6}$$

5.3.6 F1 Score

F1 Score is calculated by using equation (7), here accuracy and precision are combined to give a balanced measure of these two values. It gives a balanced quantification for the model's ability to correctly detect diseases while avoiding false positives.

$$F1\ Score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \tag{7}$$

5.3.7 AUROC

AUROC (Area under the Receiver Operating Characteristic Curve) measures a model's ability to separate positive cases from negative cases at different threshold levels. The area under the ROC curve which plots True Positive Rate against False Positive rate at varying thresholds is used to compute AUROC values.

5.3.8 Grad-CAM

Grad-CAM (Gradient-weighted Class Activation Mapping) [29] is a technique used to understand how a deep learning model makes its decisions. It uses the gradients of feature maps that go into the convolutional layers or transformer blocks to produce a heatmap that shows which areas had more importance in making a classification decision.

Grad-CAM generates a heatmap and overlays these on an image, this overlay highlights which areas it was focusing on when classifying an image. This technique helps to reveal if a model is focusing on the relevant or correct regions of an image to classify it or if it is focusing on the wrong regions and making a decision based on that. In case of fundus images, the Grad-CAM heatmap can reveal if the model is paying attention to relevant features like blood vessels, the optic disc, or lesions.

In the medical domain it is very important to have a visual and transparent explanation of the model's inner workings as they are assessing medical imaging and Grad-CAM provides this transparency through its heatmaps.

All the generated heatmaps from the Grad-CAM were used to gain a nuanced understanding of the workings of the conventional deep learning architectures and to see how their focus on an image is affected due to their architectural design. The analysis was carried out to gain an understanding of the reliability and trustworthiness of the models' decisions for the classification of the classes.

CHAPTER 6: RESULTS AND DISCUSSION

6.1 Results and Discussion of Conventional Architectures

The dataset, ODIR-5K, used for this phase of experimentations was split to 70% training, 15% testing and 15% validation.

After splitting the dataset to train, test and validation set the total number of images per class for each set is given below:

Table 6.1: Images Per Class After Data Split

Class	Train	Validation	Test
AMD	166	32	39
Cataract	190	36	36
DR	962	213	210
Myopia	159	30	38
Normal	1965	419	432

For a deeper insight into the model's performance, the class wise metrics of ConvNeXt-XLarge is given in table 6.2.

Table 6.2: Class Wise Evaluation Scores of ConvNeXt-XLarge

Class	Precision (%)	Recall (%)	F1 Score (%)	Specificity (%)	Accuracy (%)
AMD	85.0	43.5	57.6	**99.5**	73.6
Cataract	87.1	**94.4**	90.6	99.3	
DR	60.8	49.5	54.5	87.7	
Myopia	**97.1**	89.4	**93.1**	99.8	
Normal	74.9	84.9	79.6	61.9	
Macro Average	81.0	72.4	75.1		

ConvNeXt-XLarge has achieved an overall accuracy of 73.6% as shown in table 6.2. When broken down into class wise metrics, the model has received the highest accuracy (97.1%) and F1 Score (93.1%) for Myopia. This indicates that the model correctly predicted a high portion of the total cases for Myopia and the F1 score indicates the model is effective at both detecting positives and avoiding errors in classification for Myopia. It also shows a good recall score (94.4%) for Cataract, this indicates that the model is able to identify more true disease cases for this class. However, for disease classes with subtle disease markers such as DR it has achieved the lowest scores, which shows that it is not suitable in classifying this class correctly. It also shows more than 80% recall for Normal class which shows that of all actual Normal eye images, more than 80% were correctly identified as Normal class and its 85% precision for AMD shows that of all eye images predicted as AMD 85% were truly AMD cases. It also achieved more than 95% specificity for AMD, Cataract and Myopia which means it is effective at avoiding false positives for these cases.

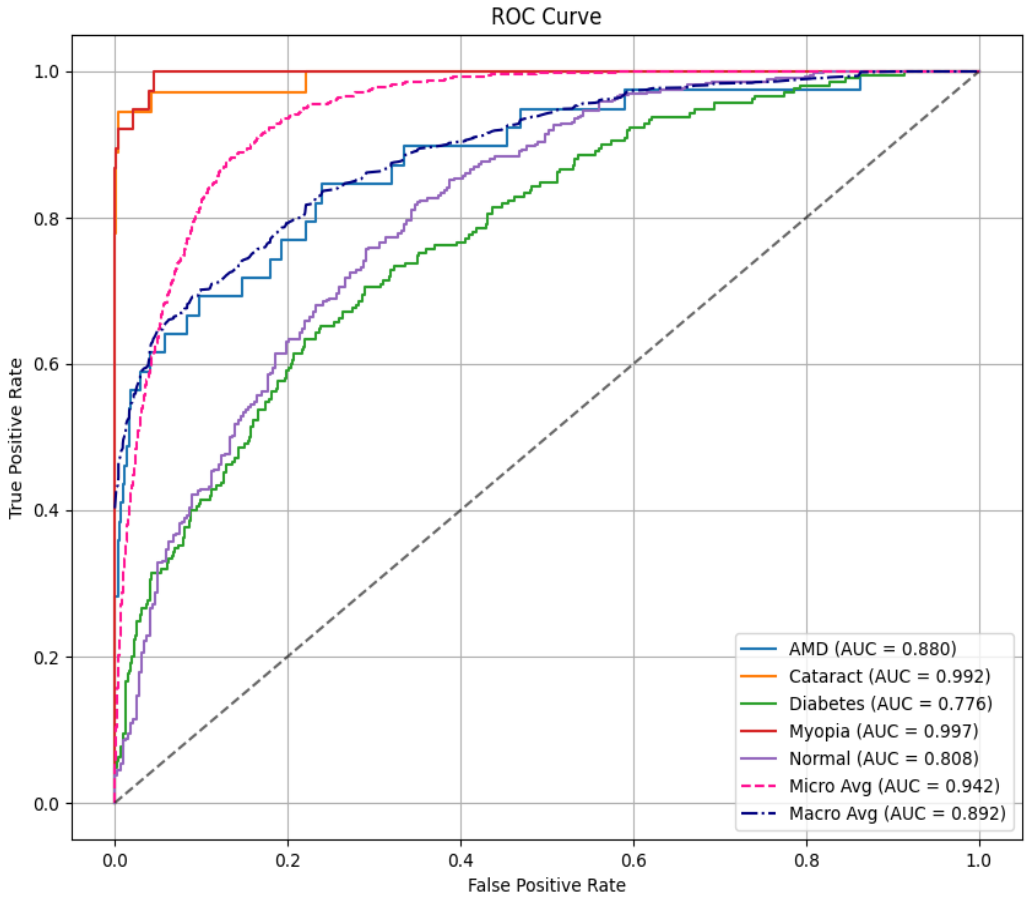

Figure 6.1: ROC Curve of ConvNeXt-XLarge

The ROC Curve gives further insight into the model's discrimination ability for each class. It shows the highest AUC values for Myopia (0.997) and Cataract (0.992) which shows that this model is effective for these two classes. Its Macro Average is 0.892, the macro average takes values of all classes and gives equal importance to each, which suggests an overall strong performance for and discriminatory power across all the five classes.

The class wise metrics of EfficientNetB3 are shown in table 6.3. The table shows the metrics scores of each class which gives a deeper insight of the model's performance for each class.

Table 6.3: Class Wise Evaluation Scores of EfficientNetB3

Class	Precision (%)	Recall (%)	F1 Score (%)	Specificity (%)	Accuracy (%)
AMD	76.4	33.3	46.4	**99.4**	69.6
Cataract	86.4	**88.8**	**87.6**	99.3	
DR	54.5	48.1	51.1	84.5	
Myopia	**87.5**	73.6	80.0	**99.4**	
Normal	72.7	81.4	76.8	59.1	
Macro Average	75.5	65.1	68.8		

The class wise metrics of EfficientNetB3 given in Table 6.3 shows highest accuracy (87.5%) and specificity (99.4%) for Myopia and highest Recall (88.8%) and F1 Score(87.6%) for Cataract suggesting its well suitability for these two classes however comparing the other classes scores of this model against the scores of ConvNeXt-XLarge from table 6.2 it can be seen that it has a low score and low macro average value across the metrics for Normal, DR and AMD class.

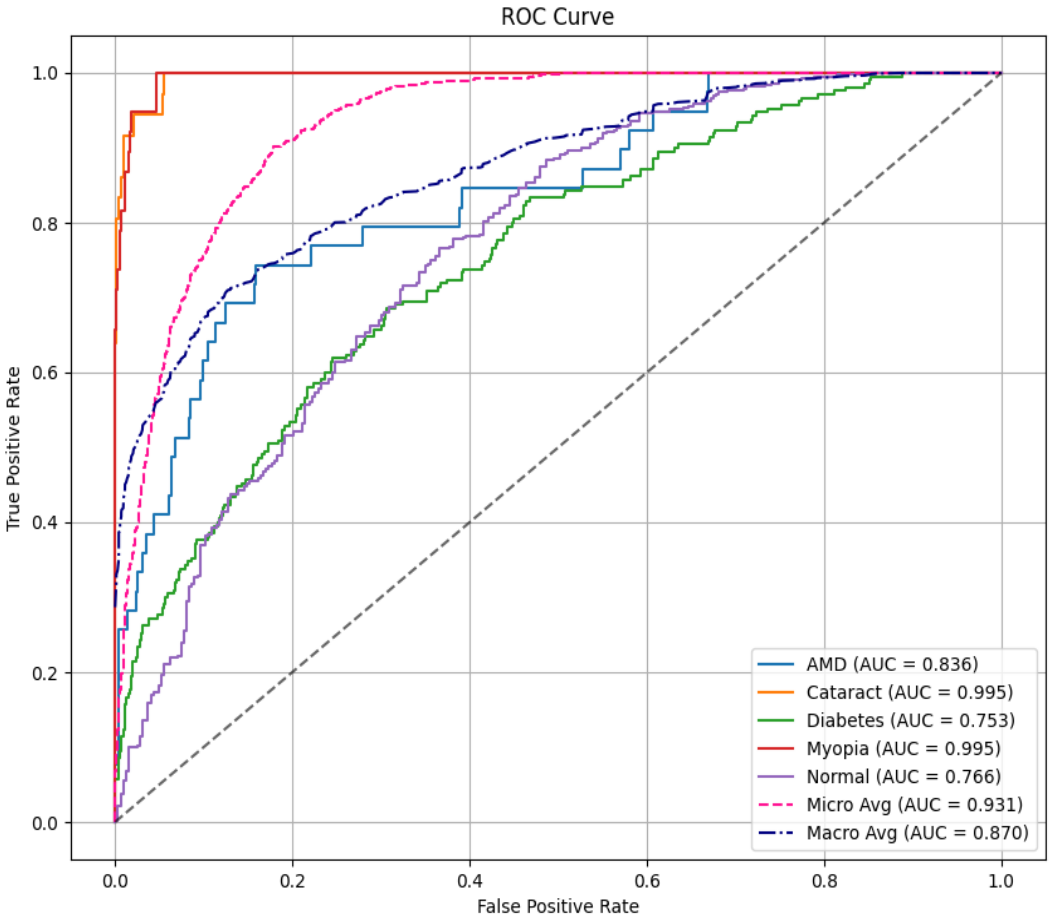

Figure 6.2: ROC Curve of EfficientNetB3

The ROC Curve further reveals the model's discrimination power. The AUC values of Myopia (0.995) and Cataract (0.995) show their strong discrimination ability for these classes and a moderate ability for AMD (0.836) however it shows a low AUC value for DR (0.753) and Normal (0.766), indicating a low differentiating power for these classes. The Macro Average value (0.87) indicates a consistent discrimination power for all the classes combined.

The class wise metrics of ViT-B-16 is shown in table 6.4. Here, it can be seen that ViT-B-16 achieved an accuracy of 71.6% meaning it was able to correctly classify approximately 71.6% of the test samples which demonstrates a moderate ability to capture disease features and classify them accordingly. The table shows the metrics scores of each class which gives a deeper insight of the model's performance for each class as accuracy alone is not sufficient to judge a model's performance.

Table 3.4: Class Wise Evaluation Scores of ViT-B-16

Class	Precision (%)	Recall (%)	F1 Score (%)	Specificity (%)	Accuracy (%)
AMD	50.0	12.8	20	99.0	71.6
Cataract	**96.7**	83.0	**89.5**	**99.8**	
DR	59.5	47.6	52.9	87.5	
Myopia	91.8	81.5	86.0	99.5	
Normal	73.0	**86.8**	79.0	57.5	
Macro Average	74.1	62.4	65.6		

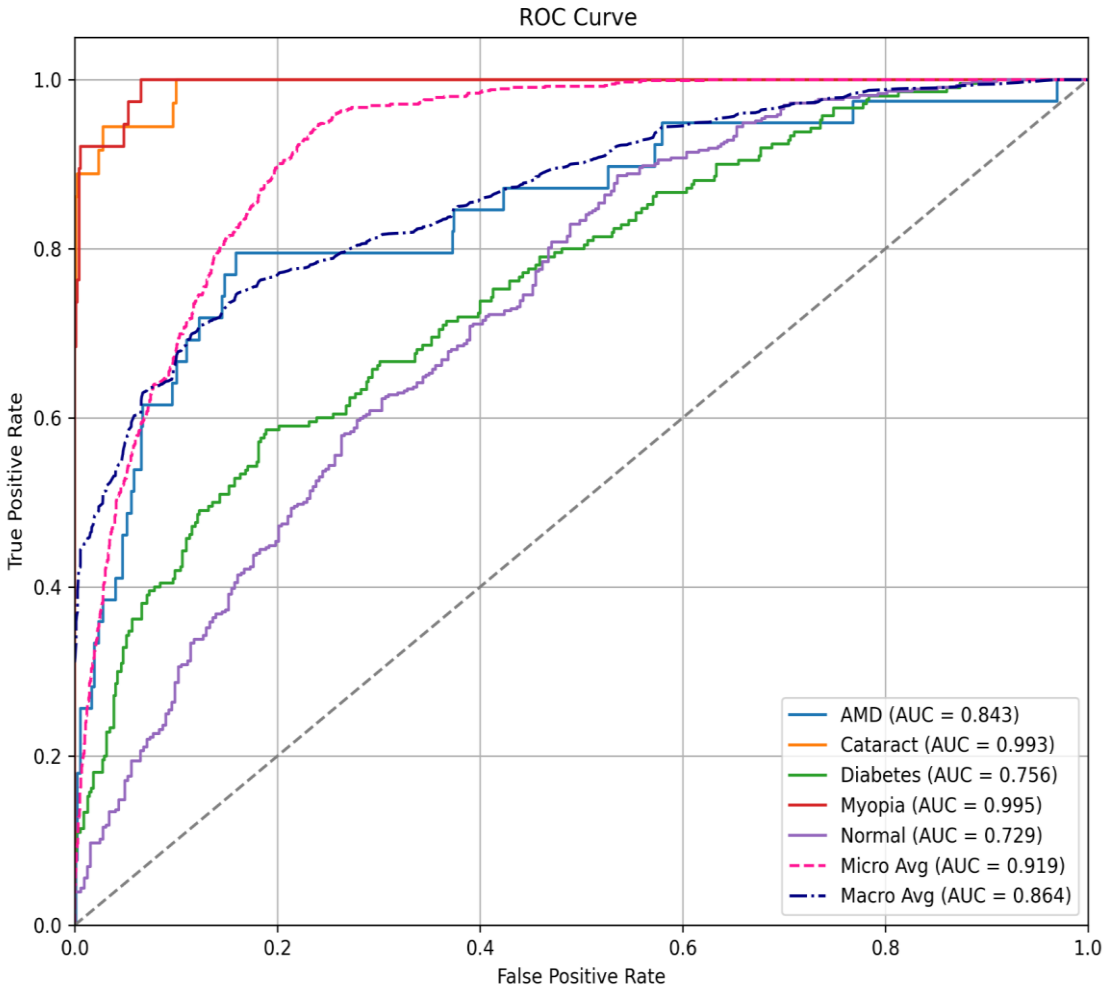

Figure 6.3: ROC Curve of ViT-B-16

The ROC Curve further reveals the model's discrimination power. The AUC values of Myopia (0.995) and Cataract (0.993) show its strong discrimination ability for these two classes and a moderate ability for AMD (0.843) however it shows a low AUC value for DR (0.756) and Normal (0.729), indicating a low differentiating power for these classes.

Comparing the scores from Table 6.2, 6.3 and 6.4 it can be seen that ConvNeXt-XLarge had the highest performance for almost all classes across the evaluation thresholds. Outperforming

EfficientNetB3 by 4% and ViT-B-16 by 2% in accuracy and it also outperforms them by 2.2% and 2.8% respectively in AUROC values.

To further investigate and understand the architecture of the models and the influence it has on the decision making of a model, we then applied Grad-CAM and analyzed the heatmaps generated by each model.

Sample heatmaps have been given below:

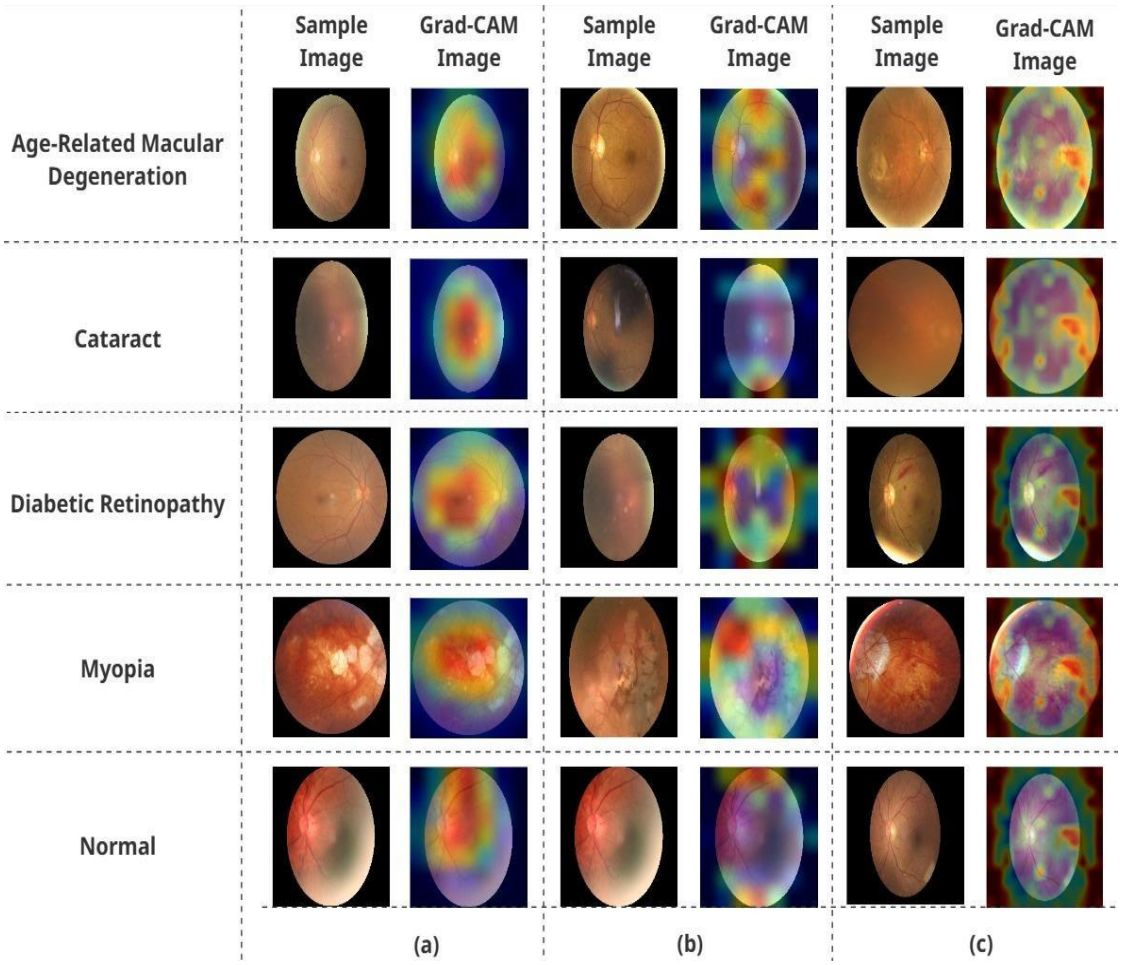

Figure 6.4: Grad-CAM Visualization. (a) EfficientNetB3; (b) ConvNeXt-XLarge; (c) ViT-B-16

As seen in Figure 6.4, the Grad-CAM of a) EfficientNetB3 shows a centralized focus on the retinal images indicated by the warm colors of the heatmap. This can be attributed to EfficientNetB3's compound scaling and kernel size, which is useful for disease markers that present themselves in the central region. However, it does not cover the other relevant regions of the image. While ConvNeXt-XLarge shows a wide focus with various wide warm regions on the heatmaps, this can be attributed to its large kernel size inspired by transformers. However, the heatmap also shows that it focuses on the darker regions of the images that are not relevant for making a classification decision. While c) ViT-B-16 shows a dispersed focus as seen from the warm regions on the heatmap. This dispersed focus can be attributed to the

vision transformer's patching of the image where the image gets divided into equal sized patches; however, this leads to the model not fully focusing on the relevant regions.

Taking all the results together, ConvNeXt-XLarge covers most of the regions of the images and achieved an accuracy of 73.6% and AUROC of 89.2%. This shows that ConvNeXt-XLarge was able to focus on broad regions, however it is still unable to utilize the features from those regions properly. Hence to make use of this broad focusing ability and to enhance the feature representation and utilize it effectively we then explored two types of ensembles and their various techniques.

6.2 Results and Discussion of Decision and Feature Ensembles

To get a better and reliable performance using CNNs we experimented in ensemble settings to understand how each ensemble strategy affects the classification performance. As more than one model is involved in these settings, our aim was to use this group-up approach to systematically analyze how combining diverse classifiers influences performance and to identify which ensemble strategy best addresses the limitations of conventional CNNs.

We tried three methods of merging features and used three different backbones (ConvNeXt-XLarge, MobileNetV3Large, and EfficientNetB3). We introduced MobileNetV3Large as the third base model for its ability to capture fine-grained local features efficiently, offering complementary representations alongside ConvNeXt and EfficientNet. Images of each class after the train, test and validation split have been given below.

Table 6.5: Images Per Class After Data Split

Class	Train	Validation	Test
AMD	223	43	41
Cataract	224	43	46
DR	1152	245	287
Myopia	178	49	29
Normal	2775	559	562

6.2.1 Results and Discussion of Feature Ensembles

Table 6.6 gives class wise scores of multiplication merging feature ensemble technique:

Table 6.6: Class Wise Evaluation Scores of Multiplication Merging

Class	Precision (%)	Recall (%)	F1 Score (%)	Specificity (%)	Accuracy (%)
AMD	56.0	34.0	42.0	98.8	72.8
Cataract	**95**.0	**84.7**	**89.6**	**99.7**	
DR	62.0	52.9	57.0	86.5	
Myopia	82.7	82.7	82.7	99.4	
Normal	75.6	84.0	79.7	62.0	
Macro Average	74.4	67.8	70.2		

Class wise accuracy for single eye classification using feature concatenation ensemble has been given below:

Table 6.7: Class Wise Evaluation Scores of Concatenation Merging

Class	Precision (%)	Recall (%)	F1 Score (%)	Specificity (%)	Accuracy (%)
AMD	62.5	36.5	46.0	99.0	73.6
Cataract	**95.0**	89.0	**92.0**	**99.7**	
DR	69.9	39.0	50.6	92.7	
Myopia	82.7	82.7	82.7	99.4	
Normal	72.6	**91.0**	80.9	52.0	
Macro Average	72.8	72.8	72.8		

Class wise accuracy for single eye classification using feature summation merging ensemble has been given below:

Table 6.8: Class Wise Evaluation Scores of Summation Merging

Class	Precision (%)	Recall (%)	F1 Score (%)	Specificity (%)	Accuracy (%)
AMD	57.0	39.0	46.0	98.7	72.4
Cataract	**95**.0	**91**.0	**93**.0	**99.7**	
DR	62.9	47.0	54.0	88.2	
Myopia	80.0	82.7	81.0	99.3	
Normal	74.0	85.5	79.5	58.8	
Macro Average	73.9	69.2	70.9		

From Table 6.6, 6.7, and 6.8, it can be seen that the feature concatenation merging technique gave the best accuracy (73.6%) among all three merging techniques. Outperforming summation technique (72.4%) by 1.2%, and feature multiplication or multiplication (72.8%) by 0.8%. It also shows good scores for almost all the classes across the evaluation metrics. It also shows 99% specificity for AMD which shows that it was able to correctly discriminate between healthy and AMD images. Summation and multiplication merging shows a good score for Cataract and Myopia, but they show low scores for AMD and DR across most metrics.

From Figure 6.5, it can be seen that the AUC values of feature concatenation are the highest among the three methods, for all the classes. Its micro average value (0.90), which is an average of the AUC values across all classes, outperforms summation (0.883) by 1.7%; it also outperforms multiplication or multiplication technique by 2.2%. The macro average AUC shows its superior discrimination ability for all classes.

Both concatenation and summation techniques achieved the same AUC (0.998) for Myopia, which shows that for this class, these two techniques have the same discrimination ability.

Hence, overall, concatenation feature ensemble shows a balanced performance when all scores are considered.

ROC Curve of the three merging techniques:

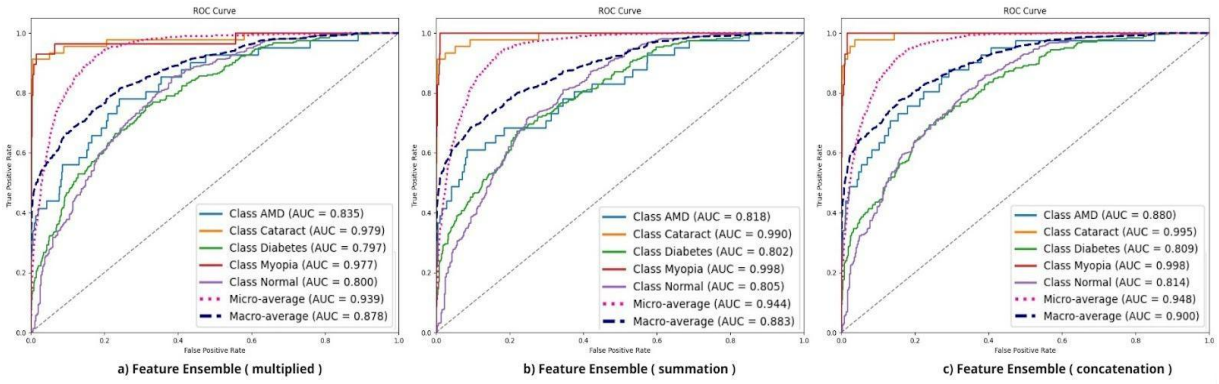

Figure 6.5: ROC Curves of Feature Ensemble. (a) Multiplication; (b) Summation; (c) Concatenation

These insights suggest the advantage of feature concatenating the feature maps from the three backbones and the prevention of information loss especially for classes DR and AMD as well as Cataract and Myopia.

Thus, from this phase of experimentation we deduced that decision ensembles do not have the suitability to make use of extracted features hence the feature ensemble's feature concatenation is well suited for better feature representation which is needed for fair classification decision.

6.2.2 Results and Discussion of Decision Ensembles

Table 6.9: Class Wise Evaluation Scores of ConvNeXt-XLarge

Class	Precision (%)	Recall (%)	F1 Score (%)	Specificity (%)	Accuracy (%)
AMD	55.0	53.6	54.0	98.0	71.6
Cataract	**80.7**	**91**.0	**85.7**	**98.9**	
DR	62.8	51.0	56.0	87.0	
Myopia	74.0	89.6	81.0	99.0	
Normal	75.0	80.7	77.8	62.7	
Macro Average	69.6	73.3	71.1		

For decision ensembles the base models were trained and tested and the confidence scores or SoftMax probabilities were used for the decision ensemble techniques. The class wise metrics of each individual base model: ConvNeXt-XLarge (Table 6.9), EfficientNetB3 (Table 6.10), MobileNetV3Large (Table 6.11) are given below. From the base model's evaluation scores, it can be seen that in this phase of experimentations, ConvNeXt-XLarge had a slightly increased score in accuracy (71.6%), outperforming EfficientNetB3 by 1.5% and MobileNetV3Large by

4.9%. The scores suggest that the architectural design and parameters of ConvNeXt-Xlarge was more effective at extracting the complex retinal disease features, which enabled better discrimination between classes compared to the other models.

Table 6.10: Class Wise Evaluation Scores of EfficientNetB3

Class	Precision (%)	Recall (%)	F1 Score (%)	Specificity (%)	Accuracy (%)
AMD	44.7	41.0	43.0	97.7	70.8
Cataract	**91**.0	**89**.0	**90**.0	**99.5**	
DR	65.5	39.7	49.0	91.0	
Myopia	77.0	82.7	80.0	99.0	
Normal	72.0	86.8	78.7	53.0	
Macro Average	70.1	67.9	68.2		

Among the three models, MobileNetV3Large had the lowest accuracy of 69.3%, which shows that its lightweight design, while computationally efficient, may be less capable of capturing the complex and subtle retinal disease features needed for accurate classification.

Table 6.11: Class Wise Evaluation Scores of MobileNetV3Large

Class	Precision (%)	Recall (%)	F1 Score (%)	Specificity (%)	Accuracy (%)
AMD	34.0	36.5	35.0	96.8	69.3
Cataract	**84**.0	**91**.0	**87.5**	**99**.0	
DR	62.0	41.0	50.0	89.0	
Myopia	81.0	75.8	78.5	99.0	
Normal	72.0	83.6	77.0	55.0	
Macro Average	66.7	65.8	65.7		

The decision ensemble methods used three base models and two of the models had very close performance while the other had a very low performance. The class wise metrics for the decision ensemble: Majority Voting (Table 6.12), Max Rule (Table 6.13), Weighted Probability Average (Table 6.14) and Probability Average (Table 6.15) are given below:

Table 6.12: Class Wise Evaluation Scores of Majority Voting

Class	Precision (%)	Recall (%)	F1 Score (%)	Specificity (%)	Accuracy (%)
AMD	59.0	46.0	52.0	98.5	72.8
Cataract	**89**.0	**91**.0	**90**.0	**99**.0	
DR	68.0	41.0	51.6	91.8	
Myopia	82.0	79.0	80.7	99.0	
Normal	73.0	88.9	80.0	54.0	
Macro Average	74.4	69.4	70.9		

Table 6.13: Class Wise Evaluation Scores of Max Rule

Class	Precision (%)	Recall (%)	F1 Score (%)	Specificity (%)	Accuracy (%)
AMD	58.0	43.9	50.0	98.5	72.9
Cataract	**91**.0	**91**.0	**91**.0	**99.5**	
DR	67.8	41.8	51.7	91.5	
Myopia	78.0	86.0	81.9	99.0	
Normal	73.0	88.7	80.0	55.0	
Macro Average	73.7	70.4	71.08		

Table 6.14: Class Wise Evaluation Scores of Weighed Probability Average

Class	Precision (%)	Recall (%)	F1 Score (%)	Specificity (%)	Accuracy (%)
AMD	62.9	41.0	50.0	98.9	72.8
Cataract	**93**.0	**91**.0	**92**.0	**99.6**	
DR	68.0	41.0	51.0	91.8	
Myopia	79.0	79.0	79.0	99.0	
Normal	72.7	89.5	80.0	53.0	
Macro Average	75.32	68.54	70.64		

Table 6.15: Class Wise Evaluation Scores of Probability Average

Class	Precision (%)	Recall (%)	F1 Score (%)	Specificity (%)	Accuracy (%)
AMD	62.9	41.0	50.0	98.9	72.8
Cataract	**93**.0	**91**.0	**92**.0	**99.6**	
DR	68.0	41.0	51.0	91.8	
Myopia	79.0	79.0	79.0	99.0	
Normal	72.7	89.5	80.0	53.0	
Macro Average	75.3	68.5	70.6		

Comparing Table 6.12, 6.13, 6.14, 6.15, it can be seen that the Max Rule decision ensemble achieved the highest accuracy (72.9%) among the four ensembles. Max Rule also achieved the highest score in all the evaluation metrics for Cataract class among the four decision ensembles.

Considering the overall accuracy scores, it can be seen that all the decision ensembles' prediction ability is within the same range, this could suggest that the models do not show diversity in their decision making which led to their similar accuracy scores. ConvNeXtXLarge and EfficientNetB3 were very close in accuracy while MobileNet had the lowest performance however despite this the decision ensembles had close results. Another insight was that decision ensembles are not the right approach for us to use for better and reliable classification as the decision ensembles always happen after a decision is made and better feature representation is necessary for reliable results. Hence Feature ensemble has better potential as it is known to prevent feature loss.

From the ROC Curves of the decision ensembles given Figure 6.6, it can be seen that b) Max Rule, a) Probability Average and c) Weighted probability average have shown close macro average AUC values of 0.90, 0.905, and 0.905 respectively. Only d) Majority Voting is behind with an AUC of 0.837. This shows that the three techniques have a similar discrimination ability for the classes which further highlights that the base models are not giving a diverse decision.

ROC Curves of Decision Ensembles:

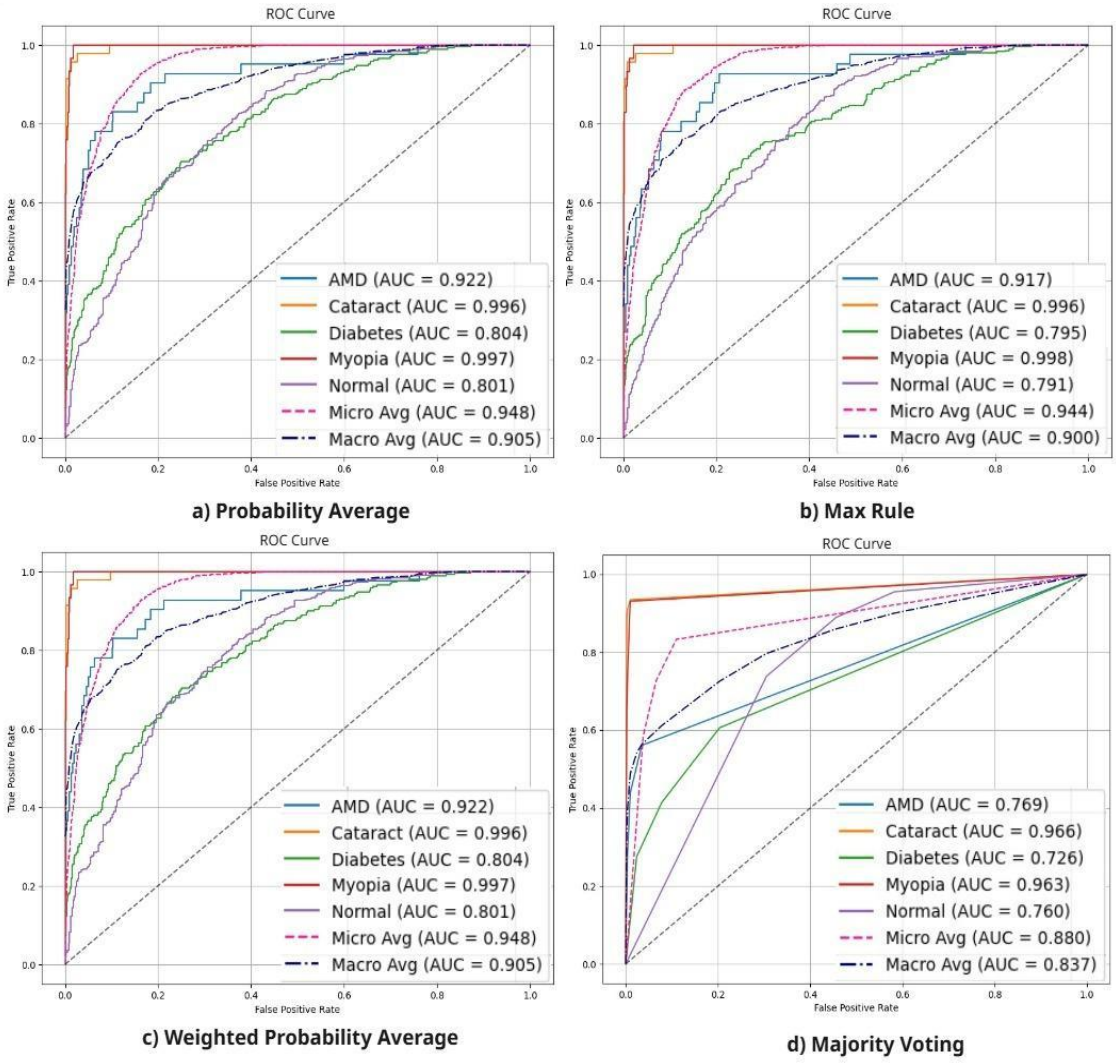

Figure 6.6: ROC Curves of Decision Ensemble (a) Probability Average (b) Max Rule (c) Weighted Probability Average (d) Majority Voting

Comparing the ensembles, the feature concatenation ensemble has the most potential as it includes all feature information from the base models and its accuracy (73.6%) also suggests its strong performance among all the techniques experimented in this study.

However, none of the methods reached a satisfactory result and one reason could be that the models are taking similar information from the same single input image which restricts the

ensemble's ability to capture diverse perspectives of the data, causing the models to reinforce similar feature biases rather than complement each other.

From the comparisons, the feature concatenation ensemble showed it had the best feature representation as seen from its scores, which we used to combine the feature maps of the broad regions covered by two ConvNeXt-XLarge backbones.

ConvNeXt-XLarge contains a large number of parameters that help it to extract rich features and its large kernel makes it have a broad focus that makes it suitable to cover most areas of the image and feature concatenation helps to combine the features without information loss hence these were selected as the main components to be used as part of our proposed method.

6.3 Results and Discussion of Proposed Bilateral Ensemble Method

For this method, complimentary information presented in subject's both eyes were considered hence this limited the sample size for training, testing and validation. The number of images per class after split are given below in table 6.16.

Table 6.16: Images Per Class After Split

Class	Train	Validation	Test
AMD	98	21	22
Cataract	88	18	20
Myopia	77	16	18
DR	476	102	103
Normal	1123	240	242

The class wise metrics of bilateral ensembles using color normalized images are given below:

Table 6.17: Class Wise Evaluation Scores with Color Normalizing Bilateral Ensemble Method

Class	Precision (%)	Recall (%)	F1 (%)	Specificity (%)	Accuracy (%)
AMD	66.0	18.0	28.0	99.0	74.5
Cataract	80.0	**100**.0	88.0	98.0	
DR	64.0	50.0	56.0	90.0	
Myopia	**100**.0	72.0	83.0	**100**.0	
Normal	76.0	88.0	81.0	58.0	
Macro Average	77.0	65.0	67.0		

From table 6.17, it can be seen that the method is the most effective for Cataract due to its 100% recall which indicates that no disease cases were missed for this class. This also suggests that the model was able to fully capture and utilize the disease features. It also showed a 100% precision score for Myopia, which means every prediction for Myopia was correct. It also showed a good performance for the Normal class, suggested by the 88% recall score, the score indicates that the model is good at identifying most of the normal eye cases. However, the

model showed a weak performance for DR and AMD which shows that the model as well the images need to be further enhanced for these two classes. The macro average specificity of 89% also shows a good distinguishing ability between healthy and disease effected images.

However, a 74.5% accuracy shows that the model is an overall balanced classifier for each of the disease classes.

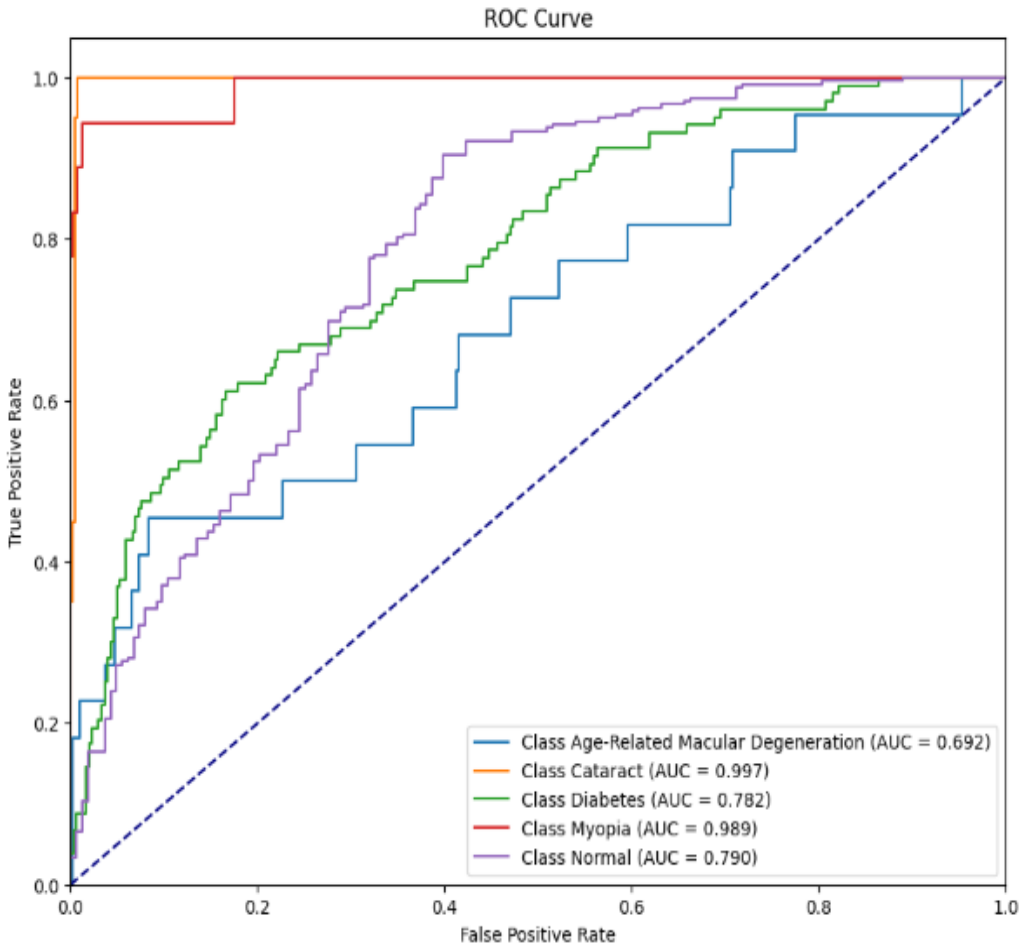

Figure 6.7: ROC Curve of Bilateral Method Using Color Corrected Images

The AUC values of AMD (0.692) and DR (0.782) further evidence the method's performance. It shows weak discrimination ability for these classes. However, AUC measures performance across all thresholds and the dataset was kept imbalanced hence this may have influenced the results. However, it shows strong discrimination power for Cataract (AUC 0.997) and Myopia (AUC 0.989) and a moderate ability for Normal class (AUC 0.79). This shows that at a chosen threshold the method performed well and had a better accuracy than single eye classification.

We then tried the bilateral ensemble using images without color normalizing to understand the effects of color normalization of the images on the method's performance and the results are reported below:

Table 6.18: Class Wise Evaluation Scores Without Color Normalizing Bilateral Ensemble Method

Class	Precision (%)	Recall (%)	F1 Score (%)	Specificity (%)	Accuracy (%)
AMD	44.0	18.1	25.8	98.0	72.1
Cataract	83.0	**100**.0	90.9	98.0	
DR	59.0	50.4	54.4	88.0	
Myopia	**100**.0	61.1	75.8	**100**.0	
Normal	75.4	85.1	80.0	58.9	
Macro Average	72.0	65.0	67.0		

Comparing Table 6.17 and Table 6.18, it can be seen that color normalization was effective as it improved the overall accuracy from 72.1% to 74.5%. It reports a 5% increase in precision as seen from its macro average precision score (72% to 77%). It also showed a slight 1% increase in specificity for AMD (99%) and 2% increase in specificity for DR (90%). This suggests that color normalization helped to enhance the method to effectively reject non-target disease images, avoiding false positive cases for these diseases.

Comparing the experiments, overall, the bilateral ensemble extracts diverse complimentary features of the disease classes from both eyes, unlike single eye classification where the models take information from the same single input image and the concatenation merging technique preserves these features while color normalization helps to enhance the image quality for better disease specific feature extractions, which together contribute to the improved accuracy and class-wise performance observed in the proposed method.

Inspired by clinical practice, the proposed method analyses the left and right eye of a patient by taking concatenated left and right eye images of a subject as input. This approach is not taken for single eye classification that is done in conventional CNN methods where each eye is treated independently without subject-wise context. This allows the model to capture complementary and correlated features between eyes, such as subtle asymmetries, shared pathological patterns, and disease consistency, which are often missed in conventional single-eye classification approaches.

CHAPTER 7: CONCLUSION

7.1 Summary

Eye diseases are a major cause of concern, especially for the aged population. Diseases such as Age-related Macular Degeneration, Diabetic Retinopathy, Myopia and Cataract are some of the common diseases that contribute to vision impairment worldwide. However, early diagnosis can prevent these diseases from permanently damaging the retina and preserving vision.

Traditional methods of diagnosis for these diseases include the manual scanning of ocular images from different imaging modalities such as OCT, FFA, and fundus imaging. Manual screening needs to be done by expert ophthalmologists, but it is a time-intensive process especially for cases that present Subtle disease features which need careful screening to detect the disease. As the population is ever growing, it has become difficult to accommodate patients and provide timely diagnosis.

To tackle such limitations, convolutional neural networks have been developed and validated across various studies by researchers, and they have achieved promising results. However, in the eye domain, many studies have used CNNs to extract information of diseases from single eye images, which has restricted the capacity as well as a reliable method as diseases can appear differently in the left and right eye due to their composition. Furthermore, there are diseases that can affect both eyes and one eye, maybe at the early stage of the disease while the other eye is at an advanced stage. This can show the same disease in different forms. To leverage the complimentary information present in both eyes, a bilateral ensemble using the ODIR-5K also known as OIA-ODIR, is proposed in this study. This dataset contains left and right eye images of 5000 patients with some images containing more than one disease. Hence, for the purpose of this study, only single labeled subject-wise images were considered. By combining two backbones using the same model: ConvNeXt-XLarge, a modernized CNN that mimics a Vision Transformer while using convolutional layers, the proposed method extracts information from both eyes at the same time. The two backbones handle left and right eye images respectively, and the feature maps of both backbones are then combined using the feature concatenation merging technique. This bilateral approach is inspired by the practice of clinicians where they often check both eye images of a patient to understand the severity of a disease and give proper diagnosis. Different backbones were experimented using single eye classification along with Grad-CAM for explainability to understand which model can help with better feature representation and then different ensembles: Feature and Decision level ensembles and their different techniques were evaluated to understand the most reliable technique for feature representation and effective utilization of these feature maps. From these experiments we combined suitable strategies to reach our proposed method. The proposed bilateral ensemble method showed an accuracy of 74.5% which shows a good classification performance however further enhancements in feature extraction, class balancing, and fine-tuning are needed to achieve clinically reliable performance.

7.2. Limitations

While the proposed bilateral ensemble method is more reliable than single-eye classification because it uses information from both eyes, it still has several limitations.

Firstly, the research was based on one, single publicly available (OIA-ODIR) dataset, yet while comprehensive, it does not necessarily reflect the diversity of retinal images seen in the clinical practice. The performance of the proposed framework may therefore vary when used for images of other populations or having different imaging characteristics. Secondly, the model's output is at the moment only possible for a pre-defined set of disease classes and may be limited for other conditions. Thirdly, the framework does not explicitly address potential challenges such as poor image quality, occlusions, or artifacts, which can affect real-world diagnostic reliability. Lastly, the backbone models used in the experiments were used as-is without any fine-tuning which also may have had a part in limiting their potential performance. Additionally, the dataset was used without balancing the classes, hence the common dominating classes may have had more influence on the results.

7.3. Future Work

In summary, our proposed method’s performance suggests strong potential, however it is still limited and requires further study, hence there is room for future experiments to make the method more efficient and effective.

The future work of this study will include handling the limitations which will be addressed in several ways. This will include expanding validation to include additional, more diverse datasets to strengthen the generalizability of the approach. Incorporating more eye disease categories, as well as rare retinal conditions, could enhance the system’s clinical utility. Furthermore, appropriate augmentation techniques will be used to handle dataset imbalance for fair results across all classes. The system will be made to gain more robustness to real-world imaging challenges through the addition of advanced image quality assessment and correction modules. The development of transparent and interpretable deep learning techniques would enable clinicians to understand model predictions, which would lead to better trust in AI-assisted diagnosis. The framework will be enhanced further so that it incorporates multi-modal ocular data such as optical coherence tomography along with retinal fundus images to enable further improvements for the use in automated retinal disease screening & management.

References

[1] World Health Organization, “World report on vision,” WHO, Geneva, Switzerland, 2019.

[2] L. L. Quinn *et al.*, “Interobserver variability studies in diagnostic imaging: a methodological systematic review,” *British Journal of Radiology*, vol. 96, no. 1148, Jun. 2023, doi: https://doi.org/10.1259/bjr.20220972.

[3] S. Resnikoff et al., “The number of ophthalmologists in practice and training worldwide: a growing gap and recommendations for action,” Ophthalmic Epidemiol., vol. 27, no. 2, pp. 79–88, 2020.

[4] H. Cui and Z. Zhang, “Interocular symmetry and asymmetry in ocular health and disease,” *Frontiers in Medicine*, vol. 12, pp. 1626329–1626329, Sep. 2025, doi: https://doi.org/10.3389/fmed.2025.1626329.

[5] M. Elkholy and M. A. Marzouk, “Deep learning-based classification of eye diseases using Convolutional Neural Network for OCT images,” *Frontiers in computer science*, vol. 5, Jan. 2024, doi: https://doi.org/10.3389/fcomp.2023.1252295.

[6] A. Shoukat, S. Akbar, Syed Ale Hassan, S. Iqbal, A. Mehmood, and Qazi Mudassar Ilyas, “Automatic Diagnosis of Glaucoma from Retinal Images Using Deep Learning Approach,” *Diagnostics*, vol. 13, no. 10, pp. 1738–1738, May 2023, doi: https://doi.org/10.3390/diagnostics13101738.

[7] Z. Liu et al., “A ConvNet for the 2020s,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), 2022, pp. 11976–11986.

[8] M. Tan and Q. V. Le, “EfficientNet: Rethinking model scaling for convolutional neural networks,” Proc. Int. Conf. Machine Learning (ICML), 2019, pp. 6105–6114.

[9] A. Dosovitskiy et al., “An image is worth 16x16 words: Transformers for image recognition at scale,” in Proc. Int. Conf. Learning Representations (ICLR), 2021.

[10] D. S. W. Ting, C. Y. Cheung, and T. Y. Wong, “Deep learning in ophthalmology: The technical and clinical-facing challenges,” Eye, vol. 33, no. 8, pp. 1205–1210, 2019.

[11] N. Congdon et al., “Causes and prevalence of visual impairment among adults in the United States,” Arch Ophthalmol, vol. 122, no. 4, pp. 477–485, 2004.

[12] D. A. Antonetti, R. Klein, and T. W. Gardner, “Diabetic retinopathy,” New England J. Med., vol. 366, no. 13, pp. 1227–1239, 2012.

[13] Centers for Disease Control and Prevention, “National Diabetes Statistics Report 2020,” U.S. Dept. Health and Human Services, Atlanta, GA, 2020.

[14] Early Treatment Diabetic Retinopathy Study Research Group, “Early photocoagulation for diabetic retinopathy. ETDRS report number 9,” Ophthalmology, vol. 98, no. 5, pp. 766–785, 1991.

[15] D. M. Brown et al. (RISE and RIDE Research Group), “Ranibizumab for diabetic macular edema: results from 2 phase III trials,” Ophthalmology, vol. 119, no. 4, pp. 789–801, 2012.

[16] “Diabetic Retinopathy, Eye Disease Specialist · Top Eye Doctor · NYC,” *Manhattan Eye Doctors & Best Rated Specialists in NYC*. https://www.eyedoctorophthalmologistnyc.com/treatment/diabetic-retinopathy-eye-disease/

[17] GBD 2019 Blindness and Vision Impairment Collaborators, “Trends in prevalence of blindness and distance vision impairment in 195 countries and territories and regions from 1990 to 2020,” The Lancet Global Health, vol. 9, no. 2, pp. e144–e160, 2021.

[18] “Cataracts - Dr. Róbert GERGELY ophthalmic surgeon,” *Dr. Róbert GERGELY*, Feb. 06, 2023. https://vitreolysis.eu/language/en/cataracts/ (accessed Nov. 20, 2025).

[19] R. D. Jager, W. F. Mieler, and J. S. Miller, “Age-related macular degeneration,” New England J. Med., vol. 358, no. 24, pp. 2606–2617, 2008.

[20] W. L. Wong et al., “Global prevalence of age-related macular degeneration and strategy for prevention,” The Lancet Global Health, vol. 2, no. 2, pp. e106–e116, 2014.

[21] J. Chaudhary, “Age-Related Macular Degeneration (AMD): Symptoms, Diagnosis & Treatment,” *EYE-Q INDIA*, Jun. 07, 2024. https://www.eyeqindia.com/age-related-macular-degeneration-amd-symptoms-diagnosis-treatment/

[22] “Dr Agarwals Eye Hospital,” *Dr Agarwals Eye Hospital*, 2022. https://www.dragarwal.com/eye-treatment/refractive-surgery/myopia/

[23] B. A. Holden et al., “Global prevalence of myopia and high myopia and temporal trends from 2000 through 2050,” Ophthalmology, vol. 123, no. 5, pp. 1036–1042, 2016.

[24] K. Ohno-Matsui et al., “Pathologic myopia,” Asia-Pacific J. Ophthalmol., vol. 10, no. 5, pp. 486-517, 2021.

[25]*Ftcdn.net*,2025.https://as1.ftcdn.net/v2/jpg/01/26/01/20/1000_F_126012021_VrMWZJ8BsqIHbraFAe4hhuBqxUE9LkIg.jpg (accessed Nov. 20, 2025).

[26] J. K. H. Goh, C. Y. Cheung, S. S. Sim, P. C. Tan, G. S. W. Tan, and T. Y. Wong, “Retinal Imaging Techniques for Diabetic Retinopathy Screening,” *Journal of Diabetes Science and Technology*, vol. 10, no. 2, pp. 282–294, Feb. 2016, doi: https://doi.org/10.1177/1932296816629491.

[27] Early Treatment Diabetic Retinopathy Study Research Group, “Grading diabetic retinopathy from stereoscopic color fundus photographs—an extension of the modified Airlie

House classification. ETDRS report number 10," Ophthalmology, vol. 98, no. 5, pp. 786–806, 1991.

[28] International Diabetes Federation, IDF Diabetes Atlas, 10th ed. Brussels, Belgium: IDF, 2021.

[29] M. Yanoff and J. S. Duker, Ophthalmology, 5th ed. Elsevier, 2019.

[30] V. Gulshan et al., "Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs," JAMA, vol. 316, no. 22, pp. 2402–2410, 2016

[31] T. Li et al., "ODIR-5K: A new ocular disease recognition benchmark dataset," in Proc. IEEE Int. Symp. Biomed. Imaging (ISBI), 2019, pp. 1034–1037.

[32] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization," *International Journal of Computer Vision*, vol. 128, no. 2, pp. 336–359, Feb. 2020, doi: https://doi.org/10.1007/s11263-019-01228-7

[33] X. Li, L. Shen, M. Shen, and C. S. Qiu, "Integrating Handcrafted and Deep Features for Optical Coherence Tomography Based Retinal Disease Classification," IEEE Access, vol. 7, pp. 33771–33777, 2019, doi: https://doi.org/10.1109/access.2019.2891975.

[34] N. Li, T. Li, C. Hu, K. Wang, and H. Kang, "A benchmark of ocular disease intelligent recognition: One shot for multi-disease detection," in International symposium on benchmarking, measuring and optimization, pp. 177–193, Springer, 2020

[35] Z. L. Teo, Y.-C. Tham, M. Yu, C.-Y. Cheng, T. Y. Wong, and C. Sabanayagam, "Do we have enough ophthalmologists to manage visionthreatening diabetic retinopathy? a global perspective," Eye, vol. 34, no. 7, pp. 1255–1261, 2020.

[36] M. Niemeijer, B. van Ginneken, S. R. Russell, M. S. Suttorp-Schulten, and M. D. Abramoff, "Automated detection and differentiation of drusen, exudates, and cotton-wool spots in digital color fundus photographs for diabetic retinopathy diagnosis," Investigative ophthalmology & visual science, vol. 48, no. 5, pp. 2260–2267, 2007.

[37] K. O'shea and R. Nash, "An introduction to convolutional neural networks," arXiv preprint arXiv:1511.08458, 2015.

[38] E. Jangam and C. S. R. Annavarapu, "A stacked ensemble for the detection of covid-19 with high recall and accuracy," Computers in Biology and Medicine, vol. 135, p. 104608, 2021.

[39] J. Faruk, S. B. Alam, S. S. T. Elma, S. Wasi, R. Rahman, and S. Kobashi, "Kidney abnormality detection using segmentation-guided classification on computed tomography images," in 2024 International Conference on Machine Learning and Cybernetics (ICMLC), pp. 414–419, IEEE, 2024

[40] T. R. Ahad, S. B. Alam, F. Ha-Mim, A. Nur, S. Wasi, and R. Rahman, “Comparative analysis of deep learning models for intracerebral hemorrhage detection from computed tomography,” in 2024 27th International Conference on Computer and Information Technology (ICCIT), pp. 328– 333, IEEE, 2024

[41] M. Wu et al., “Classification of dry and wet macular degeneration based on the ConvNeXT model,” Frontiers in computational neuroscience, vol. 16, Dec. 2022, doi: https://doi.org/10.3389/fncom.2022.1079155.

[42] F.T. J. Faria, M. B. Moin, P. Debnath, A. I. Fahim, F.M. Shah, “Explainable convolutional neural networks for retinal fundus classification and Cutting-Edge Segmentation Models for Retinal Blood Vessels from Fundus Images,” arXiv preprint, arXiv:2405.07338, 2024.

[43] R. B. J. Simanjuntak, Y. Fua, R. Magdalena, S. Saidah, A. B. Wiratama, ˆ I. Daa,ˆ et al., “Cataract classification based on fundus images using convolutional neural network,” JOIV: International Journal on Informatics Visualization, vol. 6, no. 1, pp. 33–38, 2022

[44] A.-O. Asia, C.-Z. Zhu, S. A. Althubiti, D. Al-Alimi, Y.-L. Xiao, P.- B. Ouyang, and M. A. Al-Qaness, “Detection of diabetic retinopathy in retinal fundus images using cnn classification models,” Electronics, vol. 11, no. 17, p. 2740, 2022.

[45] A. A. Jabber, A. Hadi, S. M. Wadi, G. A. Shadeed, “Cataract detection and classification using deep learning techniques,” International Journal of Computing and Digital Systems, vol. 16, no. 1, pp. 1–10, 2024.

[46] V. Thanikachalam , K. Kabilan , S. K. Erramchetty,“Optimized deep CNN for detection and classification of diabetic retinopathy and diabetic macular edema,” BMC Med. Imaging, vol. 24, no. 14, pp. 406–418, 2024.

[47] X. Ou, L. Gao, X. Quan, and H. Zhang, J. Yang, W. Li “BFENet: A two-stream interaction CNN method for multi-label ophthalmic diseases classification with bilateral fundus images,” Comput. Methods Programs Biomed., vol. 219, p. 106739, 2022.

[48] T.D. Nguyen, D.T. Le, J. Bum, S. Kim, S. J. Song, H. Choo, “Retinal disease diagnosis using deep learning on ultra-wide-field fundus images,” Translational Vision Science & Technology, vol. 13, no. 2, pp. 15–24, 2023.

[49] M. Wu et. al, “Classification of dry and wet macular degeneration based on the ConvNeXt model,” Front. Public Health, vol. 12, p. 1298, 2024.

[50] Z. Liu et al., “Swin Transformer: Hierarchical vision transformer using shifted windows,” in Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV), 2021, pp. 10012–10022

[51] V. Kotu and B. Deshpande, “Data science process,” Data science, pp. 19–37, 2019

[52] W. Bakasa and S. Viriri, “Stacked ensemble deep learning for pancreas cancer classification using extreme gradient boosting,” Frontiers in Artificial Intelligence, vol. 6, p. 1232640, 2023.

[53] G. Huo, Z. Lin, Z. Wang, R. Dai, and H. Tang, “Dms-net: Dual-modal multi-scale siamese network for binocular fundus image classification,” arXiv preprint arXiv:2504.18046, 2025.

[54] X. Ou, L. Gao, X. Quan, and H. Zhang, J. Yang, W. Li “BFENet: Bilateral feature enhancement network for multi-label ophthalmic disease classification,” Comput. Methods Programs Biomed., vol. 219, p. 106739, 2022.

[55] L. Abdel-Hamid, “Glaucoma detection from retinal images using statistical and textural wavelet features,” Journal of digital imaging, vol. 33, no. 1, pp. 151–158, 2020.

[56] U. Bhimavarapu, N. Chintalapudi, and G. Battineni, “Automatic detection and classification of hypertensive retinopathy with improved convolution neural network and improved svm,” Bioengineering, vol. 11, no. 1, p. 56, 2024.

[57] A. G. Tuwan, M. Iskandar, A. Y. Zakiyyah, and K. A. Minor, “Eye diseases multiclass classification using convolutional vision transformer,” Authorea Preprints, 2025.

[58] 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, pp. 248–255, Ieee, 2009

[59] D.To,“DucTo,” Apr.03,2022. https://tokudayo.github.io/from-resnet-to-convnext-2/ (accessed Nov. 14, 2025).

[60] D. Shah, M. A. U. Khan, M. Abrar, and M. Tahir, “Optimizing Breast Cancer Detection With an Ensemble Deep Learning Approach,” *International Journal of Intelligent Systems*,vol. 2024, no. 1, Jan. 2024, doi: https://doi.org/10.1155/2024/5564649.

[61] GeeksforGeeks, “Vision Transformer (ViT) Architecture,” *GeeksforGeeks*, Oct. 09, 2024. https://www.geeksforgeeks.org/deep-learning/vision-transformer-vit-architecture/

[62] A. Howard, M. Sandler, G. Chu, L.-C. Chen, B. Chen, M. Tan, W. Wang, Y. Zhu, R. Pang, V. Vasudevan, et al., “Searching for mobilenetv3,” in Proceedings of the IEEE/CVF international conference on computer vision, pp. 1314–1324, 2019.

[63] A. Dahou, A. O. Aseeri, A. Mabrouk, Rehab Ali Ibrahim, Mohammed Azmi Al-Betar, and Mohamed Abd Elaziz, “Optimal Skin Cancer Detection Model Using Transfer Learning and Dynamic-Opposite Hunger Games Search,” *Diagnostics*, vol. 13, no. 9, pp. 1579–1579, Apr. 2023, doi: https://doi.org/10.3390/diagnostics13091579.

[64] S. Zhou, Y.-W. Si, X. Yuan, X. Li, X. Liu, X. Zhang, C. Lin, and X. Gong, “Msconv: Multiplicative and subtractive convolution for face recognition,” 2025.

[65] C. Ju, A. Bibaut, and M. van der Laan, “The relative performance of ensemble methods with deep convolutional neural networks for image classification,” Journal of Applied Statistics, vol. 45, no. 15, pp. 2800–2818, 2018. PMID: 31631918

[66] P. Patil, G. Kale, N. Bivalkar, and A. Kolhatkar, “Comparative analysis of weighted ensemble and majority voting algorithms for intrusion detection in open stack cloud environments,” International Journal of Advanced Computer Science and Applications, vol. 14, no. 12, 2023

[67] T. Chompookham and O. Surinta, “Ensemble methods with deep convolutional neural networks for plant leaf recognition,” ICIC Express Letters, vol. 15, pp. 553–565, 06 2021