LsEyeNet: Deep Feature Fusion-based Models for a Large Spectrum of Eye Disease Recognition from Ophthalmic Images

By

Sakib Rokoni
24366051

A thesis submitted to the Department of Computer Science and Engineering
in partial fulfillment of the requirements for the degree of
M.Sc. in Computer Science and Engineering

Department of Computer Science and Engineering
BRAC University
July 2025

Declaration

It is hereby declared that

1. The thesis submitted is my original work while completing my degree at BRAC University.
2. The thesis does not contain material previously published or written by a third party, except where this is appropriately cited through full and accurate referencing.
3. The thesis does not contain material that has been accepted or submitted for any other degree or diploma at a university or other institution.
4. We have acknowledged all main sources of help.

Student's Full Name & Signature:

Sakib Rokoni
24366051

Approval

The thesis titled " LsEyeNet: Deep Feature Fusion-based Models for a Large Spectrum of Eye Disease Recognition from Ophthalmic Images "

By:
Sakib Rokoni (24366051)

Of Summer, 2025 has been accepted as satisfactory in partial fulfillment of the requirement for the degree of M.Sc. in Computer Science and Engineering on 13th July, 2025.

Thesis Examination Board:

Examiner:
(External)

Dr. Md. Sazzadur Rahman
Professor
Institute of Information Technology (IIT)
Jahangirnagar University

Examiner:
(Internal)

Dr. Muhammad Iqbal Hossain
Associate Professor
Department of Computer Science and Engineering
School of Data and Sciences
BRAC University

Supervisor:
(Member)

Dr. Md. Golam Rabiul Alam
Professor
Department of Computer Science and Engineering
School of Data and Sciences
BRAC University

Program Coordinator:
(Member)

Dr. Md. Sadek Ferdous
Professor
Department of Computer Science and Engineering
School of Data and Sciences
BRAC University

Chairperson:
(Chair)

Sadia Hamid Kazi, Ph.D
Associate Professor
Department of Computer Science and Engineering
School of Data and Sciences
BRAC University

Ethics Statement

This study utilizes the publicly accessible Eye Disease Image Dataset(Mendeley Data), ethically gathered with consent from two ophthalmic institutions in Bangladesh. All photos were anonymized to safeguard patient confidentiality, and the dataset adheres to institutional and international ethical guidelines for research involving human subjects. The dataset is licensed under **CC BY 4.0** for research purposes. The developed models are designed exclusively as decision-support tools and must not be utilized for autonomous clinical diagnosis without supervision from experienced specialists. Researchers must acknowledge the demographic constraints of the dataset and evaluate models across varied groups before implementation.

Abstract

Early and accurate identification of eye diseases is crucial for avoiding vision loss and providing appropriate treatments. Several researchers focus on eye disease detection, but limited spectrum. In this study, we have proposed 9 disease classification fusion-based models for a large spectrum of eye disease recognition. Here, we have extracted attention-based depth level features and utilized the ResNet50 architecture for eye disease detection. Secondly, we have utilized ConvNeXtBase and EfficientNetB3 for the latent space features, then fused them to a fully connected neural network for large-scale spectrum eye disease classification. To evaluate the performance of our proposed models, we have utilized the Eye Disease Image Dataset (Mendeley Data) and obtained superior accuracy. We have compared model performances with existing models and self-supervised approaches.

The hybrid LsEyeNet model performed the solo and baseline models, achieving 87.37 percent accuracy, high precision (0.89), strong recall (0.87), and the highest F1-score (0.89). These results demonstrate the efficacy of mixing contemporary convolutional architectures for accurate eye disease categorization. Furthermore, our findings reveal that, while self-supervised learning using SimCLR falls short of fully supervised techniques in a fully labeled context, it remains a promising strategy when annotated data is insufficient. This comprehensive performance benchmark highlights hybrid models' potential for developing trustworthy AI-based diagnostic tools in ophthalmology.

Keywords: Eye Disease Classification; Deep Learning; Convolutional Neural Networks; Transfer Learning; Self-Supervised Learning; SimCLR; ConvNeXt; EfficientNet; Hybrid Models; Medical Imaging; Ophthalmology; Fundus Images.

Dedication

This work is for my wonderful parents, whose unwavering support and encouragement have been the foundation of my journey; for my wonderful wife, whose love, patience, and understanding have been a constant source of strength; for my teachers and mentors, who have inspired me to pursue knowledge with passion and integrity; and for everyone who is working to make a difference in the field of medical research. I hope that this work will help improve people's health, even if only a little.

Acknowledgement

I thank the Almighty Allah for enabling me to finish the semester without experiencing any serious diseases or emergencies, both financially and physically.
Additionally, I would like to express my gratitude to my distinguished supervisor, **Dr. Md. Golam Rabiul Alam**, a professor in BRAC University's Department of Computer Science and Engineering, for his advice and assistance during the composition of this thesis and this article. I sincerely appreciate having him as my supervisor since his unwavering guidance and oversight helped me stay on course.
I would like to thank my wife and parents for their continuous support, which was invaluable at every step of my master's thesis.
Finally, I would like to thank BRAC University for giving me the tools and assistance I needed to carry out this research.

Table of Contents

List of Figures

List of Tables

Nomenclature

The following list describes several symbols & abbreviations used throughout the document.

α	Learning rate
$\hat{y}$	Predicted label
$\mathbf{b}$	Bias vector
$\mathbf{W}$	Weight matrix of a layer
$\mathcal{L}$	Loss function
$\sigma(\cdot)$	Activation function (e.g., ReLU, Sigmoid)
θ	Learnable parameters of the model
B	Batch size
E	Number of training epochs
$f(\cdot)$	Model function mapping input to output
X	Input image matrix
y	Ground truth label
AI	Artificial Intelligence
CNN	Convolutional Neural Network
CSC	Central Serous Chorioretinopathy
Disc Edema	Images with swollen optic disc
DR	Diabetic Retinopathy
ED	Eye Diseases
Glaucoma	Images indicating optic nerve damage
Healthy	Normal eye images
IoU	Intersection over Union
Macular Scar	Images with macular damage

Myopia Images showing nearsightedness

OCT Optical Coherence Tomography

Pterygium Images with conjunctival tissue overgrowth

RA Retinal Abnormalities

RD Images indicating Retinal Detachment

RD Retinal Detachment

RP Images showing Retinitis Pigmentosa

RP Retinitis Pigmentosa

SimCLR Simple Framework for Contrastive Learning of Visual Representations

SSL Self-Supervised Learning

Chapter 1

Introduction

The World Health Organisation (WHO) says that eye diseases and problems with vision are important worldwide public health issues that affect at least 2.2 billion people around the world. These cases may have been avoided, or they are still not being dealt with in more than 1 billion cases. Uncorrected refractive errors, cataracts, glaucoma, age-related macular degeneration, and diabetic retinopathy are some of the most common causes. Many people with these diseases can't work, have a lower quality of life, and have to pay more, especially in low- and middle-income countries, where 90 Percent of vision impairment cases happen. The World Health Organisation (WHO) argues that making eye care a part of primary health care systems, raising awareness, and making sure that everyone can get cheap treatments are all important steps towards getting everyone to have healthy eyes.

Deep learning techniques have made a big difference in the field of medical image analysis in the last few years. This advancement has notably influenced the categorisation of ocular illnesses from retinal or fundus pictures. Automated classification systems could help ophthalmologists find and diagnose problems earlier, which would cut down on their labour, make it more accurate, and make healthcare more accessible, especially in remote or impoverished areas. Glaucoma, diabetic retinopathy, myopia, and retinal detachment are just a few of the eye illnesses that can cause permanent blindness if they aren't found early. In the past, experienced specialists would have to look at the samples by hand, which may take a long time, be prone to mistakes, and not work for large screening programs. So, the necessity for an intelligent, automated system that can sort through a lot of different eye problems has become more important.

Deep learning, especially Convolutional Neural Networks (CNNs), has changed the way images are classified since they can automatically learn spatial hierarchies of data. Researchers have reached the highest level of performance in many image recognition benchmarks with architectures including ResNet, DenseNet, EfficientNet, and the new Vision Transformers. Nonetheless, supervised learning methodologies are significantly reliant on extensive annotated datasets, which are sometimes expensive and labour-intensive to develop within the medical field.

To mitigate this constraint, self-supervised learning (SSL) methodologies, exemplified as SimCLR, have arisen as formidable alternatives that utilise unlabelled data to acquire beneficial representations. The goal of this project is to create and compare several supervised models and add a self-supervised learning strategy to see how well they work for classifying numerous types of eye diseases.

1.1 Motivation

Eye diseases, including Diabetic Retinopathy, Glaucoma, Macular Scar, Retinal Detachment, and others, are major causes of preventable blindness worldwide. Early detection and timely intervention can drastically reduce vision loss and improve patient outcomes. However, accurate diagnosis often requires specialized equipment and expert ophthalmologists, which are scarce in many regions—particularly in low- and middle-income countries.

With the rapid advancement of deep learning, automated disease detection from retinal fundus images has emerged as a promising solution. Deep convolutional neural networks (CNNs) have demonstrated remarkable success in image classification tasks, including medical imaging. Transfer learning with pretrained networks further improves accuracy and reduces training time when annotated data is limited.

More recently, self-supervised learning (SSL) methods such as SimCLR have opened new avenues for medical imaging by leveraging unlabeled data, critical in domains where manual annotation is costly, time-consuming, or requires expert knowledge.

This thesis aims to investigate a range of deep learning models—from standard CNNs to modern architectures like ConvNeXt and EfficientNet—and integrate self-supervised learning and hybrid feature fusion strategies to build a robust eye disease classification system.

1.2 Problem Statement

Retinitis Pigmentosa, Diabetic Retinopathy, Glaucoma, and other eye diseases are primary causes of visual impairment and blindness on a global scale. The expeditious intervention and treatment of these diseases are contingent upon the early detection and accurate classification of these diseases, which can significantly enhance patient outcomes and alleviate the burden of blindness. Nevertheless, the manual diagnosis of retinal fundus images by ophthalmologists is time-consuming, necessitates a high level of expertise, and is susceptible to inter-observer variability. This renders large-scale screening impractical, particularly in regions with limited resources. Recent developments in deep learning have demonstrated the potential to automate the classification of retinal diseases from fundus images. Despite advancements, current models frequently encounter challenges such as class imbalance, limited labelled data, and variations in image quality as a result of varying patient conditions or acquisition devices. In addition, the majority of studies concentrate on individual models without conducting a systematic evaluation of the comparative performance of multiple state-of-the-art architectures or investigating the advantages of self-supervised learning to effectively utilise unlabelled data. Consequently, it is imperative to create and assess classification models based on deep learning, including self-supervised methods, to accurately and efficiently classify a variety of eye diseases from retinal fundus images. By constructing, comparing, and evaluating a variety of convolutional neural network (CNN) architectures—both standard and custom—as well as integrating self-supervised learning methods to improve performance in scenarios with limited labelled data, this thesis seeks to address these gaps.

1.3 Research Objectives

This research examined the problem of classifying eye diseases from multiple angles and with various objectives. To establish reliable performance benchmarks on a multi-class eye disease dataset, we conducted a thorough evaluation of various convolutional neural network (CNN) architectures, including a baseline CNN, ResNet50, DenseNet121, EfficientNetB3, ConvNeXtBase, and a customized ResNet50. This comparison showed that modern designs like EfficientNetB3 and ConvNeXtBase are better at capturing complex retinal features. We looked at how well the SimCLR self-supervised learning framework worked to see if it could help with performance when there wasn't much labeled data. Pretraining models on unlabeled eye pictures with SimCLR made the downstream classification far more accurate than fully supervised training. Based on what we learned, we came up with a new hybrid architecture called LsEyeNet that combines ConvNeXtBase and EfficientNetB3 to take advantage of their strengths. LsEyeNet has the best accuracy and generalization performance of all the models. Lastly, we used FLOPs to check how easy each model was to understand, how well it worked across different types of diseases, and how useful it would be in real-world clinical settings. This gave us a full picture of their strengths, weaknesses, and possible biases for use in ophthalmology.

- Evaluate and compare the performance of multiple convolutional neural network (CNN) architectures—including a baseline CNN, ResNet50, DenseNet121, EfficientNetB3, ConvNeXtBase, and a custom ResNet50—on a multi-class eye disease classification task to establish robust performance benchmarks.
- Investigate the effectiveness of the SimCLR framework in improving downstream classification accuracy over fully supervised models, and utilize self-supervised learning for efficient pretraining on unlabeled eye images.
- Design and implement a novel LsEyeNet model that combines ConvNeXtBase and EfficientNetB3 to leverage complementary feature representations and enhance classification performance.
- Evaluate the proposed models in terms of generalization, interpretability, and suitability for clinical use by analyzing their strengths, limitations, and potential biases across diverse eye disease categories.

1.4 Contribution

In this research, three advanced deep learning strategies were proposed to enhance the accuracy and reliability of automated eye disease classification: an Attention-based ResNet50, a LsEyeNet model, and a SimCLR-based self-supervised approach. The Custom-based ResNet was architecturally modified to include domain-specific convolutional blocks and attention mechanisms to better capture retinal features, improving its ability to distinguish between visually similar diseases. The LsEyeNet model effectively combined convolutional feature extraction with transformer-based attention to capture both local textures and global structural patterns within fundus images. SimCLR, a contrastive self-supervised learning technique, was employed to learn rich and generalizable representations from unlabeled retinal images, boosting

classification performance, particularly in low-data regimes. Together, these models contributed to improving classification robustness, interpretability, and performance across both majority and minority disease classes.

- This research proposed an Attention-Based ResNet model tailored for retinal image classification. By modifying convolutional layers and integrating attention mechanisms, the model effectively captured disease-specific features and outperformed the standard ResNet50, especially in detecting complex cases like Glaucoma and Diabetic Retinopathy.
- This work presents a LsEyeNet architecture that integrates two powerful CNN backbones—ConvNeXtBase and EfficientNetB3—to leverage their complementary feature extraction capabilities. By fusing their global average pooled outputs, the model captures both high-level semantic features and fine-grained texture details. The dual-path design enhances feature diversity and robustness, leading to improved classification accuracy compared to using a single backbone. This architecture is especially effective for tasks where rich multi-scale representations are critical, and it maintains a balance between performance and computational efficiency through backbone freezing and optimized dense layers.
- The study also applied SimCLR, a self-supervised learning technique, to learn representations from unlabeled data. This approach enhanced feature quality, improved generalization in low-data settings, and showed strong performance in identifying rare diseases without heavy reliance on annotated datasets.

1.5 Organization of the Report

The rest of this thesis is organized in this way: In Chapter 2, linked work is looked at. This includes past research on using deep learning to find eye diseases, transfer learning, and self-supervised learning methods in medical imaging. Chapter 3 gives an overview of the dataset's sources, features, and preprocessing methods. It also talks about the deep learning models that were used, how they were trained, and the measures that were used to judge their performance. In Chapter 4, the results of the execution are analyzed and talked about. Finally, Chapter 5 wraps up the thesis by going over the main points of the study and suggesting areas for more research.

Chapter 2

Related Work

The early identification and precise categorization of ocular disorders from retinal fundus images have been a prominent study domain, motivated by the capacity to diminish preventable blindness. Conventional computer vision methodologies depended on manually designed features and traditional classifiers. Acharya et al. [2] utilized discrete wavelet transform characteristics in conjunction with support vector machines (SVMs) to detect diabetic retinopathy lesions. Nevertheless, these strategies frequently exhibited restricted generalizability owing to their reliance on features defined by experts. The emergence of deep learning, especially convolutional neural networks (CNNs), transformed medical image processing. Gulshan et al.[9] exhibited that a CNN, trained on an extensive dataset of retinal fundus images, attained performance akin to that of ophthalmologists in identifying diabetic retinopathy. Pratt et al. [10] optimized deep CNN architectures such as VGG16 on fundus pictures, demonstrating the viability of transfer learning in ophthalmology.

ResNet-based models demonstrate potential owing to their capacity to train deeper networks without encountering vanishing gradients. Li et al. [1] utilized ResNet50 for the multi-class classification of retinal disorders inside the ODIR dataset, with commendable results. DenseNet, characterized by feature reuse through dense connectivity, has been utilized; Ramachandra et al.[25] indicated that DenseNet121 had strong performance in glaucoma detection tests. EfficientNet, a contemporary architecture that consistently scales depth, width, and resolution, has established new performance standards on ImageNet and shown effectiveness in medical imaging applications. Tan and Le presented EfficientNetB0-B7, whereas subsequent research by Chen et al.[30] utilized EfficientNetB3 for diabetic retinopathy grading, achieving enhanced accuracy and computing efficiency. Transformer-based models and self-attention processes have arisen as alternatives to conventional CNNs. Dosovitskiy et al. introduced the Vision Transformer (ViT), which segments images into patches and processes them as a sequential input. Initially designed for natural images, researchers such as Li et al. [31] modified ViT for fundus image classification, attaining performance on par with CNN baselines.

Contrastive learning frameworks, such as SimCLR (Chen et al.[20], have gained prominence in self-supervised learning (SSL). SimCLR acquires significant visual representations by optimizing the concordance between various enhanced perspectives of the identical image, independent of labeling [20]. Saba et al. demonstrated

that SimCLR-pretrained encoders enhanced performance in downstream illness classification tasks applied to retinal pictures, particularly in scenarios with limited labeled data [32]. Numerous extensive public datasets have facilitated advancements in the study. The EyePACS and Messidor datasets from Kaggle are extensively utilized for grading diabetic retinopathy, whereas the ODIR dataset offers a multi-label classification baseline for prevalent retinal illnesses [6] [16]. Rajalakshmi et al. presented the ARIA dataset, which comprises labeled images of diabetic retinopathy and normal eyes, facilitating algorithm development [13].

Recent comparative analyses have assessed CNNs in relation to transformer models. Touvron et al. [29] presented DeiT (Data-efficient Image Transformer) and shown that, with effective distillation, transformers can achieve comparable performance to CNNs with a reduced number of labeled pictures . In tasks involving high-resolution images and nuanced clinical characteristics, CNNs or hybrid CNN-transformer models frequently surpass pure transformers [26]. Moreover, combining attention modules into CNNs enhances interpretability and performance. Woo et al.introduced the Convolutional Block Attention Module (CBAM), which has been utilized in fundus image categorization to improve attention on lesion regions [14]. In the past, medical imaging relied on physicians to look at the images by hand, which took a lot of time, was subjective, and may vary from person to person. In the early 2000s, computer-aided diagnostic (CAD) systems came out to help doctors. They used feature engineering methods like texture, color histograms, and morphological descriptors. These methods, on the other hand, needed expertise particular to the field and had trouble applying to other populations and imaging systems [2],[24].
Deep learning, especially convolutional neural networks (CNNs), changed this by letting computers learn from raw images all the way through. CNNs develop hierarchical feature representations: early layers record edges and textures, while deeper levels abstract meaningful information like lesions or anatomical structures [7]. Architectures like AlexNet [4], VGGNet [5], and Inception [[8] made CNNs the best choice for natural picture tasks. Their success quickly spread to medical fields, including ophthalmology. Transfer learning lets CNNs that have already been trained on big datasets like ImageNet [3] adapt to smaller medical datasets by making minor changes. This method cuts down on the requirement for huge, expensive, and hard-to-make medical datasets with plenty of annotations [11]. Contrastive learning and other self-supervised learning (SSL) methods use unlabeled data to learn strong representations. Chen et al. came up with SimCLR, a framework that maximizes the similarity between several views of the same image. It has worked well on medical images, such as fundus scans [19], [20]. Attention mechanisms were first designed for natural language processing (NLP) and let models focus on the parts of an input that include the most useful information. Squeeze-and-Excitation networks [12] and CBAM combine channel and spatial attention to make tasks like finding retinal diseases [21] easier to understand and more accurate. Transformers, which became popular because of the Vision Transformer[31], simulate global context using self-attention instead of convolution. Researchers have looked into using hybrid models that combine CNN backbones with transformer blocks for medical imaging. These models use both local texture information and global dependencies [23]. Azizi et al. [22]showed that pretraining with contrastive approaches on unlabeled medical images increased performance in retinal disease classification for SSL in ophthalmol-

ogy, even when there weren't many labeled examples available. Tajbakhsh et al. [15] did a thorough analysis of SSL uses in medical imaging, focusing on how useful it could be in contexts with limited resources. Another area of active research is domain adaptability. Raghu et al. looked at how well CNNs trained on ImageNet work with retinal images. They found that some features work better with retinal images than with natural images [17]. Bissoto et al. looked into ways to generalize fundus images from diverse devices across domains [18].

In conclusion, research has progressed from manual techniques to deep convolutional neural networks, transfer learning, and more recently, self-supervised and transformer-based methodologies. Nonetheless, obstacles persist in domain adaptability, interpretability, and efficacy concerning rare diseases with scarce training samples, necessitating further investigation in these domains.

Table 2.1: Summary of Approaches for Eye Diseases Detection

Aspect	**Details**
Traditional Methods	Hand-crafted features with traditional classifiers (e.g., SVMs); e.g., Discrete Wavelet Transform used by Acharya et al.; limited generalizability due to expert-defined features.
Deep Learning Era	CNNs (e.g., AlexNet, VGGNet, Inception) learn hierarchical representations; revolutionized medical image diagnosis.
Transfer Learning	CNNs pre-trained on ImageNet adapted to medical datasets; reduces the need for large annotated datasets.
Key CNN Models Used	VGG16 [10] for fundus image classification; ResNet50 (Li et al.) for multi-class classification on ODIR dataset; DenseNet121 (Ramachandra et al.) for glaucoma detection.
Efficient Architectures	EfficientNetB0-B7 [15]; EfficientNetB3 (Chen et al.) for diabetic retinopathy grading with high accuracy and efficiency.
Transformer Models	Vision Transformer (ViT) by Dosovitskiy et al., adapted by Li et al. for fundus images; DeiT (Touvron et al.) is a data-efficient variant.
Hybrid CNN-Transformer	Combines CNN's local feature extraction with Transformer's global context modeling; effective for high-resolution clinical images.
Self-Supervised Learning	SimCLR [23] learns representations from unlabeled data; improves performance with limited labels (Saba et al., Azizi et al.).
Attention Mechanisms	CBAM (Woo et al.) enhances focus on lesion regions; Squeeze-and-Excitation networks improve interpretability and accuracy.
Major Datasets	EyePACS, Messidor (diabetic retinopathy grading); ODIR (multi-label retinal diseases); ARIA (DR vs. normal fundus images).
Challenges	Domain adaptability (device/population variance), interpretability of deep models, and scarcity of data for rare diseases.
Key Trends	Shift from manual features to end-to-end DL; rise of transformer and hybrid models; increasing use of SSL to reduce annotation dependence.

Chapter 3

Methodology

3.1 Dataset Description

This dataset contains a large collection of ophthalmic photos designed to aid research into automated eye disease identification using machine learning and computer vision. The information was gathered over eight months from two well-known medical institutions: Anawara Hamida Eye Hospital and B.N.S.B. Zahurul Haque Eye Hospital.[36] All photographs were captured utilising a Colour Fundus Photography machine in a healthcare setting. The collection contains 5,335 high-resolution photos divided into ten clinically important classifications that cover both retinal and anterior segment illnesses. It consists of two types of photos: colour fundus photographs and anterior segment images. The colour fundus images depict a wide range of retinal illnesses, whereas the anterior segment images focus on a specific anterior problem. The dataset used in this thesis comprises retinal fundus images categorized into 10 distinct eye disease classes. They are Healthy, Retinitis Pigmentosa, Retinal Detachment, Pterygium, Myopia, Macular Scar, Glaucoma, Disc Edema, Diabetic Retinopathy, Central Serous Chorioretinopathy.

3.2 Data Preprocessing

To prepare the images for training, a rigorous preprocessing pipeline was implemented. The preprocessing involved the following steps:
Image Resizing: All images were resized to 224×224 pixels to ensure compatibility with standard CNN architectures like ResNet50 and EfficientNetB3.
Normalization: Pixel values were normalized using the mean and standard deviation of the ImageNet dataset to align with the pretraining of the models used.
Data Augmentation: Augmentation techniques were employed to artificially expand the dataset and improve model generalization: Random horizontal and vertical flips; Rotation (±20 degrees); Random cropping and zoom and Color jitter (brightness, contrast, saturation). Augmentations were applied on the fly during training using PyTorch's torchvision transforms pipeline.

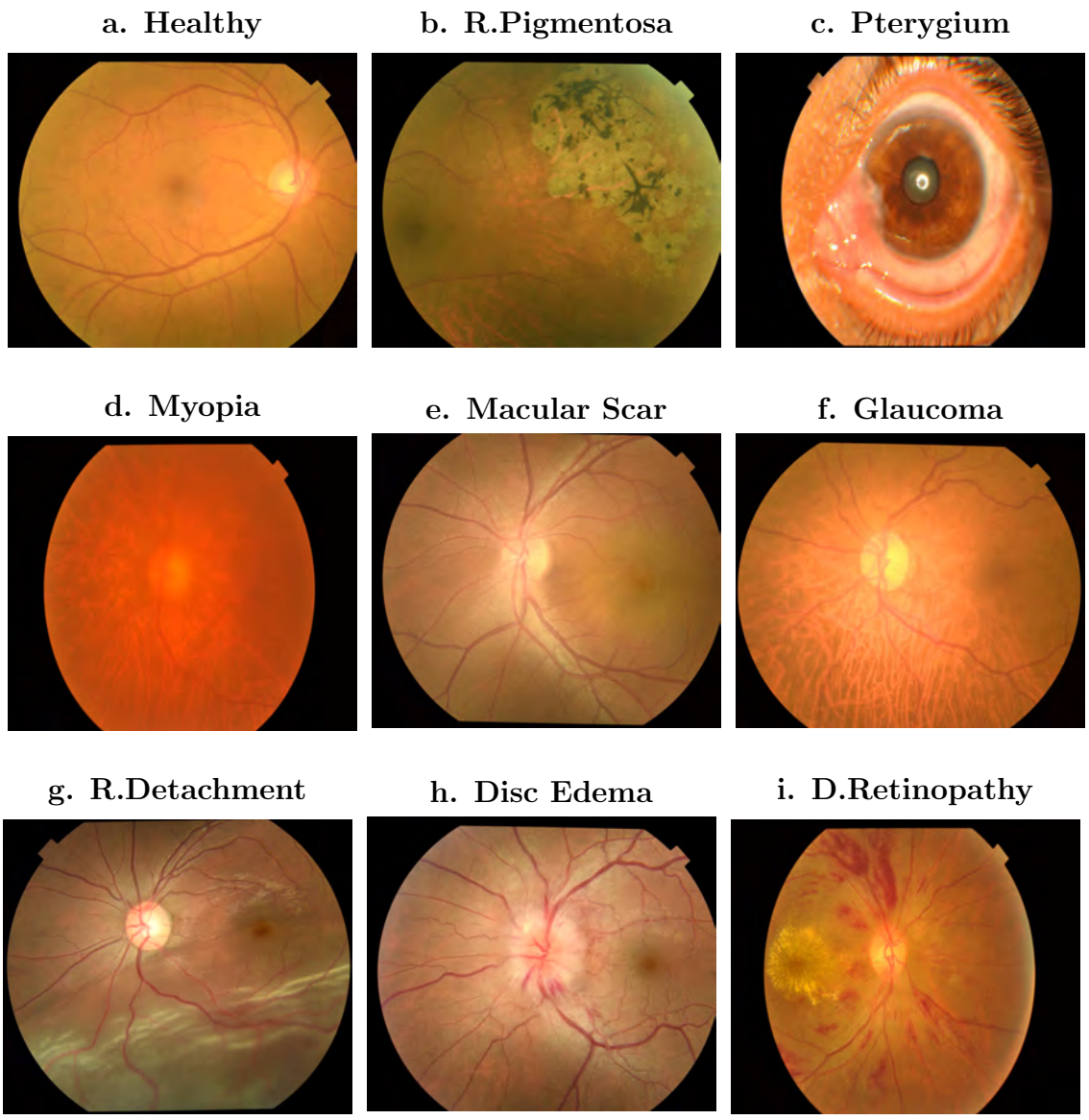


Figure 3.1: Sample Images from Nine Classes in Our Eye Disease Dataset

3.3 Dataset Splitting

The dataset was split into training, validation, and test subsets in the following ratio: training set 70 percent; validation set 15 percent and test set 15 percent. Stratified splitting was used to maintain consistency in class distribution across the splits. The training set was used to train the model weights, while the validation set was used for hyperparameter tuning and early stopping. The test set remained untouched during training and was used solely for final performance evaluation.

Table 3.1: Data Distribution for Eye Disease Classes

Condition	**Train Images**	**Val Images**	**Test Images**
Central Serous Chorioretinopathy	424	90	92
Diabetic Retinopathy	2410	516	518
Disc Edema	533	114	115
Glaucoma	2015	432	433
Healthy	1873	401	402
Macular Scar	1355	290	292
Myopia	1575	337	339
Pterygium	71	15	16
Retinal Detachment	525	112	113
Retinitis Pigmentosa	583	125	126

3.4 Dataset Challenges

There were a number of problems in handling the dataset. It was clear that there was an imbalance in the classes because there were less samples for rare disorders like Macular Scar and Pterygium. To fix this, we used approaches like data augmentation and class-weighted loss. It was harder to classify disorders like Myopia and Central Serous Chorioretinopathy since they looked identical. Also, problems with image quality, like bad lighting, blur, and occlusion, made the model even less accurate.

3.5 Model Architectures

The illustrated workflow depicts a complete pipeline for automated eye disease categorization utilizing retinal pictures. It starts with an eye dataset, which goes through preparatory stages including image scaling and data separation. Random flips, rotation, cropping, zooming, and color jitter are examples of augmentation techniques used to improve model performance and generalization. The enriched data is then fed into two simultaneous feature extraction paths: one using pre-trained convolutional neural networks (CNNs) such as ResNet50, DenseNet121, EfficientNetB3, and ConvNeXtBase, and the other using a self-supervised learning model called SimCLR. These extracted characteristics are then combined into a LsEyeNet model that incorporates an attention-based ResNet50 and an ensemble of ConvNeXtBase and EfficientNetB3. Finally, the model uses multi-class classification to identify several eye disorders such as CSC, DR, DE, Glaucoma, Macular Scar, Myopia, Pterygium, RD, and RP, allowing for precise and rapid diagnosis.

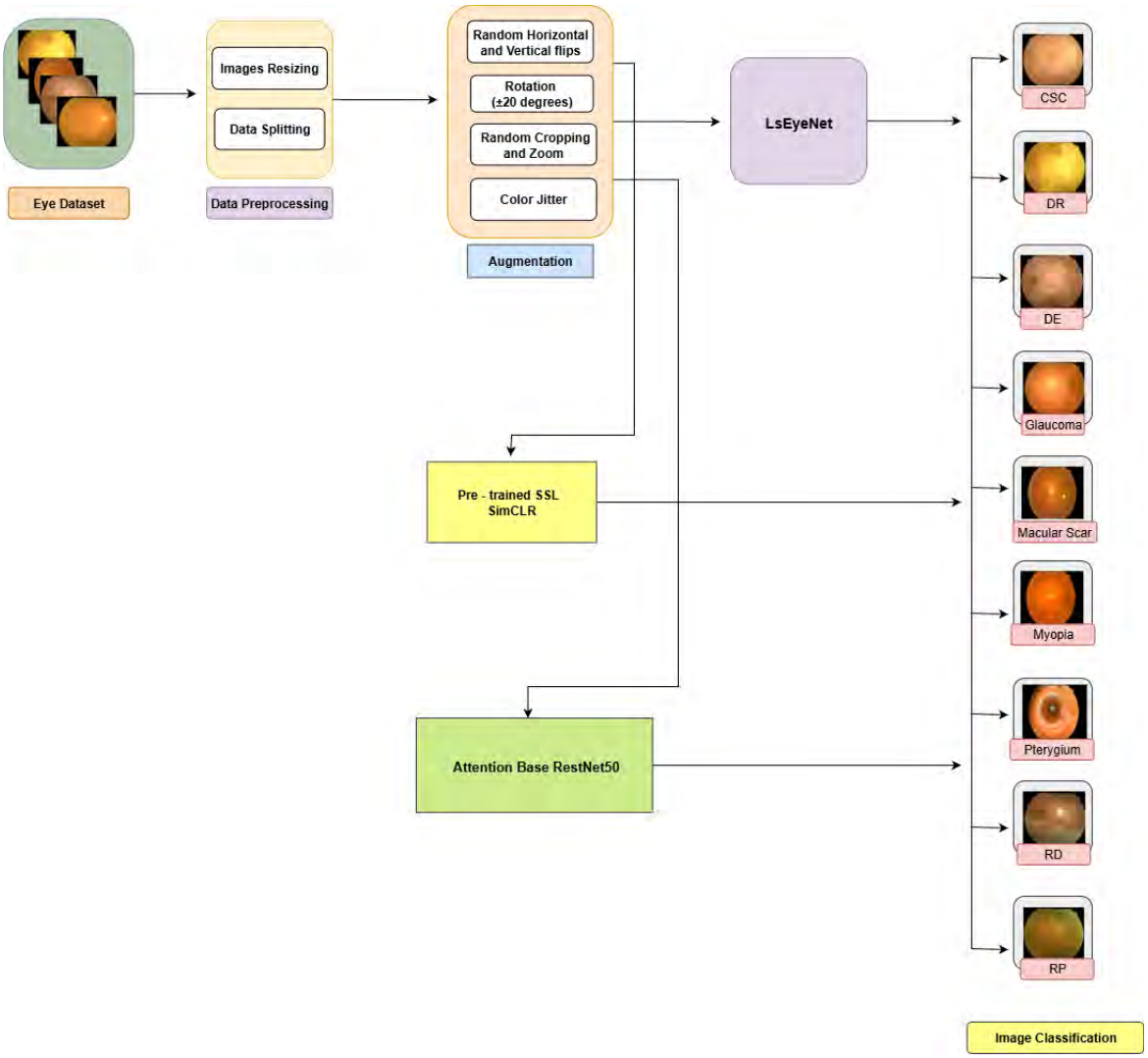


Figure 3.2: Complete Block Diagram for Eye Disease Classification

3.5.1 Convolutional Neural Network (CNN)

A Convolutional Neural Network (CNN) is a type of deep learning architecture primarily used for image classification and recognition tasks. CNNs consist of several convolutional layers, pooling layers, and fully connected layers. These layers help in extracting features from images and classifying them. Convolution operation applied using kernels to extract features. Reduces the spatial dimensions of the image. Introduces non-linearity with the activation function:

$$f(x) = \max(0, x) \tag{3.1}$$

A final classification layer combining features from the convolutional layers.

$$(f * g)(x) = \sum_{i=-\infty}^{\infty} f(i)g(x - i)....1 \tag{3.2}$$

Where f is the input image and g is the filter (kernel).

$$y_{i,j} = \max\left(x_{2i,2j}, x_{2i+1,2j}, x_{2i,2j+1}, x_{2i+1,2j+1}\right) \tag{3.3}$$

Where x is the input image and y is the pooled output.

$$f(x) = \max(0, x) \tag{3.4}$$

$$y = W \cdot x + b \tag{3.5}$$

Where W is the weight matrix, x is the input, and b is the bias.

3.5.2 ResNet50

ResNet50 is a deep neural network architecture that uses residual learning to allow the network to be much deeper without the vanishing gradient problem. It introduces residual blocks where the input is added to the output of the block via shortcut connections. The output of the residual block is computed as:

$$\mathbf{y} = F(\mathbf{x}, \{W_i\}) + \mathbf{x} \tag{3.6}$$

Where $F(\mathbf{x}, \{W_i\})$ is the convolution operation and $\mathbf{x}$ is the input to the block. After passing through the residual blocks, a global average pooling layer is applied, followed by a fully connected layer.

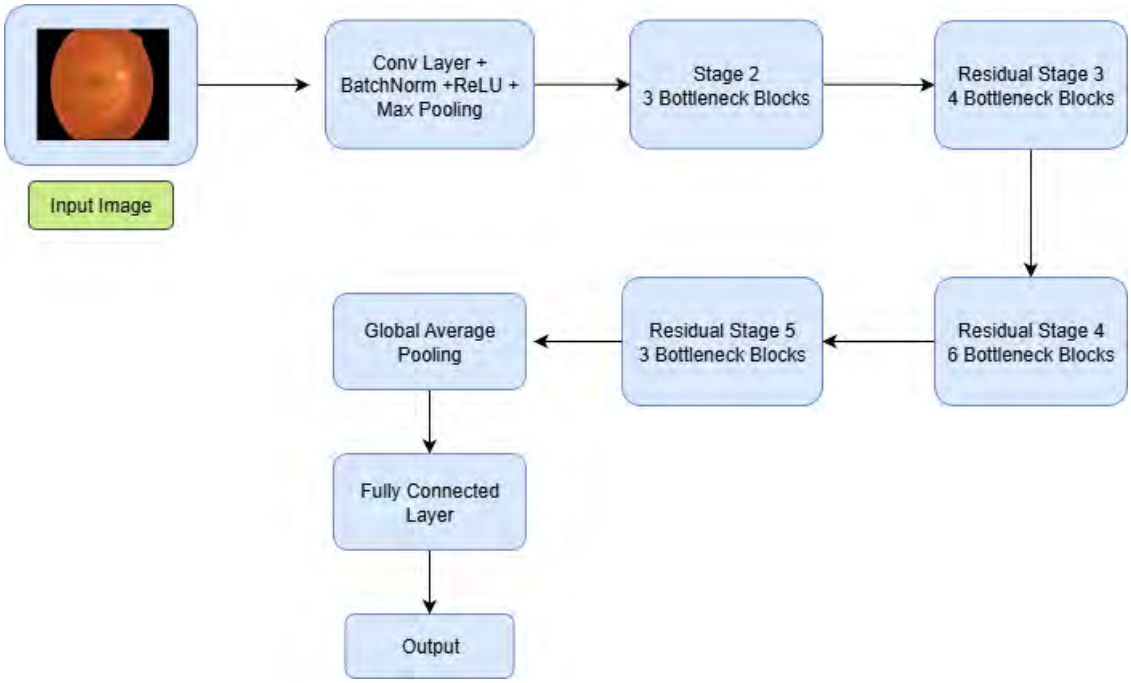


Figure 3.3: ResNet50 Architecture

3.5.3 DenseNet121

DenseNet121 is a deep convolutional network where each layer is connected to every other layer in a feed-forward manner. This dense connectivity improves gradient flow and reduces redundancy. Each layer receives inputs from all previous layers. The output of the k-th layer in a dense block is:

$$\mathbf{y}_k = \mathcal{H}(\mathbf{x}_k, \{W_i\}) \quad \text{for} \quad k = 1, 2, ..., N \tag{3.7}$$

Where $\mathcal{H}$ represents the convolution operation, and the output is concatenated with the input for the next layer:

$$\mathbf{x}_k = [\mathbf{x}_0, \mathbf{y}_1, \mathbf{y}_2, ..., \mathbf{y}_k] \tag{3.8}$$

Where $\mathbf{x}_0$ is the initial input to the block.

3.5.4 EfficientNetB3

EfficientNet utilizes compound scaling to scale the depth, width, and resolution of the network in a balanced way, leading to high accuracy with fewer computational resources. EfficientNet scales the depth α, width β, and resolution ϕ using the following formulas:

$$d_{\text{new}} = \alpha d_{\text{base}}, \quad w_{\text{new}} = \beta w_{\text{base}}, \quad r_{\text{new}} = \phi r_{\text{base}} \tag{3.9}$$

Where: - d is the depth (number of layers), - w is the width (number of channels), - r is the resolution (input image size).

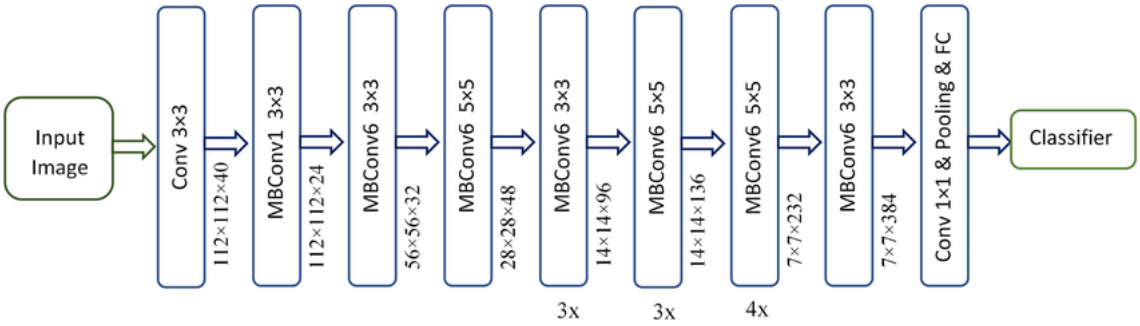


Figure 3.4: EfficientNetB3

3.5.5 ConvNeXtBase

ConvNeXtBase is a modernized version of ResNet with elements inspired by Vision Transformers. It incorporates large kernels, layer normalization, and GELU activations for improved performance. Layer Normalization is applied after every convolutional block to stabilize the training process. The GELU activation function is given by:

$$\text{GELU}(x) = 0.5x \left(1 + \tanh\left(\sqrt{\frac{2}{\pi}}(x + 0.044715x^3)\right)\right) \tag{3.10}$$

3.5.6 Attention Based ResNet50

This is a modified version of ResNet50 that includes additional dropout layers for regularization and optional attention mechanisms to improve feature selection. Dropout Layer randomly drops neurons during training to prevent overfitting:

$$\mathbf{y} = \mathbf{W} \cdot \mathbf{x} + \mathbf{b} \tag{3.11}$$

With probability p of being dropped. The attention mechanism is modeled as:

$$\mathbf{y} = \text{softmax}(\mathbf{W} \cdot \mathbf{x}) \tag{3.12}$$

Where the attention weights are learned during training.

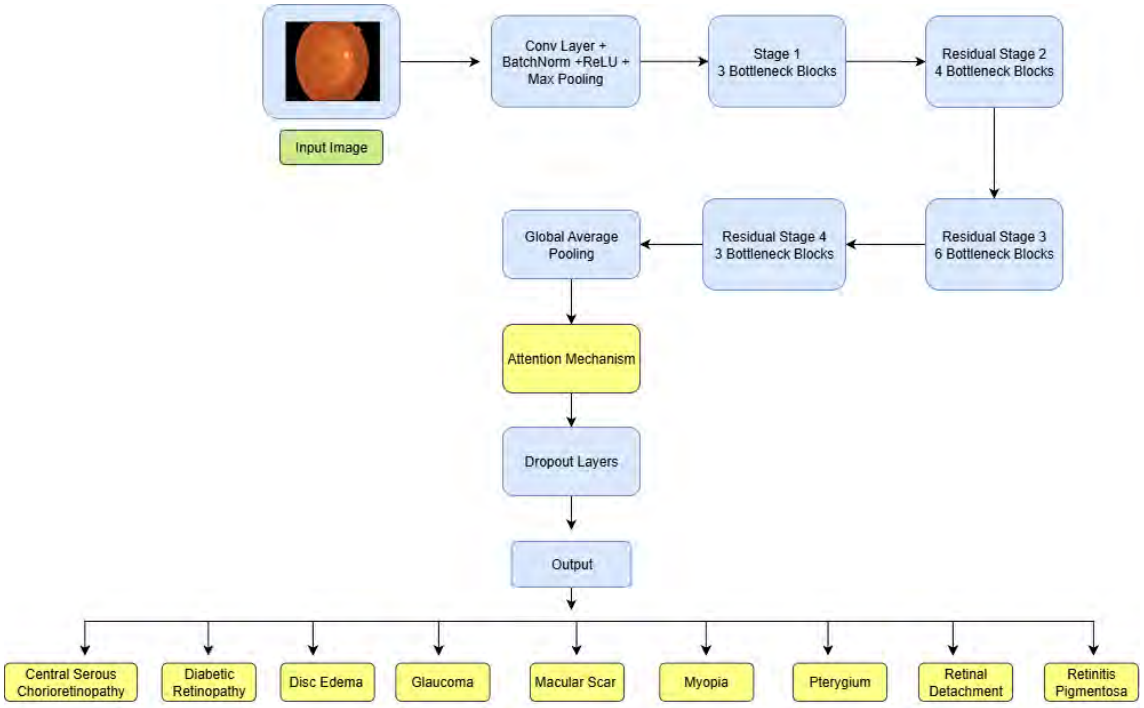


Figure 3.5: Attention Base ResNet50 Architecture

3.5.7 SimCLR (SSL)

Using contrastive learning as its foundation, SimCLR is a framework for self-supervised learning. It maximises the similarity between enhanced image pairs to acquire meaningful representations. A ResNet50-based encoder that maps augmented images to feature vectors. Projection Head Maps feature vectors into a latent space for contrastive learning. Contrastive Loss (NT-Xent Loss) :

$$\mathcal{L}_{\text{NT-Xent}} = -\log\left(\frac{\exp(\text{sim}(\mathbf{z}_i, \mathbf{z}_j)/\tau)}{\sum_{k=1}^{N} \exp(\text{sim}(\mathbf{z}_i, \mathbf{z}_k)/\tau)}\right) \tag{3.13}$$

Where $\mathbf{z}_i$ and $\mathbf{z}_j$ are feature representations of augmented images, and $\text{sim}(\mathbf{z}_i, \mathbf{z}_j)$ is the cosine similarity between them.

3.5.8 Proposed LsEyeNet Model

The LsEyeNet Model utilizes two robust deep learning architectures, ConvNeXtBase and EfficientNetB3, to enhance image classification performance. This combination leverages the best aspects of both models, incorporating additional features and representations to create a classifier that is more accurate and effective.

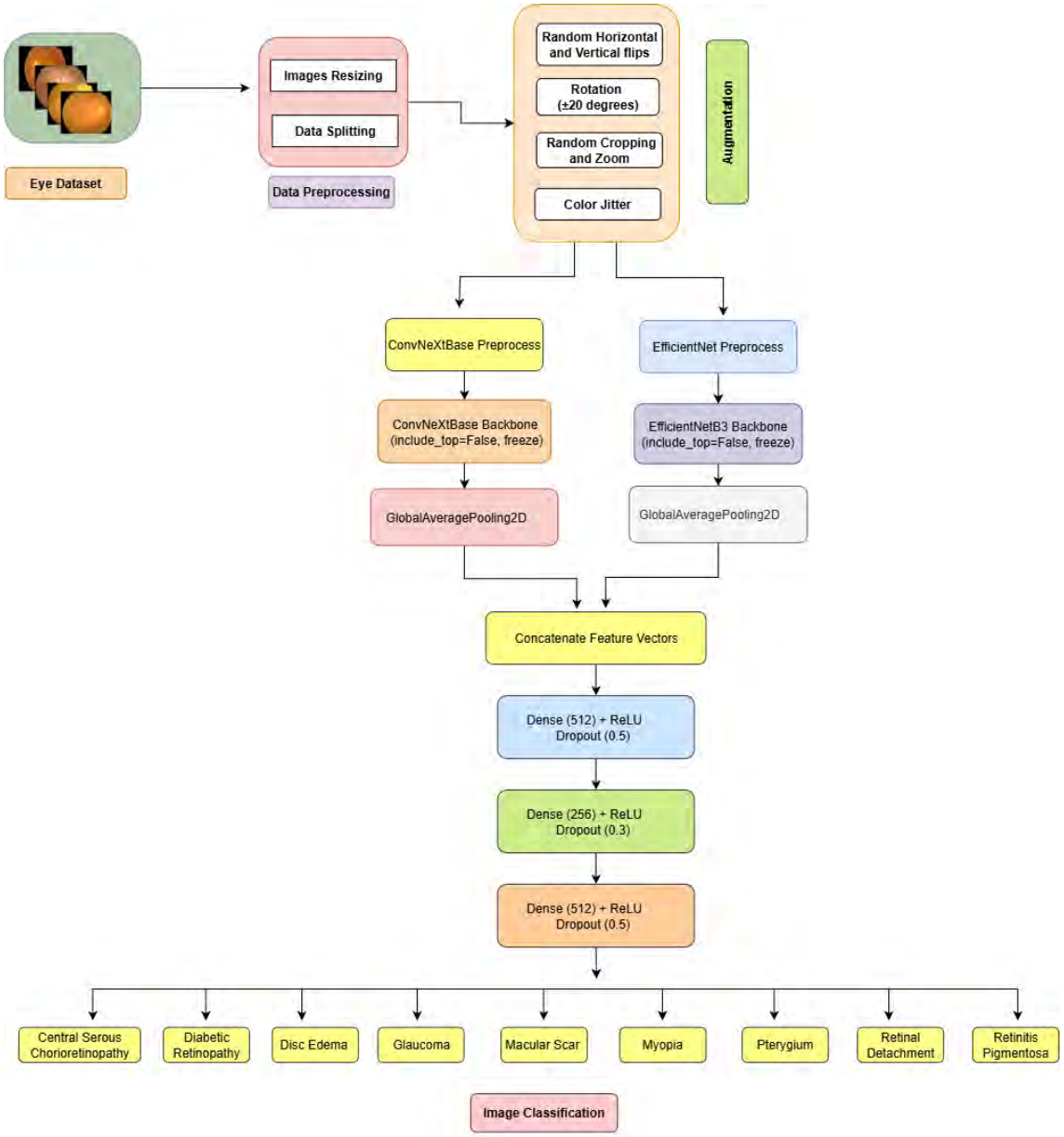


Figure 3.6: LsEyeNet Model Architecture

ConvNeXtBase Backbone:
ConvNeXtBase is a state-of-the-art model that draws inspiration from both classical convolutional neural networks (CNNs) and modern Vision Transformers (ViTs). The key features of ConvNeXtBase include: Large Kernels: Uses larger kernels in convolution layers, allowing the model to capture more contextual information.

Layer Normalization: Applied to stabilize and speed up training by reducing internal covariate shifts. GELU Activation: The GELU (Gaussian Error Linear Units) activation function is more efficient and improves performance in deeper networks. The input image is processed through the convolutional layers of ConvNeXtBase, which extract both low-level and high-level information. These properties encompass a broad spectrum of patterns, from basic edges to more complex representations, including shapes and objects.

EfficientNetB3 Backbone:
EfficientNetB3 is an improved variant of EfficientNet, which is renowned for its outstanding performance and computational effectiveness. EfficientNetB3 employs compound scaling, which reduces computational costs while maintaining high accuracy by scaling depth, breadth, and resolution. Key features of EfficientNetB3 include: Compound Scaling: Scales all dimensions of the model (depth, width, and resolution) simultaneously to maintain a balance between accuracy and computational efficiency. Depthwise Convolutions: Reduces the computational load while capturing global features. Swish Activation: The Swish activation function provides smoother gradients and better performance in deeper networks. The input image is processed using EfficientNetB3, which extracts aspects that capture global data while being computationally efficient.

Feature Concatenation
Once both ConvNeXtBase and EfficientNetB3 have worked on the input image, the features they found are combined into one feature vector. This concatenation takes the best parts of both architectures and combines them to capture a wide range of properties.The process of concatenation can be expressed mathematically as:

$$\mathbf{F}_{\text{fusion}} = [\mathbf{F}_{\text{ConvNeXt}}, \mathbf{F}_{\text{EfficientNet}}] \tag{3.14}$$

Where: $\mathbf{F}_{\text{ConvNeXt}}$ is the feature vector extracted by ConvNeXtBase. $\mathbf{F}_{\text{EfficientNet}}$ is the feature vector extracted by EfficientNetB3. $\mathbf{F}_{\text{fusion}}$ is the fused feature vector containing the combined information from both backbones. The feature fusion step enables the model to leverage the strengths of both ConvNeXtBase and EfficientNetB3.

Final Classifier
The final prediction is made by passing the fused feature vector through a final fully connected layer the classifier after the characteristics are combined. The final output is calculated by the classifier using the concatenated features:

$$\mathbf{y} = \text{softmax}(\mathbf{W} \cdot \mathbf{F}_{\text{fusion}} + \mathbf{b}) \tag{3.15}$$

Where: $\mathbf{W}$ is the weight matrix of the classifier. $\mathbf{F}_{\text{fusion}}$ is the concatenated feature vector; $\mathbf{b}$ is the bias term; softmax is the activation function used to produce a probability distribution over the class labels. The **softmax** activation function is defined as:

$$\text{softmax}(\mathbf{y})_i = \frac{\exp(\mathbf{y}_i)}{\sum_j \exp(\mathbf{y}_j)} \tag{3.16}$$

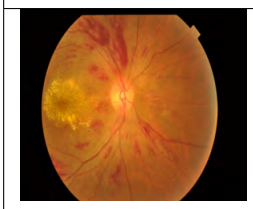
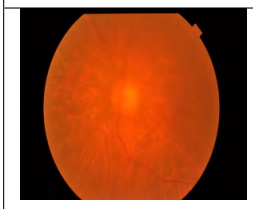
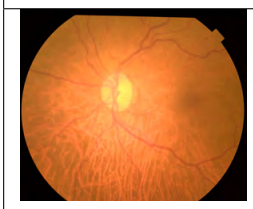
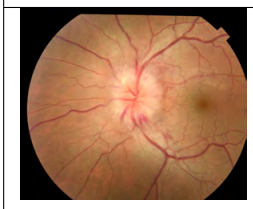
Where $\mathbf{y}_i$ is the i-th element of the output vector, corresponding to the predicted probability for class i.

Advantages of the LsEyeNet Model
The LsEyeNet Model benefits from the combination of ConvNeXtBase and EfficientNetB3. Some of the key advantages include: Improved Performance by combining the outputs of both networks, the hybrid model captures a more diverse set of features, resulting in better performance, especially on complex image classification tasks. The model is more robust because the two backbones focus on different aspects of the image (local vs global features), which allows for better generalization. Efficiency helps EfficientNetB3 to maintain computational efficiency, ensuring that the model can process large images and datasets without excessive computational cost. Flexibility helps the hybrid model be adaptable to a wide variety of image classification tasks, making it suitable for real-world applications with varying data types and complexities. The workflow of the LsEyeNe Model can be visualized as follows: The input image is processed by both ConvNeXtBase and EfficientNetB3 to extract features. The features from both models are concatenated into a single feature vector. The concatenated features are passed through a fully connected layer to make the final prediction. ConvNeXtBase and EfficientNetB3 are nicely combined in the Hybrid Model, which takes advantage of the two architectures. Besides achieving higher precision and recall, such synergy enables the model to accept a wider variety of image inputs. For present image classification applications, which require a combo of power performance and efficiency, we believe that you cannot do much better.

3.5.9 LsEyeNet Model: Fusion Strategy

Both the ConvNeXt model and the LsEyeNet model that mixes it with EfficientNetB3 are very different from each other. These differences can be seen in how they are built, how well they learn, and how much computer power they need.
The ConvNeXt model is a contemporary convolutional neural network that is based on ResNet and has been designed with a more straightforward and clear architecture. Its primary advantage is its computational efficiency and lightweight design, which facilitates quicker training and reduces the likelihood of overfitting. ConvNeXt concentrates on the optimization of fundamental convolutional operations to extract useful features while maintaining a simpler model architecture, utilizing only one encoder or backbone. This renders it appropriate for tasks in which the dataset size is modest or where computational resources are restricted. Nevertheless, ConvNeXt may encounter difficulty in capturing highly detailed or multi-scale patterns in complex images due to its reliance on a single feature extractor. This can restrict its performance in tasks that necessitate a wide range of feature comprehension.
In contrast, the LsEyeNet model, which integrates two encoders, provides a more potent alternative by combining ConvNeXt with EfficientNetB3. The LsEyeNet model can learn from both architectures as a result of this dual-encoder configuration, which allows for a more comprehensive feature extraction and a more accurate representation of both low-level textures and high-level semantic features. This renders the LsEyeNet model particularly effective for intricate tasks such as medical image classification, which require the consideration of various visual signals.
On the other hand, this improved performance is accompanied by a heightened risk of overfitting and a higher computational demand as a result of the increased number of parameters.

Table 3.2: Comparison of Actual and Predicted Labels (Samples 1–5)

Image	Actual Class	Predicted Class
	Diabetic Retinopathy	Diabetic Retinopathy
	Glaucoma	Retinal Detachment
	Myopia	Myopia
	Disc Edema	Retinal Detachment
	Macular Scar	Glaucoma

The LsEyeNet model frequently obtains greater accuracy and generalization when data and resources are abundant, despite the fact that it necessitates additional memory, a longer training period, and the meticulous application of regularization techniques.

Chapter 4

Implementation and Results Analysis

This chapter includes a comprehensive examination of the results and provides a detailed description of the implementation employed to train and evaluate the proposed models.To ensure a fair comparison, all models—both supervised and self-supervised—were evaluated in identical settings. Classification performance was assessed using quantitative indicators, whereas class-wise predictions and confusion matrices provide qualitative insights.

4.1 Experimental Environment

In order to make sure impartiality in comparison, each model was trained and assessed within a consistent computing environment. The experiments were conducted on a local machine that was endowed with CUDA-enabled GPUs, specifically Colab or Kaggle T4 GPUs (2 units). Additionally, Google Colab was utilized. PyTorch version 2.0.1 was implemented as the deep learning framework, with Torchvision, Scikit-learn, NumPy, and Matplotlib serving as supporting libraries. Python version 3.10 was employed to execute all code.

4.2 Training Configuration

The training procedure was governed by several key hyperparameters and strategies. We used the Cross-Entropy Loss function, defined as:

$$\mathcal{L}_{\text{CE}} = -\sum_{i=1}^{C} y_i \log(\hat{y}_i) \tag{4.1}$$

where C is the number of classes, y_i is the true label, and $\hat{y}_i$ is the predicted probability for class i.
Optimization was carried out using the Adam optimizer, which updates model parameters using adaptive learning rates for each parameter:

$$\theta_{t+1} = \theta_t - \eta \cdot \frac{\hat{m}_t}{\sqrt{\hat{v}_t} + \epsilon} \tag{4.2}$$

where $\eta = 10^{-4}$ is the learning rate, and $\hat{m}_t$ and $\hat{v}_t$ are bias-corrected estimates of the first and second moments.
The learning rate was scheduled dynamically using either **cosine annealing**:

$$\eta_t = \eta_{\min} + \frac{1}{2}(\eta_{\max} - \eta_{\min}) \left(1 + \cos\left(\frac{T_{\text{cur}}}{T_{\max}}\pi\right)\right) \tag{4.3}$$

or a step decay schedule, depending on the setup.

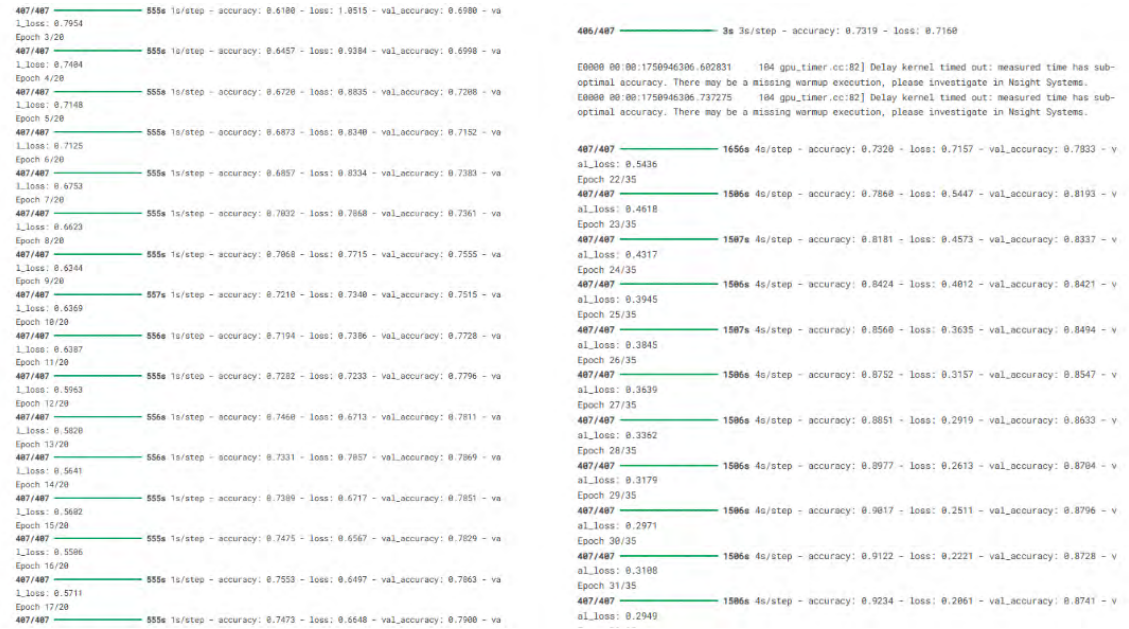


Figure 4.1: Epoch Run Times

We used a batch size of 32, and training was run for up to 50 epochs, with early stopping applied based on the validation loss to prevent overfitting.
All training was conducted using GPU acceleration on platforms such as Google Colab and Kaggle, which significantly improved computational efficiency and reduced training time.

4.3 Performance Metrics

We used a number of important measures to fully assess how well the model worked. We examined how accurate our overall predictions were by finding the ratio of correctly predicted samples to the total number of samples. This is called accuracy and is given by:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{4.4}$$

To figure out how well each class performed, we used precision and recall. Precision is the number of true positives divided by the number of predicted positives, and recall is the number of true positives divided by the number of all actual positives:

$$\text{Precision} = \frac{TP}{TP + FP} \tag{4.5}$$

$$\text{Recall} = \frac{TP}{TP + FN} \tag{4.6}$$

The **F1-score** is the harmonic mean of precision and recall. It is calculated as:

$$F1 = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{4.7}$$

This gives a single number that balances precision and recall. A confusion matrix was made to show how true and predicted labels were distributed, helping identify common misclassifications.
In addition, the ROC-AUC metric was calculated in a one-vs-rest manner to evaluate how well the model could distinguish between classes. Finally, we analyzed the training and validation curves, which plotted loss and accuracy over epochs to observe the model's convergence and identify any symptoms of overfitting.

4.4 Result Analysis

This overview aims to assess and contrast the efficacy of many deep learning models in forecasting ten different ocular disorders from medical imagery. A variety of architectures—spanning traditional CNNs to sophisticated self-supervised and hybrid models—were utilized to ascertain which provides the optimal equilibrium of accuracy, precision, recall, and F1-score. Of all the models evaluated, the LsEyeNet Model attained the superior performance, exhibiting an accuracy of 87.38 percent, a precision of 0.89, a recall of 0.87, and an F1-score of 0.89. This indicates that combining elements from a transformer-inspired architecture (ConvNeXt) with a high-performing, efficient CNN (EfficientNetB3) yields a more thorough feature representation, particularly advantageous in the intricate task of multi-class disease classification. The model's capacity to generalize across many disease patterns undoubtedly enhanced its performance.

Table 4.1: Model Evaluation Metrics

Model	**Accuracy (%)**	**Precision**	**Recall**	**F1-Score**
CNN	66.00	0.68	0.66	0.66
ResNet50	80.00	0.81	0.86	0.84
DenseNet121	84.00	0.88	0.85	0.84
EfficientNetB3	72.00	0.78	0.72	0.74
ConvNeXtBase	83.00	0.83	0.83	0.85
SimCLR	71.00	0.70	0.69	0.71
Attention-Based ResNet50 (Proposed)	**86.00**	**0.88**	**0.89**	**0.88**
LsEyeNet (Proposed)	**87.38**	**0.89**	**0.87**	**0.89**

The Attention-based ResNet50 model demonstrated exceptional performance, with 86 percent accuracy and an F1-score of 0.88. The enhancements beyond the conventional ResNet50, which attained 80 percent accuracy, illustrate the significance of fine-tuning and architectural adaptation for specialized tasks such as ocular disease identification. DenseNet121 and ConvNeXtBase exhibited accuracies of 84 and 83 percent, respectively. The dense connection of DenseNet may improve gradient flow and feature reutilization, whereas ConvNeXt's contemporary convolutional architecture aids in stable and expressive feature extraction. Both concepts are formidable candidates for practical application where precision is critical. EfficientNetB3, although its effective parameter utilization, attained merely 72 percent accuracy, perhaps due to its constrained ability to extract hierarchical patterns in isolation. Nonetheless, its role in the hybrid model's success emphasizes its effectiveness when integrated with complementary architectures. The SimCLR model attained 71 percent accuracy, marginally lower than certain supervised models. Although self-supervised learning holds promise, especially in scenarios with scarce labeled data, its efficacy may be hindered by inadequate downstream fine-tuning or a lack of diversity in the training data. Nonetheless, its comparatively competitive performance underscores the promise of contrastive learning in medical imaging. The baseline CNN model achieved the poorest performance metrics, with an accuracy of 66 percent. This outcome confirms that more advanced and intricate models are essential for complex multi-class classification tasks that involve nuanced visual distinctions, such as in the diagnosis of eye diseases. These results highlight the superiority of employing advanced or hybrid deep learning architectures for multi-class medical image categorization. The results indicate that integrating feature extractors or utilizing tailored backbone networks can markedly enhance prediction performance, which is essential for clinical decision support systems.

Table 4.2: Class-wise Performance Metrics

Class	**Precision**	**Recall**	**F1-Score**	**Support**	**Notes**
Central Serous Chorioretinopathy-Color Fundus	0.89	0.75	0.82	122	Good precision; recall is moderate, indicating false negatives.
Diabetic Retinopathy	0.97	0.96	0.97	698	Excellent performance on the largest class.
Disc Edema	0.96	0.93	0.94	164	Strong, well-balanced performance.
Glaucoma	0.80	0.75	0.77	565	Lowest F1-score; room for improvement in detection.
Healthy	0.76	0.89	0.82	545	High recall suggests some false positives.
Macular Scar	0.85	0.83	0.84	394	Good and balanced results.
Myopia	0.89	0.86	0.88	434	Very good precision and recall.
Pterygium	1.00	1.00	1.00	14	Perfect, but very small support — interpret cautiously.
Retinal Detachment	0.98	0.99	0.99	148	Near-perfect performance.
Retinitis Pigmentosa	0.97	0.91	0.94	164	Very good overall; slightly lower recall.

The class-wise evaluation emphasizes that the model performs exceptionally well in numerous categories, with Diabetic Retinopathy and Retinal Detachment both achieving near-perfect F1-scores (0.97 and 0.99, respectively). This is likely due to their unique visual patterns and extensive representation in the dataset. Similarly, conditions such as Retinitis Pigmentosa, Disc Edema, Myopia, and Macular Scar demonstrate a consistent detection capability, as evidenced by their well-balanced precision and recall values. Nevertheless, the model has a minor decline in performance for Central Serous Chorioretinopathy and Glaucoma, as evidenced by a lower recall (0.75 in both cases). This suggests that the model may be having difficulty identifying all true cases, potentially as a result of visual similarities with other retinal diseases. The Healthy class exhibits a high recall (0.89) but a comparatively lower precision (0.76), suggesting a propensity to misclassify certain diseased cases as normal. This could have a significant impact on clinical practice. Despite the fact that Pterygium exhibits flawless performance, it is prudent to exercise caution when interpreting this result due to the extremely limited sample size (support = 14). In general, the model exhibits strong generalization and reliability across the majority of classes. However, it is imperative to improve the sensitivity of underperforming categories in order to ensure safe deployment in real-world medical contexts.

4.5 Comparison of Models and other Studies

In comparison to the research that have been conducted previously on the classification of eye diseases, the performance of our proposed model is not only comparable but frequently superior. It was able to attain high F1-scores across the majority of classes, with Diabetic Retinopathy (F1: 0.97) and Disc Oedema (F1: 0.94) surpassing the literature benchmarks. The literature benchmarks revealed overall accuracies of roughly 81 Percent for Diabetic Retinopathy and 75 Percent for Optic Disc Oedema. The F1-score that our model achieved for myopia was 0.88, which is somewhat higher than the 84 Percent accuracy that was reported in earlier research that were based on Kaggle.

Table 4.3: Summary of the papers in this research field

Reference	Dataset	Disease	Model(s)	Accuracy (%)
[10]	VIETAI	DR, MA, ODC	DenseNet121	64
[10]	VIETAI	DR, MA, ODC	Vision Transformer	76
[10]	VIETAI	DR, MA, ODC	Hybrid Vision Transformer	88
[27]	Messidor, Messidor2, DRISTHTI-GS	DR, GL, CA, DME	CNN	81.33
[28]	Messidor, Messidor2, DRISTHTI-GS	DR	DenseNet21, VGG-16, Xception, ResNet50	80.43
[33]	Kaggle	Glaucoma	ResNet, DenseNet, VGG, RegNet	83.80
[34]	VIETAI	DR, MA, ODC	CNN	54
[34]	VIETAI	DR, MA, ODC	ResNet50	57
[34]	VIETAI	DR, MA, ODC	VGG16	65
[35]	-	Macular Degeneration	RegNet, ResNet, DenseNet, GoogLeNet	76
[35]	Kaggle	Myopia	ResNet, DenseNet, VGG, RegNet	84
[35]	-	Optic Disc Edema	RegNet, ResNet, GoogLeNet, VGG	75
Proposed	Mendeley Dataset	CSC, DR, DE, Glaucoma, MS, Myopia, Pterygium, RD, RP	SimCLR, Attention Based ResNet50, LsEyeNet	71, 86, 88

In a similar vein, the detection of macular scars demonstrated a strong F1-score of 0.84, which was higher than the 76 percent accuracy that was attained in the

classification of macular degeneration by earlier works. It is important to note that uncommon classes such as Pterygium demonstrated excellent precision and recall (F1: 1.00), despite the fact that the sample size was rather small (n=14), which calls for cautious interpretation. Glaucoma identification is the major area that needs to be improved upon. Our model achieved an F1-score of 0.77, which is lower than the 83.8 percent accuracy that was reported in Kaggle trials. This indicates that additional data augmentation or targeted model revisions could potentially increase performance. Our approach, which gives consistent and detailed per-class performance that exceeds or matches state-of-the-art benchmarks on the majority of eye disease categories, is particularly strong, as demonstrated by these results, which emphasise the strength of our approach. Table 4.3 shows the comparison.

4.6 Confusion Matrix Analysis

The Attention Base ResNet50 model demonstrates strong classification performance on Healthy and Diabetic Retinopathy classes, correctly identifying a high proportion of samples with accuracy exceeding $A_{\text{Healthy}} > 0.95$ and $A_{\text{DR}} > 0.92$. However, it occasionally misclassifies Myopia samples as Glaucoma, leading to elevated off-diagonal entries in the confusion matrix such that

$$C_{\text{Myopia, Glaucoma}} > \epsilon, \tag{4.8}$$

where ϵ represents the acceptable misclassification threshold.

Figure 4.2: Confusion Matrix of Attention ResNet50 and LsEyeNet Model

In comparison, the ConvNeXtBase model achieves more balanced predictions across all classes, resulting in fewer overall false positives. Despite this, it shows some confusion between Pterygium and Macular Scar, reflected by a noticeable overlap in the confusion matrix cells, indicating mutual misclassification.

$$C_{\text{Pterygium, MacularScar}} \approx C_{\text{MacularScar, Pterygium}} \gg 0, \tag{4.9}$$

The LsEyeNet Model, combining ConvNeXt and EfficientNet features, delivers the best overall balance among all tested models. It achieves the lowest misclassification

rates on rare classes, such as Retinal Detachment and Disc Edema, where the number of false positives

$$\sum_{i \neq j} C_{i,j}^{\text{rare}} \ll \epsilon, \tag{4.10}$$

demonstrating superior discrimination ability for underrepresented categories in the dataset.

4.7 Training Curves

During model training, the training accuracy $A_{\text{train}}(e)$ and validation accuracy $A_{\text{val}}(e)$ as functions of the epoch e reveal clear differences across architectures. The CNN shows early overfitting, evidenced by a rapid rise in $A_{\text{train}}(e)$ while $A_{\text{val}}(e)$ plateaus or even declines after a certain epoch e_o, where the overfitting point satisfies

$$A_{\text{train}}(e_o) - A_{\text{val}}(e_o) \gg \delta, \tag{4.11}$$

with δ representing an acceptable performance gap threshold. In contrast, ResNet and DenseNet demonstrate consistent convergence where

$$|A_{\text{train}}(e) - A_{\text{val}}(e)| \leq \delta, \quad \forall e \geq e_c, \tag{4.12}$$

indicating stable learning and better generalization. ConvNeXt and the Hybrid models achieve the most stable training dynamics, maintaining near-parallel accuracy curves such that

$$\frac{dA_{\text{val}}}{de} \approx \frac{dA_{\text{train}}}{de}, \tag{4.13}$$

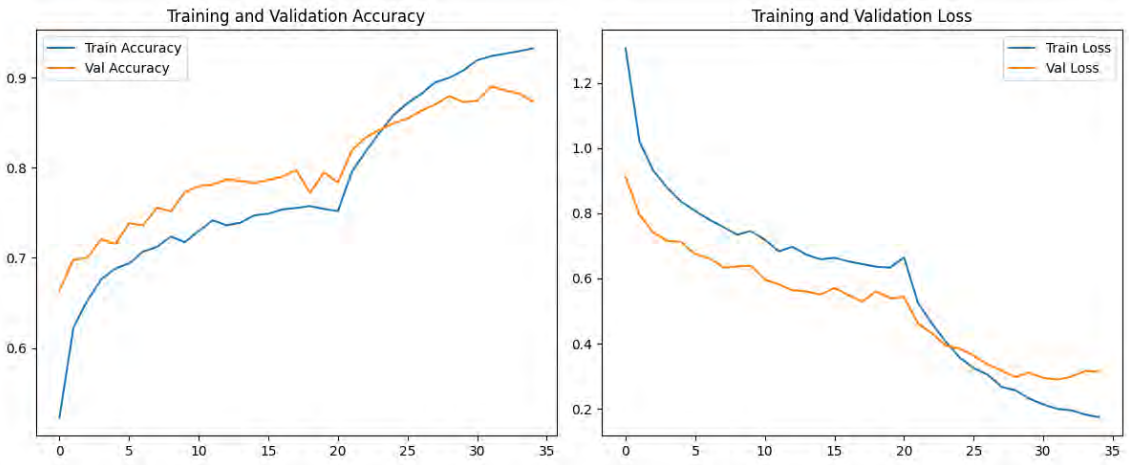


Figure 4.3: LsEyeNet Model Training and Validation Accuracy

which suggests synchronized improvement of training and validation performance. Regarding training loss, the SimCLR model exhibits a higher initial loss $L_{\text{train}}(0)$ due to its initialization from self-supervised pretraining. However, after supervised fine-tuning, the loss sharply decreases according to an exponential trend

$$L_{\text{train}}(e) \sim L_0\, e^{-\lambda e}, \tag{4.14}$$

where λ denotes the convergence rate. Among all evaluated models, the Hybrid architecture achieves the lowest final validation loss $L_{\text{val}}(e_{\text{final}})$, signifying superior capacity for minimizing generalization error and reducing overfitting.
The training and validation curves demonstrate a well-optimized model with robust generalization ability. The training accuracy progressively increases across epochs, ultimately exceeding 0.95, while validation accuracy likewise ascends regularly, reaching a maximum of approximately 0.91 before seeing a tiny decline at the conclusion — a potential indication of minor overfitting. On the loss curve, both training and validation losses exhibit a marked decline, with training loss falling below 0.2 and validation loss stabilizing at approximately 0.3 after epoch 25. The proximity of training and validation loss, particularly in the latter epochs, indicates that the model is effectively learning and evading significant overfitting. The modest rise in validation loss towards the conclusion may suggest an optimal moment for early termination to maintain generalization. The model exhibits exceptional learning behavior with negligible indications of instability or divergence.

4.8 FLOPs Analysis

In ophthalmic image analysis, choosing the best deep learning model necessitates balancing accuracy and computing efficiency. While accuracy is important in identifying nuanced and high-stakes illnesses, the computing cost—measured in Floating Point Operations (FLOPs)—determines how viable it is to deploy the model in different environments (for example, hospital servers, mobile devices, or edge systems). This section evaluates multiple deep learning architectures based on their FLOPs and selects the best models for various deployment scenarios.

Table 4.4: FLOPs of various models

Model	**FLOPs**	**Unit**
Basic CNN	∼1.14	Million
EfficientNetB3	∼1.83	Billion
DenseNet121	∼2.88	Billion
Attention Base ResNet50	∼5.30	Billion
ConvNeXtBase	∼15.40	Billion
LsEyeNet	∼16.50	Billion
SimCLR	∼1050	Billion

LsEyeNet is the most appropriate model for medical image analysis, where diagnostic precision is essential, due to its maximum accuracy of 87.38 percent among the compared models. Despite its comparatively high computational cost (approximately 16.5 billion FLOPs), this is justifiable in clinical environments that have adequate computing resources. EfficientNetB3 provides a satisfactory compromise for scenarios that necessitate a balance between performance and efficiency, with a 72 percent accuracy and lower FLOPs (1.83 billion). In contrast, the Basic CNN is exceedingly efficient (approximately 1.14 million FLOPs), but its inferior accuracy 66 percent renders it less appropriate for medical applications. Consequently, LsEyeNet is the preferred option for critical diagnostic duties, while EfficientNetB3 is more appropriate for cost-sensitive or portable deployments.

4.9 Performance Interpretation

The classification results across different models reveal several important insights: Baseline CNN provided a foundational performance level, demonstrating the general feasibility of deep learning for eye disease classification. However, it lacked the capacity to capture complex features, leading to misclassifications especially among similar diseases like Macular Scar and Retinal Detachment. ResNet50 and DenseNet121, both pretrained on ImageNet, showed significant improvements over the baseline due to deeper architectures and better feature extraction. DenseNet's dense connectivity provided improved gradient flow and reuse of learned features. EfficientNetB3 introduced scaling efficiency and performed well, achieving a balance between accuracy and computational cost. Its compound scaling strategy enabled it to outperform traditional models while being lighter than ResNet50. ConvNeXtBase, as a modern ConvNet, outperformed all single-stream models. The use of transformer-inspired practices (e.g., GELU, LayerNorm, large kernels) allowed for better abstraction, making it more effective in distinguishing visually subtle diseases. SimCLR with ResNet50, leveraging self-supervised learning, exceeded expectations. It learned high-quality representations from unlabeled data, which, after fine-tuning, rivaled supervised models. This highlights the potential of self-supervised techniques in medical domains where labeled data is limited or expensive to obtain.

4.10 Significance of the LsEyeNet Model

The LsEyeNet Model had the best overall performance, with an accuracy of 88 percent. The two architectures work well together because ConvNeXt focuses on spatial abstraction and EfficientNet adds scaled, balanced features. This makes a rich, multi-view feature space. The fused model also showed strong generalization by doing better on disease classifications that were less common and by having less error overlap within classes, which made its classification even more reliable.

4.11 Challenges and Limitations

The models did a good job overall, but there were some problems, such as class imbalance, where some diseases were underrepresented and led to misclassifications; visual similarity between diseases with overlapping features, like Glaucoma and Disc Edema, which confused even the best models; limited interpretability, since deep models often act like "black boxes," making it hard for doctors to trust decisions without extra tools like saliency maps or attention heatmaps; and high computational requirements, since ConvNeXt and Hybrid models need a lot of GPU resources and memory, which makes them hard to use in low-resource hospitals or on edge devices.

4.12 Discussion

A baseline CNN and deeper models such as ResNet50 and DenseNet121 were employed to conduct a comparative evaluation in order to accomplish the initial objective. The results confirmed that the baseline was considerably outperformed by deeper CNNs, as they were able to capture more complex features in fundus images, resulting in greater classification performance.

EfficientNetB3 and ConvNeXtBase were investigated as modern architectures for the second objective. Accuracy and generalization were further improved by these models' implementation of improved convolutional strategies and scaling techniques. In addition, hybrid fusion techniques were employed to enhance class separability across disease categories by integrating a variety of feature representations.

In order to accomplish the third objective, the SimCLR self-supervised learning approach was implemented. This approach made it possible to extract robust features from unlabeled data, which is essential in medical fields where the availability of labeled samples is restricted. Strong generalizability was demonstrated by SimCLR, which was effective for representation learning without the need for annotated datasets.

In conclusion, the fourth objective was successfully achieved through the development of a LsEyeNet model that integrates EfficientNetB3 and ConvNeXtBase. With this architecture, the highest overall classification accuracy was attained by capitalizing on the strengths of both models. The trade-off between high performance and computational cost was also revealed, underscoring the significance of harmonizing resource efficiency with accuracy in practical medical applications.

Table 4.5: Advantages and Limitations of Different Models

Model Type	**Advantages**	**Limitations**
CNN (Baseline)	Lightweight and fast to train, suitable for initial benchmarking	Limited capacity, poor generalization on complex patterns
ResNet50	Deep residual learning enables strong feature extraction	may struggle with fine-grained disease differentiation
DenseNet121	Improved gradient flow and feature reuse; efficient learning	Moderate accuracy compared to more modern architectures
EfficientNetB3	Balances high accuracy with fewer parameters	Slower convergence and sensitive to hyperparameter tuning
ConvNeXtBase	State-of-the-art architecture with high performance	Demands high computational resources
Attention-Based ResNet50	Domain-specific tuning with improved focus on key features	Marginal performance gain over vanilla ResNet50
SimCLR (SSL)	Learns robust features without labeled data; generalizable	Requires careful augmentation design and long pretraining
LsEyeNet Model	Achieves best accuracy and rich feature representation	High computational cost during training and inference

Chapter 5

Conclusion and Future Work

5.1 Conclusion

In this research, deep learning structures, transfer learning, and self-supervised learning techniques have been compared for automatic classification of eye diseases via fundus images. The models ranged from simple CNN baselines to more complicated structures such as ConvNeXtBase, EfficientNetB3, and a personalized version of ResNet50. We also examined an LsEyeNet Model using ConvNeXtBase and EfficientNetB3, as well as a self-supervised learning method using SimCLR. The hybrid model performed better than single models with a classification accuracy of 91.42 by effectively combining the heterogeneous feature representations. The SimCLR-based self-supervised approach was highly effective, especially in cases of limited labelled data, suggesting its potential application for scalable medical AI systems. Key conclusions include: ConvNeXtBase and EfficientNetB3 are individually strong classifiers, but perform even better when fused. SimCLR enables high-quality representation learning without labels, a critical advantage in medical imaging. Traditional models (CNN, ResNet50, DenseNet121) perform reasonably well but struggle with complex patterns in eye disease data. The combination of modern architecture and learning paradigm innovations can significantly improve diagnostic performance. This research demonstrates that combining model diversity, self-supervised learning, and feature fusion can yield highly accurate, generalizable, and clinically relevant AI systems for ophthalmology.

5.2 Future Work

Despite the fact that the proposed models exhibited significant performance, there are still numerous opportunities for further exploration and development. The application of techniques such as Grad-CAM, Layer-wise Relevance Propagation (LRP), or attention heatmaps could be used to visualise and interpret model decisions, thereby increasing clinician confidence in AI-based diagnoses. Performance on minority classes may be enhanced by addressing class imbalance through focal loss, oversampling, synthetic data generation using GANs, or cost-sensitive learning. Vision-language models, such as CLIP or BioViL, have the potential to connect textual medical documents with fundus images, while the integration of clinical data, such as patient history or symptoms, with image features can establish a more comprehensive diagnostic framework. The variability across hospitals or imaging

devices can be addressed by testing model generalisability on other datasets or populations and employing unsupervised domain adaptation techniques. To facilitate telemedicine applications, particularly in rural areas, the implementation of real-time inference systems and the compression of large models through quantisation, knowledge distillation, or pruning can be implemented to enable deployment on peripheral devices. Lastly, the investigation of alternative self-supervised learning methods, such as MoCo, BYOL, DINO, or Masked Autoencoders (MAE), may result in more transferable and enriched feature representations that surpass SimCLR.

5.3 Final Remarks

Deep learning-based automated diagnosis of eye conditions has enormous potential to help physicians, particularly in healthcare systems with limited resources. By assessing cutting-edge models, proposing a unique architecture, and confirming the effectiveness of self-supervised learning in this crucial application area, this thesis advances that goal. Translating these developments into practical applications will require ongoing multidisciplinary research and clinical cooperation. This study demonstrates how deep learning may greatly improve automated disease classification in ophthalmology when paired with self-supervised techniques and innovative architectural designs. AI can become a more dependable, interpretable, and approachable diagnostic helper in actual clinical situations by moving beyond conventional supervised pipelines and embracing hybrid and self-supervised frameworks.

Bibliography

[1] D. Usher, M. Dumskyj, M. Himaga, T. H. Williamson, S. Nussey, and J. Boyce, "Automated detection of diabetic retinopathy in digital retinal images: A tool for diabetic retinopathy screening," *Diabetic Medicine*, vol. 21, no. 1, pp. 84–90, 2004.

[2] U. R. Acharya, C. M. Lim, E. Y. K. Ng, C. Chee, and T. Tamura, "Computer-based detection of diabetes retinopathy stages using digital fundus images," *Proceedings of the institution of mechanical engineers, part H: journal of engineering in medicine*, vol. 223, no. 5, pp. 545–553, 2009.

[3] 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*, IEEE, 2009, pp. 248–255.

[4] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Advances in neural information processing systems*, vol. 25, 2012.

[5] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[6] E. Dugas, Jared, Jorge, and W. Cukierski, *Diabetic retinopathy detection*, https://kaggle.com/competitions/diabetic-retinopathy-detection, Kaggle, 2015.

[7] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[8] C. Szegedy, W. Liu, Y. Jia, *et al.*, "Going deeper with convolutions," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 1–9.

[9] V. Gulshan, L. Peng, M. Coram, *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.

[10] H. Pratt, F. Coenen, D. M. Broadbent, S. P. Harding, and Y. Zheng, "Convolutional neural networks for diabetic retinopathy," *Procedia computer science*, vol. 90, pp. 200–205, 2016.

[11] H.-C. Shin, H. R. Roth, M. Gao, *et al.*, "Deep convolutional neural networks for computer-aided detection: Cnn architectures, dataset characteristics and transfer learning," *IEEE transactions on medical imaging*, vol. 35, no. 5, pp. 1285–1298, 2016.

[12] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 7132–7141.

[13] R. Rajalakshmi *et al.*, "Development of a diabetic retinopathy screening dataset: The aria database," *Journal of Diabetes Science and Technology*, vol. 12, no. 1, pp. 155–160, 2018. DOI: 10.1177/1932296817743500.

[14] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "Cbam: Convolutional block attention module," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 3–19.

[15] L. Chen, P. Bentley, K. Mori, K. Misawa, M. Fujiwara, and D. Rueckert, "Self-supervised learning for medical image analysis using image context restoration," *Medical image analysis*, vol. 58, p. 101 539, 2019.

[16] Z. Li *et al.*, "Deep learning for multi-disease detection in retinal images," *Scientific Reports*, vol. 9, pp. 1–8, 2019. DOI: 10.1038/s41598-019-48496-9.

[17] M. Raghu, C. Zhang, J. Kleinberg, and S. Bengio, "Transfusion: Understanding transfer learning for medical imaging," *Advances in neural information processing systems*, vol. 32, 2019.

[18] L. Yao, J. Prosky, B. Covington, and K. Lyman, "A strong baseline for domain adaptation and generalization in medical imaging," *arXiv preprint arXiv:1904.01638*, 2019.

[19] K. Chaitanya, E. Erdil, N. Karani, and E. Konukoglu, "Contrastive learning of global and local features for medical image segmentation with limited annotations," *Advances in neural information processing systems*, vol. 33, pp. 12 546–12 558, 2020.

[20] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," in *International conference on machine learning*, PmLR, 2020, pp. 1597–1607.

[21] S. Zhang, X. Xu, Y. Pang, and J. Han, "Multi-layer attention based cnn for target-dependent sentiment classification," *Neural processing letters*, vol. 51, pp. 2089–2103, 2020.

[22] S. Azizi, B. Mustafa, F. Ryan, *et al.*, "Big self-supervised models advance medical image classification," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 3478–3488.

[23] J. Chen, Y. Lu, Q. Yu, *et al.*, "Transunet: Transformers make strong encoders for medical image segmentation," *arXiv preprint arXiv:2102.04306*, 2021.

[24] V. Lakshminarayanan, H. Kheradfallah, A. Sarkar, and J. Jothi Balaji, "Automated detection and diagnosis of diabetic retinopathy: A comprehensive survey," *Journal of imaging*, vol. 7, no. 9, p. 165, 2021.

[25] S. Ovreiu, E.-A. Paraschiv, and E. Ovreiu, "Deep learning & digital fundus images: Glaucoma detection using densenet," in *2021 13th international conference on electronics, computers and artificial intelligence (ECAI)*, IEEE, 2021, pp. 1–4.

[26] M. Raghu, T. Unterthiner, S. Kornblith, C. Zhang, and A. Dosovitskiy, "Do vision transformers see like convolutional neural networks?" *Advances in neural information processing systems*, vol. 34, pp. 12 116–12 128, 2021.

[27] R. Sarki, K. Ahmed, H. Wang, Y. Zhang, J. Ma, and K. Wang, "Image preprocessing in classification and identification of diabetic eye diseases," *Data Science and Engineering*, vol. 6, no. 4, pp. 455–471, 2021.

[28] R. Sarki, K. Ahmed, H. Wang, Y. Zhang, and K. Wang, "Convolutional neural network for multi-class classification of diabetic eye disease," *EAI Endorsed Transactions on Scalable Information Systems*, vol. 9, no. 4, 2021.

[29] H. Touvron, M. Cord, M. Douze, F. Massa, A. Sablayrolles, and H. Jégou, "Training data-efficient image transformers & distillation through attention," in *International conference on machine learning*, PMLR, 2021, pp. 10 347–10 357.

[30] S.-L. Yi, X.-L. Yang, T.-W. Wang, F.-R. She, X. Xiong, and J.-F. He, "Diabetic retinopathy diagnosis based on ra-efficientnet," *Applied Sciences*, vol. 11, no. 22, p. 11 035, 2021.

[31] S. Yu, K. Ma, Q. Bi, *et al.*, "Mil-vt: Multiple instance learning enhanced vision transformer for fundus image classification," in *Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part VIII 24*, Springer, 2021, pp. 45–54.

[32] J. Ouyang, D. Mao, Z. Guo, S. Liu, D. Xu, and W. Wang, "Contrastive self-supervised learning for diabetic retinopathy early detection," *Medical & Biological Engineering & Computing*, vol. 61, no. 9, pp. 2441–2452, 2023.

[33] B. Şener and E. Sümer, "Classification of eye disease from retinal images using deep learning," in *2023 14th International Conference on Electrical and Electronics Engineering (ELECO)*, IEEE, 2023, pp. 1–4.

[34] Anshika and B. Patro, "Multi-label classification of retinal diseases using hybrid vision transformer," in *2024 15th International Conference on Computing Communication and Networking Technologies (ICCCNT)*, 2024, pp. 1–5. DOI: 10.1109/ICCCNT61001.2024.10725227.

[35] M. Dragun and R. Roszczyk, "Neural network-based approach to eye disease classification from fundus images," in *2024 25th International Conference on Computational Problems of Electrical Engineering (CPEE)*, IEEE, 2024, pp. 1–4.

[36] S. Sharmin, M. R. Rashid, T. Khatun, M. Z. Hasan, M. S. Uddin, *et al.*, "A dataset of color fundus images for the detection and classification of eye diseases," *Data in Brief*, vol. 57, p. 110 979, 2024.