CAPITAL UNIVERSITY OF SCIENCE AND
TECHNOLOGY, ISLAMABAD

CALF-Net: A CNN Attention-Leveraged Transformer Network for Diabetic Retinopathy Detection

by

Hafsah Mahmood

A thesis submitted in partial fulfillment for the
degree of Master of Science

in the

Faculty of Computing
Department of Computer Science

2026

CERTIFICATE OF APPROVAL

CALF-Net: A CNN Attention-Leveraged Transformer Network for Diabetic Retinopathy Detection

by

Hafsah Mahmood

(MCS243001)

THESIS EXAMINING COMMITTEE

S. No.	Examiner	Name	Organization
(a)	External Examiner	Dr. Asif Muhammad	FAST, Islamabad
(b)	Internal Examiner	Dr. Aamer Nadeem	CUST, Islamabad

Dr. Nadeem Anjum
Thesis Supervisor
May, 2026

Dr. M. Masroor Ahmed
Head
Dept. of Computer Science
May, 2026

Dr. M. Abdul Qadir
Dean
Faculty of Computing
May, 2026

Author's Declaration

I, **Hafsah Mahmood** hereby state that my MS thesis titled "**CALF-Net: A CNN Attention-Leveraged Transformer Network for Diabetic Retinopathy Detection**" is my own work and has not been submitted previously by me for taking any degree from Capital University of Science and Technology, Islamabad or anywhere else in the country/abroad.

At any time if my statement is found to be incorrect even after my graduation, the University has the right to withdraw my MS Degree.

(Hafsah Mahmood)

Registration No: MCS243001

Plagiarism Undertaking

I solemnly declare that research work presented in this thesis titled "**CALF-Net: A CNN Attention-Leveraged Transformer Network for Diabetic Retinopathy Detection**" is solely my research work with no significant contribution from any other person. Small contribution/help wherever taken has been duly acknowledged and that complete thesis has been written by me.

I understand the zero tolerance policy of the HEC and Capital University of Science and Technology towards plagiarism. Therefore, I as an author of the above titled thesis declare that no portion of my thesis has been plagiarized and any material used as reference is properly referred/cited.

I undertake that if I am found guilty of any formal plagiarism in the above titled thesis even after award of MS Degree, the University reserves the right to withdraw/revoke my MS degree and that HEC and the University have the right to publish my name on the HEC/University website on which names of students are placed who submitted plagiarized work.

(Hafsah Mahmood)

Registration No: MCS243001

Acknowledgement

First and Foremost, I am grateful to Almighty Allah for giving me the courage, endurance, and direction I needed to finish this thesis. This voyage was made possible by his countless blessings. Also, I would like to dedicate this thesis to my dear husband, Talha Khan, who has been so supportive and encouraging in my higher education journey. Your faith in me, your support that has always given me a push, and being there with me when I needed you the most is priceless. Thanks to your support, your encouragement, and your help in making this a possibility. I also dedicate this work to my family whose support, patience and encouragement have enabled me to persue my dreams. Your love and support have been my foundation, and I am deeply grateful for everything you have done to help me succeed.

My supervisor, Dr. Nadeem Anjum, has my sincere gratitude for his unwavering support, kind words of encouragement, and insightful advice during this endeavour. It meant more to me than words can say that you trusted me. I want to express my sincere gratitude to my family and friends for their unwavering love, support, and prayers, which helped me get through challenging times.

(Hafsah Mahmood)

Abstract

Diabetic Retinopathy (DR) is a serious condition that occurs in type two diabetic patients and can lead to irreversible vision loss if not treated in time. The visual indicators for automated detection of DR are minute lesions, which are categorized as microaneurysms, hemorrhages, and exudates. However, the detection remains challenging due to the similarities of DR lesions with other retinal pathologies, leading to low sensitivity (recall/ increased false negatives) despite achieving high accuracy and specificity. This study presents an integrated model, CALF-Net, for improved DR detection in terms of sensitivity in the presence of other retinal pathologies by combining Convolutional Neural Networks (CNNs) with Vision Transformers (ViTs) in order to effectively capture the local lesion features with global contextual information.

The proposed approach consists of careful, extensive preprocessing steps for enhancing retinal images. The features of the fundus images are extracted through a selected CNN backbone, and global spatial dependencies are learned from transformer integration, along with a spatial attention gate to focus only on lesions and reduce retinal background noise. The proposed CALF-Net is evaluated using a publicly available dataset named RFMiD, chosen due to the presence of multiple retinal diseases and to mirror the actual clinical settings by effectively detecting DR in the presence of other similar ocular pathologies. The model was evaluated against accuracy, sensitivity, specificity, and Area under the ROC curve (AUC-ROC), while image quality assessment was done through Peak Signal to Noise Ratio (PSNR) and Structural Similarity Index (SSIM).

According to experimental results, CALF-Net outperforms current methods on the RFMiD dataset by achieving **91.88%** accuracy, **96.77%** sensitivity, and **90.89%** specificity. The results indicate a clinically favorable trade-off where the model prioritizes sensitivity over specificity. This is important as increased sensitivity(recall) reduces the false negatives and lowers the chance of missed diagnosis, which can lead to irreversible vision loss. The key contribution of the proposed methodology lies in improving the sensitivity and reducing the false negatives by

effectively capturing the local features with their global context in the presence of other visually similar retinal pathologies. Additionally, the suggested model attains an AUC of **0.9750**, demonstrating a high capacity for discrimination. A PSNR of **31.26 dB** and an SSIM of **0.88** are obtained from the preprocessing pipeline's image quality evaluation, showing the successful enhancement of retinal features without degradation that are essential for accurate lesion detection.

Overall, this thesis shows that automated DR detection is greatly improved by combining global context with local feature learning. With great potential for practical implementation in computer-aided ophthalmic screening systems, the suggested CALF-Net architecture offers a scalable and clinically useful solution.

Contents

List of Figures

List of Tables

Abbreviations

AMD/ARMD	Age related Macular Edema
AGC	Adaptive Gamma Correction
ADHE	Adaptive Histogram Equalization
AdamW	Adam with Decoupled Weighted Decay
APTOS2019	Asia Pacific Tele-Ophthalmology Society 2019
BRVO	Branch Retinal Vein Occulusion
CALF-Net	CNN Attention-Leveraged Transformer
CALHE	Contrast Adaptive Limited Histogram Equalization
CONV	Convolutional
CRVO	Central Retinal Vein Occulusion
CTNet	Convolutional-Transformer Network
DDoG	Dilated Difference of Gaussian
DR	Diabetic Retinopathy
DRD	Diabetic Retinopathy Dataset
DRNet	Diabetic Retinopathy Network
DnCNN	Denoising Convolutional Neural Network
DoG	Difference of Gaussian
DTNet	DenseNet and Transformer Network
ESIHE	Exposure based Sub-Image Histogram Equalization
EyePACS	Eye Picture Archive Communication System
FC	Fully Connected
FEB	Feature Extraction Block
FFN	Feed Forward Network
GAN	Generative Adversarial Network

GAP	Global Average Pooling
GPB	Grading Prediction Block
Grad-CAM	Gradient-weighted Class Activation Mapping
HE	Histogram Equalization
HIRD-Net	Hierarchical-Inception Residual-Dense Network
HSI	Hue, Saturation, Intensity
HSV	Hue, Saturation, Value
IoMT	Internet of Medical Things
IDRiD	Indian Diabetic Retinopathy Image Dataset
MHSA	Multi-Head Self-Attention
MLP	Multi-Layer Perceptron
NiN	Network in Network
PSNR	Peak Signal to Noise Ratio
RFMiD	Retinal Fundus Multi-Disease Image Dataset
RGB	Red, Green and Blue
ReLU	Rectified Linear Unit
ResNet	Residual Network
ResVit	ResNet with Vision Transformers
SECA	Squeez and Excitation Attention
SOP	Standard Operating Procedure
SSIM	Structural Similarity Index
SVM	Support Vector Machine
VGGNet	Visual Geometry Group Network
ViTs	Vision Transformers

Chapter 1

Introduction

1.1 Background

Vision is one of the fundamental senses of human beings, which acts as a source to navigate from one place to another. It is a medium for interacting with the environment through perceiving light and shape [1]. It helps a person to perform their daily routine activities and enhance the quality of their life. The functioning of the human eye is a complex mechanism of a biological imaging system in which light signals are converted to neural signals with the help of the retina [2]. The retina is a sensory layer that is located at the back of the human eyeball. It is sensitive to light, activating nerve impulses that are transmitted to the brain through the optic disc, with the help of which a visual image is created. Due to its unique properties, the retina is examined to diagnose vision-related ocular diseases [3]. Among these, Diabetic Retinopathy is a very severe retinal disease that is one of the causes of worldwide vision impairment.

Diabetic Retinopathy (DR) is found in type two diabetic patients due to prolonged high blood sugar levels [4]. It affects the blood vessels of the retina by damaging them and creating small lesions categorized as microaneurysms, hemorrhages, and exudates. If it is not detected in time, it keeps on spreading in the eye, which can lead to irreversible vision loss [5]. Tan and Wong [6] reported in 2024 that 589 million people belonging to the age group of 20 to 79 years are living with diabetes,

and this number is going to increase up to 853 million by 2050, indicating a total increase by 45%. It is estimated that 22% to 35% of diabetic patients will develop DR in some stage of their lifetime [7]. Figure 1.1 and 1.2 represent the retina of a healthy patient and the one suffering from DR.

FIGURE 1.1: Fundus image of a Healthy Eye [8]

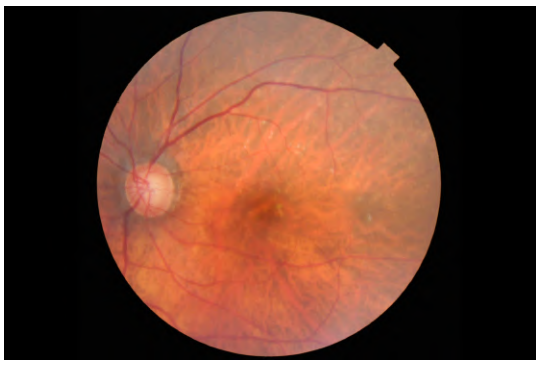

FIGURE 1.2: Fundus image containing Diabetic Retinopathy [8]

1.2 Visual Indicators of Diabetic Retinopathy in Fundus Images

The small retinal lesions are formed during DR, which serve as indicators of the disease while detecting it. The lesions that are associated with DR, in particular, are

i. Microaneurysms: These are formed in the capillaries of the retina. These capillaries are dilated, and sometimes tiny bulges appear on them. Visually, they appear like small red spots in the fundus images as shown in Figure 1.3. Although they are small in size, they play a vital role in the initial progression of DR [9].

ii. Hemorrhages: These occur particularly when the blood vessels get ruptured, and the blood leaks out of these vessels.

    The size of these lesions depends on the leakage. The occurrence of hemorrhages depicts the worsening of the condition. Figure 1.4 shows the Hemorrhages in DR[10].

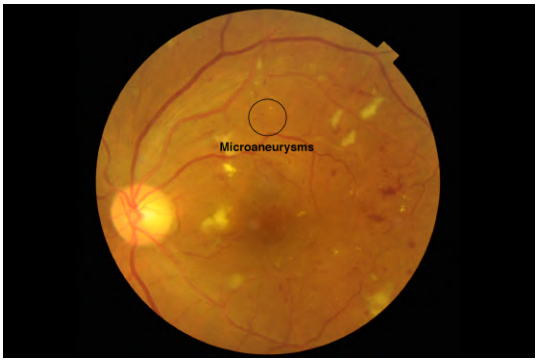

FIGURE 1.3: Microaneurysms in Diabetic Retinopathy [8]

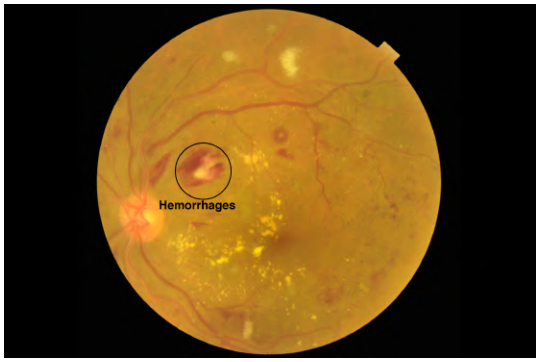

FIGURE 1.4: Hemorrhages in Diabetic Retinopathy [8]

iii. Exudates: The above-mentioned abnormal vessels lead to the deposit of protein or lipids, appearing in a yellow color. The presence of exudates refers to the advanced stages of DR. Figure 1.5 and 1.6 show two types of exudates, named as soft and hard, respectively[11].

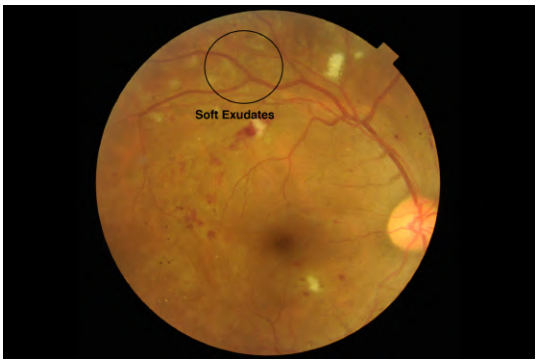

FIGURE 1.5: Soft Exudates in Diabetic Retinopathy [8]

FIGURE 1.6: Hard Exudates in Diabetic Retinopathy [8]

Altogether, these pathological symptoms form lesions in DR whose size, appearance, color, and arrangement on the retina depict the severity of the disease.

1.3 Disease Progression and Spatial Dependency

The appearance of these lesions in DR is not developed in an isolated pattern; rather there spatial location on the retina also explains the presence of the disease. For instance, a single red spot appearing on the retina does not mean it is a microaneurysm; it could be normal vasculature, or this could appear due to camera

artifacts as noise [12]. Hence, there is a need to look at the global distribution, as in the early stages, these tiny red bulges are formed in scattered positions, usually near the macula and sometimes in a clustered pattern [13]. With the progression of the disease, the hemorrhages and the exudates also appear on the retina, leading to different severity levels of DR. The accurate detection of it required both the local and the global location of these indicators in order to reduce the false negatives. Hence, the detection of these lesions on the pixel level is not sufficient, and it requires information about the distribution and clustering of lesions with respect to the retinal anatomy [14].

1.4 Challenges in Fundus Image Analysis

Fundus images act as a non-invasive method of analyzing the anatomical pattern present in the retina. However, there are a number of factors that make the analysis challenging. These include lightening and illumination distortions, low contrast imaging due to various camera devices, and patient-specific differences [15]. Even though the pupil of the patient is dilated with the help of eye drops before image acquisition, which allows more light to enter the eye, enabling the camera to capture a wider and clearer view of the retina [16].

Also, in some cases, the coloring agent is injected through the patient's arm in order to increase the visibility of blood vessels [17]. Still, the final images are unclear and can sometimes miss the minute details, including small pathological indicators or thin retinal blood vessels that are required for the diagnosis [18]. Among them are lesions, specifically microaneurysms and exudates, which are difficult to view because of their small size and appearance similar to the natural anatomy of the retina. The identification becomes even harder due to the presence of noise, poor contrast, and lens artifacts [19].

Therefore, the preprocessing of the fundus images is a vital step before the automation of the diagnosis, which includes contrast enhancement, denoising filters, and steps to make small anatomical structures appear wider so that they can be distinguishable as well as to enhance the overall model performance [20].

1.5 Automated DR Detection using Convolutional Neural Networks and Vision Transformers

Traditionally, DR is mostly detected manually with the help of trained ophthalmologists who routinely analyze retinal fundus images. They manually look at the signs of the DR, which can sometimes lead to erroneous results, specifically when large-scale screenings are conducted. This is even amplified in the area where resources are limited, along with a shortage of experienced clinicians [21].

In order to overcome these problems, researchers are working on automated detection of DR with the help of computer-aided systems. Deep Learning Convolutional Neural Networks have emerged as a tool for automatic extraction of the local spatial features from images, increasing the success rate of detecting various diseases in medical imaging [3]. CNNs use the hierarchical representation of features for learning the patterns, including lesions to detect DR. Moreover, Vision Transformers have also emerged due to their ability to capture spatial dependencies of the lesions in the retina, which helps in detecting DR more accurately [17].

1.5.1 Working of Convolutional Neural Networks and Vision Transformers

The deep learning models consist of Convolutional Neural Networks that are particularly designed for the data, which can be represented in the form of grids or matrices known as images. CNNs are especially employed in medical imaging due to their ability to automatically extract the local feature patterns. The pixels that are confined to a small number of surrounding pixels are termed local visual patterns. The word local is used for them because their identification can be done with the help of a limited region [22].

The building blocks of CNNs are Convolutional layers, which are built up of kernels. The basic architecture of the CNNs is shown in Figure 1.7. The kernels

consist of learnable filters. Let us suppose there is an image with Height and Width represented with H and W. The Channels of the image are denoted by C. For example, in the case of an RGB (Red, Green, Blue) image, the number of channels will be equal to three. The following equation represents a simple image in RGB colors.

$$I \in \mathbb{R}^{H \times W \times C} \tag{1.1}$$

A set of learnable filters is applied with the help of a Convolutional layer represented as

$$K \in \mathbb{R}^{f \times f \times C} \tag{1.2}$$

where K is the kernel and f is the size of the filter.

The fxf kernel window slides on the whole image step by step. The weighted sum of the image pixels with respect to the kernel is calculated by taking the dot product of overlapping pixels. The result after sliding the kernel on the image will produce a feature map indicating one of the features learned from the image [20].

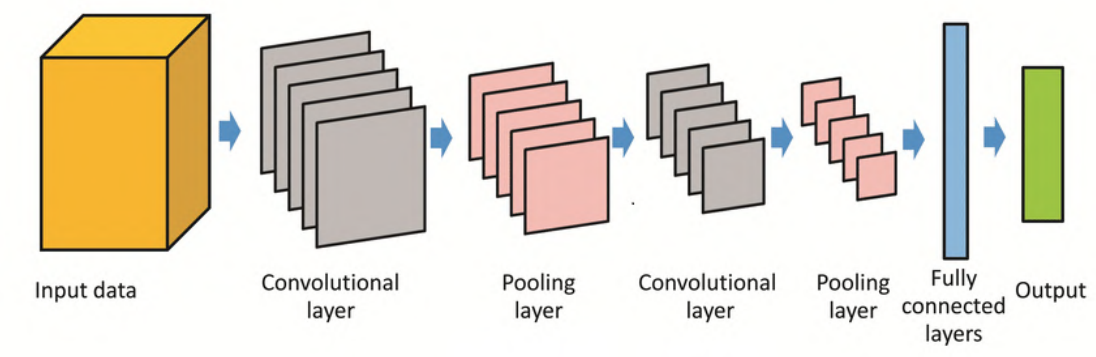

FIGURE 1.7: Basic Architecture of Convolutional Neural Networks [23]

In the initial stages of the Convolutional layer, the edges and boundaries are learned by the CNNs, and progressively, by the end of reaching the last Convolutional layer, the model learns all the features like lesions, blood vessels, optic disc, exudates, and many more.

However, it is seen that the primary focus of the CNNs is on the receptive fields present locally on the retina, meaning they can detect the features on the pixel level [20].

This is useful for identifying individual lesions, but their dependencies on the other anatomical representations are overlooked by them, resulting in misclassifications, especially in advanced stages of the DR.

For this purpose, transformers that were created for natural language processing are utilized successfully for image analysis as well. They are termed as Vision Transformers (ViTs). The concept is the same as that used in traditional transformers by extracting the information from small discrete units, termed as tokens. In natural language processing, tokens are the words of the sentences. Similarly, in images, tokens are the small patches that are extracted from them. After the formation of these, each patch is flattened into 1D vectors. Additionally, the positional information of the pixels is attached to each patch and is termed as positional embeddings in order to maintain the spatial record associated with patches, which is required for understanding the global context of an image [24].

Transformers understand the context with the help of Multi-Head Self-Attention (MHSA). Weights are assigned to the patches based on the global relevance regardless of the spatial distance that exists among them. The weights define which portion of the image should be given extra importance while training [24]. The basic architecture of transformers is shown in Figure 1.8.

Several studies have explored the strengths of vision transformers and have reported an increase in performance when it comes to understanding long-range global dependencies. This type of architecture improves the detection rate of diabetic retinopathy, especially in detecting the severe cases of diabetic retinopathy, which require information about the placement of the lesion within the anatomical structure of the human retina [25].

However, Transformers depend on a huge amount of data for their training, testing, and in order to generalize the context. This results in their limited application in the field of medical imaging.

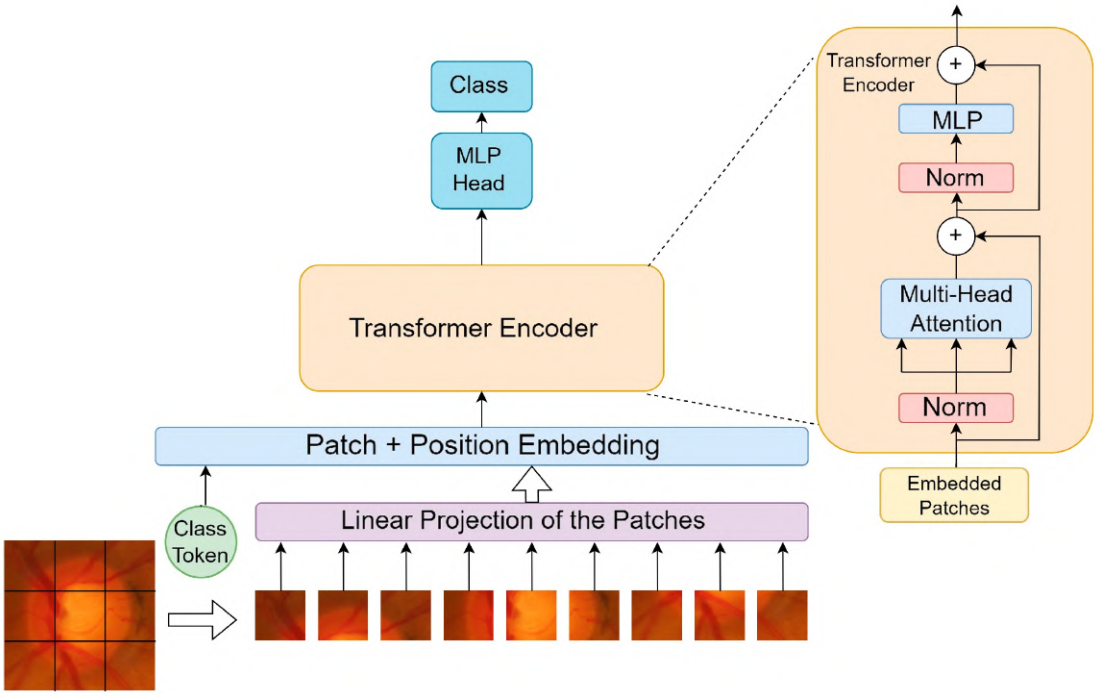

FIGURE 1.8: Basic Architecture of Transformers [26]

1.5.2 Hybrid Approach

Recent studies have explored numerous ways to successfully integrate both the strengths of CNNs and Transformers by amalgamating the local features extraction and global reasoning, respectively [18]. These models are suitable for the detection of DR due to several reasons that are listed below.

i. These models are able to detect small lesions due to the inherent nature of CNNs.

ii. These models are able to understand the global relationship of these lesions with the other anatomical structures of the retina.

iii. They improve the detection of the DR, specifically in the existence of other pathologies as well.

1.6 Clinical Perspective on Diagnostic Errors

Accuracy is not the sole metric when it comes to the performance evaluation of a medical diagnosis of an automated screening system. It is decided on the basis of numerous types of classification outcomes that impact clinical decision-making. These outcomes are accuracy, sensitivity (recall), specificity, and the area under the curve (AUC).

Special attention is given to the false negatives in medical applications [20]. In the case of DR [10], it means a person with DR is misclassified as a healthy patient, which leads to a delay in treatment, resulting in progression of the disease, leading to irreversible vision impairment. However, compared to this, a false positive may require additional examination by a clinician, leading to a lower level of risk when compared with false negatives.

The false negatives are quantified as Recall, also known as Sensitivity, in medical terms. It indicates the correct identification of the disease cases by the system. High sensitivity means that the number of false negatives is less, which is most important in the automated medical screening of diseases [6]. It makes the system more cautious about the disease cases and tries not to overlook them.

To summarize, DR is a severe disease of the retina leading to vision loss. The disease is spread through the lesions termed as microaneurysms, hemorrhages, and exudates. The key to detecting DR is the extraction of these lesions, which are used as the main diagnostic markers.An effective representation allows the model to differentiate DR from other retinal diseases with visual similarities. Consequently, improves classification accuracy, leading to a higher sensitivity and lower false negatives, as the model can better distinguish between subtle disease pathologies in similar retinal conditions.

Therefore, an automated approach is required to effectively extract the lesion features while understanding not only the local retinal features but also the global context of them in order to increase the sensitivity of DR detection. The goal

is to minimize the false negatives in a setting where multiple clinical pathologies coexist with each other.

1.7 Problem Statement

The detection of DR remains challenging in the presence of other retinal pathologies due to visually similar lesion patterns accross different diseases, leading to an increase in false negatives, resulting in lower sensitivity.

1.8 Research Objectives

i. To improve the DR lesion detectability by applying a lesion-aware preprocessing technique in Retinal fundus images.

ii. To propose a framework that can improve lesion representations to reduce the false negatives noted in the classification of DR, particularly in a multi-disease dataset.

1.9 Research Questions

RQ1: How does lesion-aware preprocessing enhance the structural quality and visibility of diabetic retinopathy lesions in fundus images with different levels of noise and illumination?

RQ2: How much can the sensitivity be improved and false negatives be reduced by improving lesion detection in Diabetic Retinopathy?

1.10 Research Contributions

The following are the main contributions of this study:

i. Lesion-Aware Preprocessing: To enhance the visibility and recognition of DR lesions, by applying lesion-aware image preprocessing techniques.

ii. Enhanced Sensitivity for Diabetic Retinopathy diagnosis: To reduce false negatives and increase sensitivity by improving DR-specific lesion extraction and enhancing DR diagnosis in the presence of visually similar retinal diseases.

1.11 Organization of the Thesis

The rest of the thesis is organized as follows. Chapter 2 is on Literature Review, focusing in detail on the studies about preprocessing of the fundus images for improving the detection of DR. It also covers the papers focusing on classification of DR through CNNs, ViTs, and existing hybrid approaches. Chapter 3 covers the Methodology of the thesis, including the dataset description and implementation details, followed by Chapter 4, Results and Discussions, stating the performance of the proposed approach compared to the existing approaches. It also covers the detailed ablation study conducted to verify the methodology even further. Finally, Chapter 5 is Conclusion and Future Work, which summarizes the contributions and points to the limitations along with the future direction for the proposed work.

Chapter 2

Literature Review

2.1 Introduction

DR is a serious retinal condition that, if detected early, can prevent vision loss worldwide. Traditional detection of DR was performed by trained ophthalmologists, resulting in prolonged screening time and erroneous results, particularly in sites with limited access to specialists, especially rural and underdeveloped areas [19].

Recently, a gradual inclination has been seen towards automated screening in order to meet operational efficiency, along with enhancing accuracy and patient safety in a shorter amount of time. Automated systems are capable of handling a large amount of data compared to traditional manual screening [7].

In this regard, Convolutional Neural Networks (CNNs) have depicted improved performance due to their inherent nature to extract the local features pixel-wise automatically with the help of their hierarchical structure. On the other hand, Vision Transformers (ViT) have also shown promising results because they tend to understand the long spatial global relationship among numerous anatomical structures. Additionally, hybrid approaches have outperformed the above-mentioned by utilizing the strengths of both techniques. The purpose of the literature review

is to provide a detailed analysis of the existing techniques that are employed for DR detection.

The main focus is on the research that has used deep learning approaches, including CNNs and ViTs, along with their integration and the augmentation techniques presented in them. The chapter also focuses on the preprocessing techniques that are important while working with fundus images, and the data augmentation techniques to obtain bias-free results.

The review is conducted on the research papers that are carefully selected after applying the inclusion and exclusion criteria. This ensures that only the most relevant and impactful studies are selected, highlighting the need for innovative and influential contributions in the DR detection.

The rest of the chapter is organized in the following way. The selection criteria of the studies are discussed in detail, followed by the detection of DR through deep learning approaches, mainly focusing on three categories, including CNNs, ViTs, and Hybrid approaches. This is followed by a section focusing on the preprocessing of fundus images specifically for DR detection. The chapter is concluded by stating the research gap, followed by the specific problem this research is addressing, along with the research objectives, consequently leading us to the research question that this research will answer, motivating the need for the proposed methodology.

2.2 Study Selection: Inclusion and Exclusion Criteria

A total of 70 research papers were selected for the study. The following inclusion criteria were applied to the selected papers as shown in figure 2.1.

2.2.1 Inclusion Criteria

The inclusion criteria for the literature review is isted below.

i. Studies published in the years 2021 to 2026.

ii. Studies proposing DR detection using Convolutional Neural Networks.

iii. Studies proposing DR detection using Vision transformers.

iv. Studies proposing hybrid approaches combining CNNs and ViTs for better DR detection.

v. Studies explaining preprocessing techniques.

After applying the inclusion criteria, the number of papers was narrowed down to 60.

2.2.2 Exclusion Criteria

The exclusion criteria cover the following points.

i. Studies that do not suggest end-to-end pipelines for DR detection.

ii. Studies that do not cover the preprocessing techniques that are specific to fundus images. Those papers in which the preprocessing techniques are not tested through classification are also excluded.

The papers left after applying the exclusion criteria were 45-50 in number. These were the most relevant studies for this research.

2.3 Review of DR Detection Techniques

2.3.1 CNN-based Approaches

Abbasi et al. [1] proposed an extensive preprocessing technique to handle cataract-type images and to detect DR in them. After applying the preprocessing on the images, they were classified as DR vs Non-DR by using 6 known Convolutional

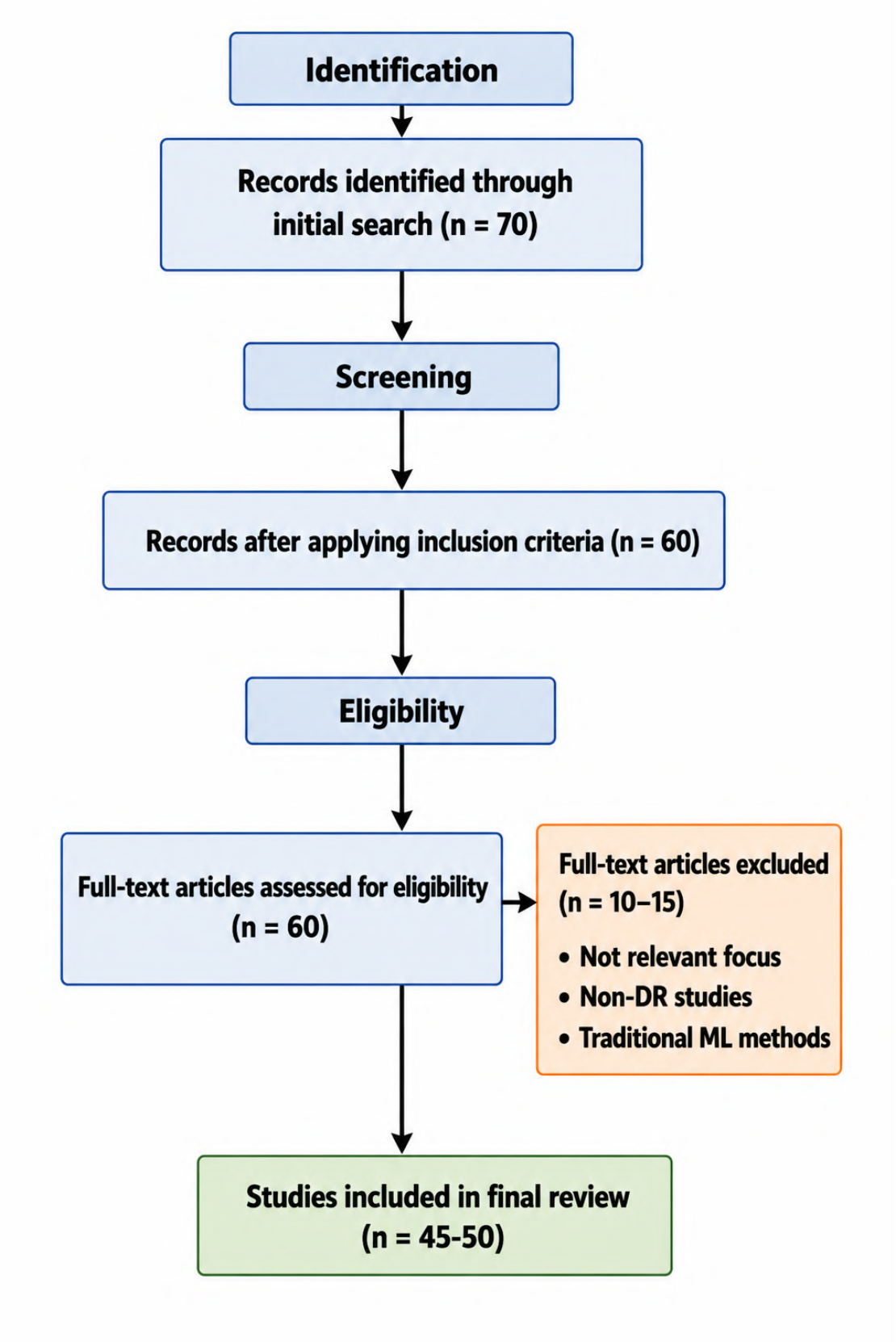

FIGURE 2.1: Inclusion and Exclusion Criteria of Literature Review

Neural Network models, including ResNet-152, AlexNet, GoogleNet, VggNet-s, VggNet-16, and VggNet-19. Three datasets were used for this purpose, including Messidor, RFMiD, and EyePACS. The outperformer among them was ResNet-152 with 91.56% accuracy. However, the sensitivity was reported to be 85.48%. The architecture of ResNet-152 is explained in the Figure 2.2 below.

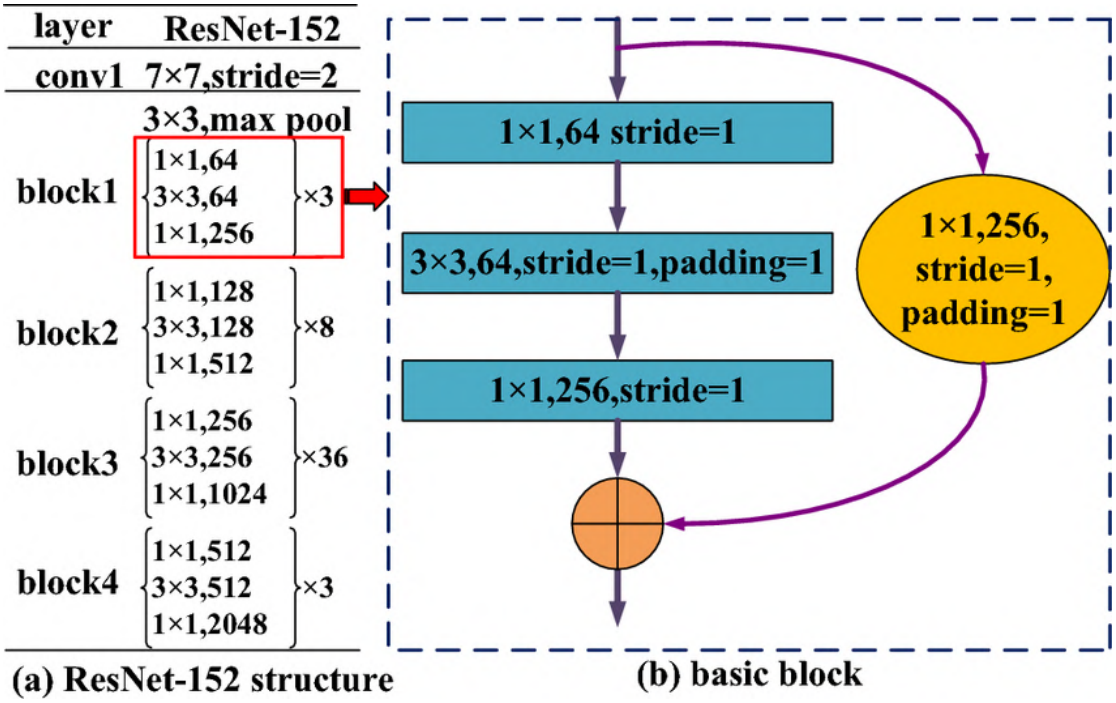

FIGURE 2.2: Architecture of ResNet-152 [27]

It is seen that although the accuracy is achieved, the sensitivity is not high which means that many false negatives are present. This might be due to the reason that the model is only capturing the localized feature, and hence, they are unable to detect the disease severity.

Also, the paper results were better on Messidor and EyePACS datasets, but the technique struggled when it came to the RFMiD dataset, which is a multi-disease dataset, and in the presence of other clinical pathologies that are quite visually similar to DR, the detection of DR becomes even harder.

Kumar et al. [2] also conducted research on RFMiD for the classification of DR using four convolutional neural network models. A total of 15 epochs for training

were used with 60 iterations each. The models were VGG-19, VGG-16, EfficientnetB0, and ResNet152, with the highest accuracy of 90.00% and sensitivity of 87.14% recorded on ResNet-152.

Ashraf et al. [3] worked on the classification of the DR by introducing a novel end-to-end approach named Hierarchical-Inception-Residual-Dense Network (HIRD-Net).

The paper utilizes Contrast Limited Adaptive Histogram Equalization (CLAHE) and Dilated Difference of Gaussian (DoG) to enhance the lesions of the fundus images. The proposed model not only contains the hierarchical feature representation but also contains the side-by-side residual dense blocks for an enhanced mechanism of feature learning.

Along with this, the model is implementing Squeeze and Excitation Channel Attention (SECA) in order to improve the Feature Maps.

SECA is applied after each global average pooling (GAP) layer. The proposed framework was validated on IDRiD-APTOS2019, DDR, and EyePACS datasets. The highest accuracy achieved was 93.46% on IDRiD-APTOS2019, with a sensitivity of 87.02%.

The paper also added the element of explainability by adding the Grad-CAM at the end as shown in figure 2.3. It is seen that the model is tested only on the benchmark datasets that only contain DR images vs healthy images. There is a need to test the proposed methodology on more challenging dataset which are containing other clinical pathologies as well to recreate the actual clinical settings.

Suedumrong et al [4] were able to detect DR in the EyePACS dataset with a classification accuracy of 90.60%, and sensitivity of 87.64% by applying numerous preprocessing techniques, including conversion of images to grayscale, removal of background, and data augmentation techniques to address the severe class imbalance that was present in the dataset.

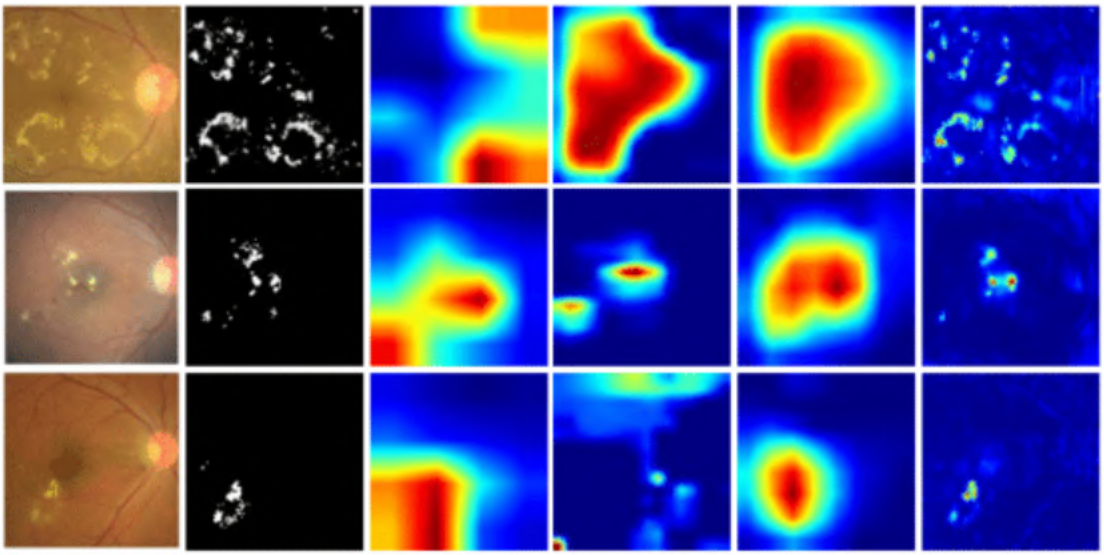

FIGURE 2.3: Grad-Cam visualization of Fundus Images [3]

The model utilizes fewer parameters compared to Inception V3, VGGNet, and ResNet.

A survey on the Convolutional Neural Networks working for the detection of the DR was conducted by Abushawish et al [5]. It covers cross-dataset deployment for the detection of DR and its severity levels by using transfer learning, end-to-end learning, and existing hybrid approaches.

Twenty-six pretrained models of CNNs on ImageNet were evaluated by the authors for the sake of comparison in this study. DRGF is the dataset that was employed in this study. The highest accuracy of 78.12% and sensitivity of 78.00%. was obtained on DenseNet121 architecture.

Although generic preprocessing techniques are discussed, the study does not explains the techniques that are required for the reduction of unwanted bias from the dataset.

Khan et al [6] introduced a technique that uses very few parameters in order to increase the training time of the model and the time of convergence. For this, they proposed a model named VGG-NiN, which in-cooperates spatial pyramid pooling layer of VGG16 along with the Network in Network layer. Both of these are stacked on each other to make the model capable of handling any scale of DR image.

The addition of NiN introduces a non-linear nature in the model, which can result in better classification. The average sensitivity reported in the paper was 55.6% with total parameters of 45.49 million.

It is noted that although the number of parameters is reduced, the accuracy and sensitivity are also compromised, which is one of the most important matrices in medical image classification.

Shamrat et al. [7] proposed a novel model named DRNet for the classification of diabetic retinopathy and compared the results with 16 CNN models. The study created a Dataset of the real time patients from the hospital in India, containing 3662 images.

In order to remove the class imbalance, the images were expanded to a total of 7500 by applying data augmentation of flipping, zooming, and cropping.

The preprocessing steps included image noise removal by a median filter and image enhancement technique by applying gamma correction. The proposed model used 3 convolutional layers with pooling layers in between and a dropout layer between two Fully Connected (FC) layers named as FC Layer 1 and FC Layer 2.

The model results were the accuracy of 97% and the sensitivity of 98%. The results are only tested on the dataset, which is not publicly available. There is a need to check the generalizability of the proposed model on the other datasets as well to test its applicability in the real time settings.

Another study proposed a revised structure of ResNet-50 [9] in order to overcome the problem of over-fitting in ResNet-50. The weights of the layers are adjusted by an adaptive learning rate in each generation. The study employed the Standard Operating Procedure (SOP) in its methodology.

The best features are obtained from the ResNet-50 with the help of a visualization tool shown in figure 2.4, which is used in the study, instead of the end result features after the pooling layer.

The results depict that the revised structure of ResNet-50 performs better than the original ResNet-50. The proposed modified architecture was tested on the

EyePACS dataset, and it produced an accuracy of 74.32%. In this paper, the model focuses on a single disease dataset and needs to be tested on the multi-disease datasets as well.

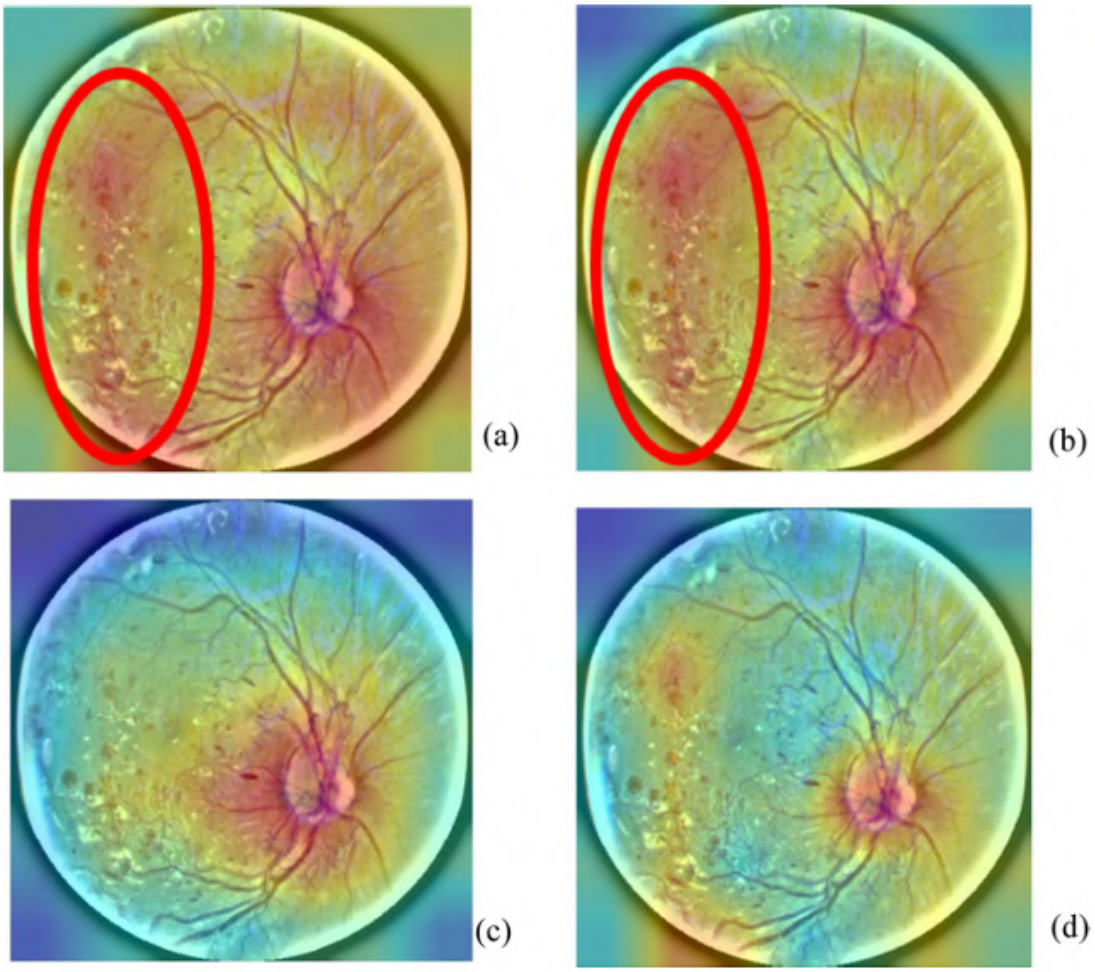

FIGURE 2.4: Features from multiple Convolutional Layers of ResNet-50 [9].

Das et al. [10] conducted a study that evaluates a total of 26 Deep learning models, including CNNs. The results depict that the model, which was highly affected by the problem of over-fitting, was ResNet-50 when compared to Inception V3. The models were tested on the EyePACS dataset.

The best results were reported by EfficientNetB4, whereas InceptionResNetV2, NasNetLarge, and DenseNet169 also performed well in detecting DR. The highest accuracy achieved was 79.11% on the test dataset. The sensitivity of the proposed model towards the disease class was not reported.

Das et al. [11] proposed a novel approach for the detection of DR by creating a Compact CNN Design. The study was verified on 4 publicly available datasets, including DRD, Messidor-2, and IDRiD.

The results presented the highest accuracy of 96.74% on Messidor-2, which is a single disease small dataset. It is noted that although the model does not require high computational power due to its compact structure, the sensitivity was compromised, especially in complex large datasets.

It is also noted in several studies that the pooling effects on the images in CNN suppress the fine-grained, small lesions, which are required for early detection of DR [28]. This effect is visually depicted in the figure 2.5.

Global Average Pooling (GAP) converts the final output tensor of CNN into a single 1D vector by taking the average of every feature map, resulting in a single value against each channel separately.

All the regions of the fundus image are equally treated by it. On the other hand, Global Max Pooling (GMP) considers the single maximum value for each channel, resulting in focusing on the strongest activation of that particular location. [29]

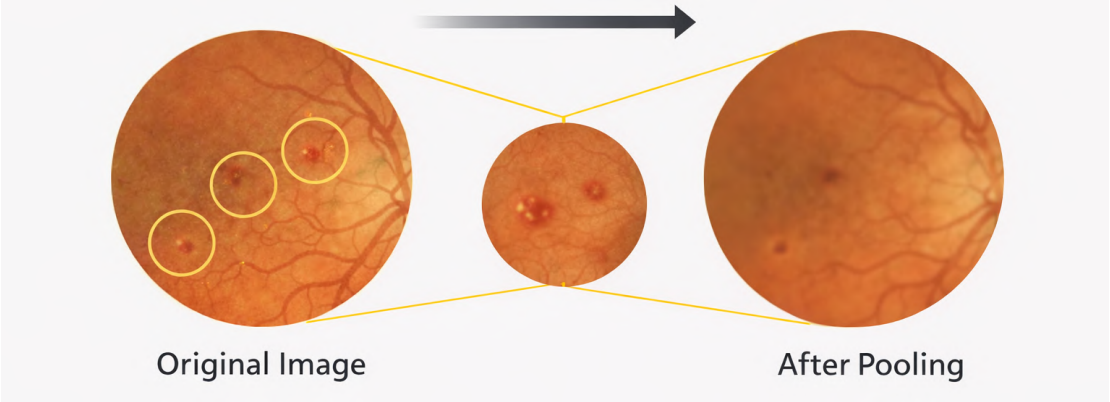

FIGURE 2.5: Effect of Global Average Pooling on fundus images[28]

A comparison of top five CNN techniques used for DR detection in the literature review is summarized in Table 2.1 given below.

TABLE 2.1: Summary of DR Detection Papers Using CNN Architectures

Ref	Architecture	Datasets	Key Strengths	Limitations	Key Results
[1]	Adaptive preprocessing with CNNs	Messidor, RFMiD, EyePACS	Improves low-quality fundus images, including cataract samples	Global contrast enhancement suppresses lesion-level details	Acc: 95.88%, 91.56%, 89.81%, Sen: 86.90%, 85.48%, 82.33%
[2]	VGG-19, VGG-16, EfficientNetB0, ResNet-152	RFMiD	Multi-disease settings with augmentation applied	Low sensitivity; model inclined towards Non-DR class	Acc: 90.00%, Sen: 87.14%
[3]	Hird-Net: hierarchical inception residual dense network	IDRiD, APTOS, DDR, EyePACS	Squeeze-and-Excitation improves feature learning	Noise misclassified as lesions; DR-only datasets used	Acc: 93.46%, 82.45%, 79.94%, Sen: 95.14%, 87.02%, 85.14%
[4]	Parameter-efficient CNN framework	EyePACS	Reduced complexity with competitive performance	Not assessed against Multi-disease settings	Acc: 90.60%, Sen: 87.64%
[5]	Survey of CNN models	DRGF	Analysis of 26 CNN architectures	Class imbalance not addressed	Acc: 78.12%, Sen: 78.00%

2.3.2 Transformer-based Approaches

Karkera et al. [12] utilize a fine-tune transformer-based model for the automated detection of DR in order to understand the severity level of DR.

The dataset that they used was APTOS2019. The results were quite promising, giving an accuracy of 94.63% with a sensitivity of 91%. The model was evaluated on the APTOS-2019 dataset only, which limits its generalizability to work better in the multi-disease conditions as well. The basic architecture of Vision Transformers is shown in figure 2.6.

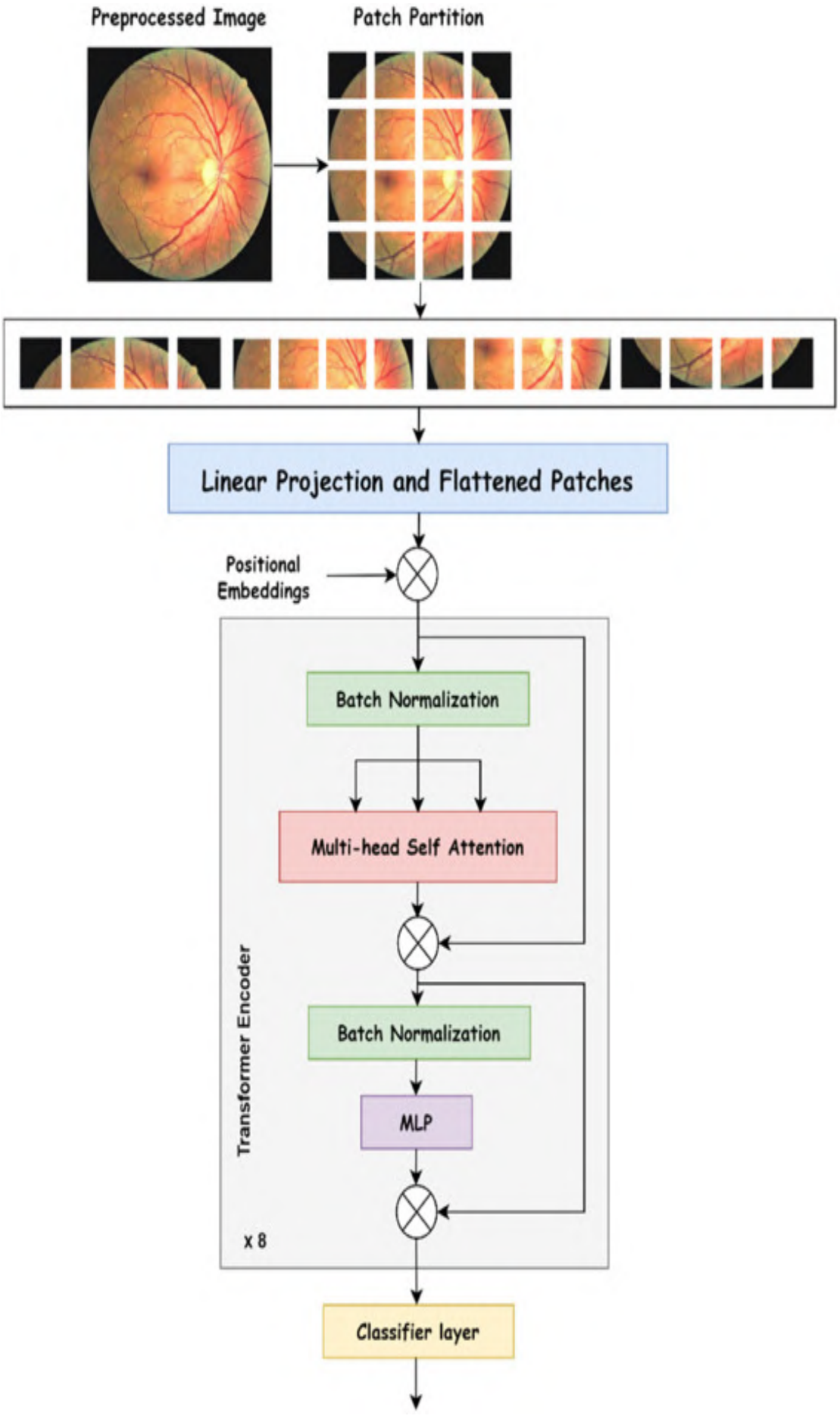

FIGURE 2.6: Vision Transformer Architecture [25]

Mohan et al. [13] proposed a Vision Transformer (ViT) for the DR classification on the fundus images. The dataset that the paper utilizes is EyePACS.

First, the images were converted into patches. Then, only the non-overlapping patches were considered further. All the patches were flatten in a 1D array before applying linear and positional embedding on them.

Multi-head self-attention mechanisms were used to understand the global relationship of the patches. Softmax was applied in the final classification layer.

The results showcase the accuracy of 91.4% and the sensitivity of 92.6%. This explains that Transformers are better than the CNN models because of their capabilities to understand the relationship between different image spatial locations.

Another study [14] indicated the use of the Vision Transformers (ViT) in the detection of DR by using two main modules in their proposed framework.

One was termed as FEB (Feature Extraction Block) and GPB (Grading Prediction Block). The model was able to capture various spatial relationships among macula, optic discs, and lesions to understand the DR severity level.

The paper also conducted a detailed ablation study for the verification of the methodology even further. The accuracy was 82.23%, and the sensitivity was 81.40% on the DDR Dataset.

It is noted that transformers work even better in detecting the grading level of DR.

Wu et al. [15] also employed vision transformers for DR detection. They also used a multi-head self-attention mechanism with positional embeddings that are used to provide the location details of the patches.

The results reveal an accuracy of 91.4% and sensitivity of 92.6% on a dataset. The dataset name was not explicitly mentioned, which makes the study difficult to reproduce.

The summary of all the above transformer-based approaches is summed up in the following table 2.2.

TABLE 2.2: Summary of DR Detection Papers using Vision Transformers

Ref	Datasets	Key Strengths	Limitations	Key Results
[12]	APTOS2019	Utilized the global spatial context of the images	DR detection is not tested in multiple disease environments	Acc: 94.63%, Sen: 91%
[13]	EyePACS	Only non-overlapping patches were selected before flattening	No details on preprocessing and augmentation were stated	Acc: 91.4%, Sen: 92.6%
[14]	DDR	Contains two modules named FEB and GPB	Architecture not tested on multiple disease clinical settings	Acc: 82.23%, Sen: 81.40%
[15]	Not Specified	Detailed Multi-Head Self-Attention Mechanism (MHSA) used	No details on dataset and positional embeddings given	Acc: 91.4%, Sen: 92.6%

2.3.3 Hybrid based Approaches

Bala et al. [16] produce a novel hybrid approach of combining CNNs with transformers in order to combine both the strengths of local and global context to detect DR. The paper named the model CTNet.

The first step consists of a CNN layer containing residual connections to extract local features, which are further converted into patches with the help of a transformer module to handle long spatial relations regarding how a lesion is related to the other lesions present sparsely, with the help of the Self Attention Mechanism.

The pooling layer is not performing the max pooling on the tokens; rather, it is being implemented on the patch level to make the model more energy efficient. The model was tested on APTOS2019, and the results reveal very promising results with an accuracy of 98% and 98.8% sensitivity, which is quite high. The limitation of the technique lies in the selection of the dataset.

As APTOS2019 is specific to only DR, and in real clinical environments contain other diseases as well.

Sekar et al. [17] also worked on the integration of CNN with the transformers for more accurate detection of DR.

DT-Net was the name of the proposed hybrid architecture, which used the DenseNet121 with Transformer. The model was trained and tested on the APTOS2019 dataset. The proposed method without using the class weights in cross-entropy results in a sensitivity of 85%, whereas when class weights are used in cross-entropy, a gain of 10% in sensitivity is seen.

Darapaneni et al. [18] implement an integrated approach of CNNs with transformers. The framework was tested on EyePACS.

The results reveal the accuracy of 96.3% with the sensitivity of 96.7%. Initially, the paper discussed the issue of multiple diseases present in the retina, which are similar to DR, and the model can classify them as DR.

But the training and testing of their technique were particularly using the datasets, which were disease-specific and only contained one disease, which is DR vs healthy images.

Fan et al. [19] introduced a hybrid approach by combining ResNet-50 with Vision Transformers (ViTs). The integrated model was tested on the EyePACS dataset available on Kaggle. The results displayed the accuracy of 88.40% with a sensitivity of 86%.

Ikram et al. [20] suggested an integrated model of ResNet 50 with transformers named as ResViT in order to automate the process of DR screening accurately.

The model utilizes the APTOS2019 dataset for the particular study. The paper applied data augmentation to remove class imbalance from the dataset and to generalize the results. The augmentation techniques consisted of flipping in horizontal directions, along with rotation, and zooming.

In order to justify the results produced by the ResViT, further explainable AI was implemented using Grad-CAM to analyze the classification. The accuracy of the proposed novel technique was 89.4% with a sensitivity of 93%. The results depicts the over-fitting problem of ResNet-50 was solved by amalgamating it with

transformers to give them a better understanding of the lesions present in fundus images.

Yamuna et al. [30] introduces a novel deep learning approach for the detection of diabetic retinopathy in the RFMiD dataset, combining convolutional feature learning with transformer-based attention mechanisms.

Their approach uses CNN blocks to extract local texture features, and a transformer block to capture global interactions between retinal regions. This approach enhances feature extraction for multi-class retinal disease classification.

The model achieves better classification performance than traditional CNN-based methods, especially in sensitivity and AUC, showcasing its capability to capture subtle lesion patterns in fundus images.

Kuruba et al. [31] proposes a hybrid design of deep convolutional networks and transformer encoders for diabetic retinopathy classification. The CNN is used for multi-scale feature extraction, and the transformer encoder improves global context recognition of retinal lesions.

The architecture is tested on the RFMiD dataset and performs well in terms of sensitivity and specificity. Yet, the paper does not specifically mention confusion among multiple pathologies in retinal diseases, an issue often encountered in clinical practice where similar-looking lesions can co-exist.

Singh et al. [32] research proposes a vision transformer-based CNN for detecting diabetic retinopathy in the RFMiD dataset.

The model integrates CNN feature extraction and vision transformer layers to exploit both localised and global information in fundus images. The approach achieves high accuracy and is more robust than CNN-only models.

But the inclusion of transformers makes the model computationally heavy, which could hinder its use in real-world clinical settings.

The hybrid approaches for DR detection can be summarized in the following table 2.3

TABLE 2.3: Summary of Hybrid CNN–Vision Transformer Based DR Detection Papers

Ref	Datasets	Key Strengths	Limitations	Key Results
[16]	APTOS2019	Pooling applied on patches improves energy efficiency	Dataset selection limits generalization	Acc: 98% Sen: 98.8%
[17]	APTOS2019	Class-weighted cross-entropy enhances learning	No synthetic data generation for augmentation	Acc: 90% Sen: 95%
[18]	EyePACS	Validated across multiple datasets	Not evaluated in real-time or multi-pathology settings	Acc: 96.3% Sen: 96.7%
[19]	EyePACS	ViT integration reduces ResNet-50 overfitting	Class imbalance handling was not discussed	Acc: 88.40% Sen: 86%
[20]	APTOS2019	Noise removal and Grad-CAM enhance explainability	Not tested on multi-disease datasets	Acc: 89.4% Sen: 93%
[30]	RFMiD	Combines local CNN features with global attention	Computational complexity increases due to ViT	Acc: 92.10% Sen: 90.20%
[31]	RFMiD	Fusion of hierarchical CNN features and global attention	Limited focus on multi-pathology confusion	Acc: 90.85% Sen: 89.70%
[32]	RFMiD	Spatial CNN extraction with context modeling	High computational cost due to transformer modules	Acc: 88.40% Sen: 86%

2.4 Preprocessing Techniques for Fundus Images in Diabetic Retinopathy Detection

Although the above-mentioned studies were already using some sort of preprocessing on the fundus images to improve the performance of the detection.

But some studies were explicitly stating the preprocessing steps in detail to address the wide range of problems that exist in fundus images, as already explained in Chapter 1. Some of the closely related studies on the enhancement of fundus images for DR detection are discussed below.

Abbasi et al. [1] proposed a preprocessing technique that was even able to enhance the cataract images in such a way that DR could be detected from them. Cataract is a disease of the eye in which a gray cloud starts appearing in the natural lens of the human eye, as shown in figure 2.7. This hinders the light from passing

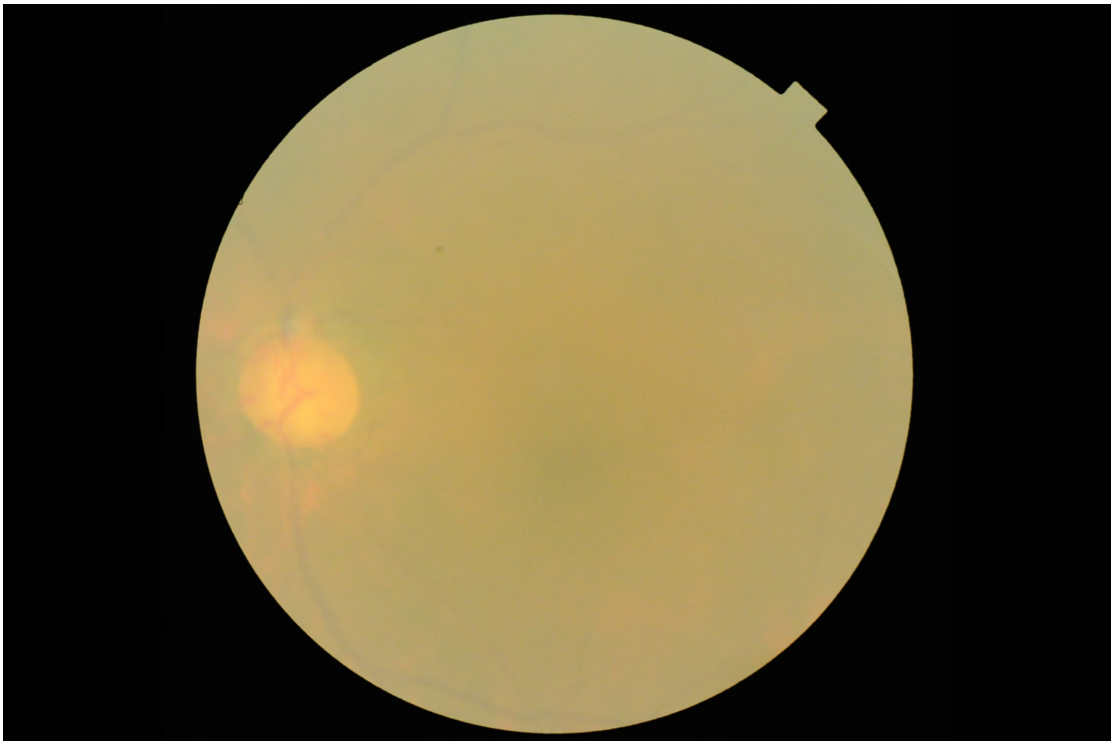

FIGURE 2.7: Fundus image of a patient having Cataract [8]

through the lens and hence greatly impacts the illumination that is required while capturing the fundus images. The lesions of DR become almost invisible in severe cases of cataract.

The study proposed an extensive image processing technique utilizing adaptive gamma correction, intensity normalization, and histogram equalization for the enhancement of the images. It was difficult to obtain the cataract pair of images, so the study produced the synthetic cataract images using Generative Adversarial Networks (GANs), and then the preprocessing technique was applied to those images.

The results are shown in figure 2.8. The technique was verified on three of the datasets, named Messidor, RFMiD, and EyePACS.

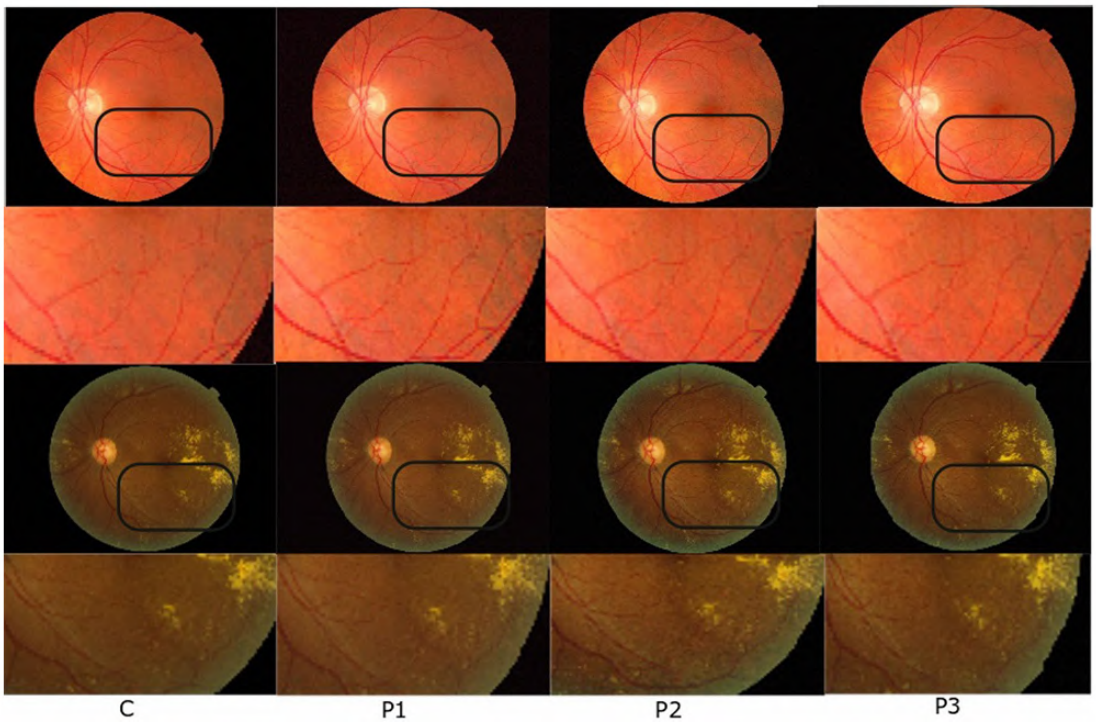

FIGURE 2.8: Effects of preprocessing contrast enhancement [1]

Abbood et al. [33] propose an image enhancement algorithm for improving the quality of the fundus images for the detection of DR. The technique first crops the image to remove any unnecessary details that may hinder the process of enhancement.

A Gaussian blur is applied to remove noise from the images. The technique is evaluated on two of the datasets, named Messidor and EyePACS. The results were also validated in the real-time environment of smart hospitals as an Internet of Medical Things (IoMT) application. The results reveal that with enhancement, the classification accuracy of DR was improved.

Balashunmugam et al. [34] conducted a thorough review of the retinal fundus images that contain many retinal diseases. Hence, the preprocessing of these images are essential for pinpointing the disease present at various location of the retinal layer.

For this purpose, the paper studied the histogram equalization (HE) technique that improves the overall contrast of the image, Contrast Limited Adaptive Histogram Equalization (CLAHE) technique, which is a more refined version of HE because it is applied adaptively to each pixel without over-amplifying the contrast of a pixel.

The paper also focused on the filter-based techniques, including Gaussian, Median, and other filters related to the spatial locality of the pixels.

Contrast adjustment approaches, along with edge and structure preservation mechanisms, were also reviewed. This comprehensive study explained that the choice of technique depends upon the nature of the pathology, as each pathology contains different indicators that need to be detected in the retina of the human eye.

Another novel preprocessing technique was developed by Naz and Ahuja [35], which combined a modified form of Fuzzy C-means clustering (soft-clustering) approach with Support Vector Machine (SVM).

Firstly, the images were converted from the RGB channel to the HSI channel in order to retain the color information. Secondly, noise was reduced from the images by using the median filter. Intensity Histogram equalization was used to improve the contrast of the images.

False microaneurysms were detected with the help of connected components. The optic disc was also removed from the images due to its similarity with microaneurysms through morphological operations. The results reveal better performance in DR screening compared to images without any preprocessing.

Alaguselvi and Murugan [36] introduce a denoising technique utilizing CNN termed as DnCNN for enhancing the fundus images containing DR. DnCNNs improved the denoising architecture of images. Numerous histogram-based techniques, including HE, Adaptive Histogram Equalization (ADHE), CLAHE, and Exposure-based Sub-Image Histogram Equalization (ESIHE), are utilized for the preprocessing of the DR images.

The results were evaluated through measuring Peak Signal to Noise Ratio (PSNR) and Structural Similarity Index (SSIM). The results reported the PSNR value to be 25.43 and the SSIM to be 0.3584 on a DRIVE dataset.

The major limitation of the technique is that it has not been verified on a publicly available dataset, limiting the reproducibility of the results. The results are summarized in table 2.4 given below.

TABLE 2.4: Summary of Preprocessing and Image Enhancement Techniques for DR Detection

Ref	Technique	Datasets	Key Strengths	Limitations
[1]	Adaptive Gamma Correction, Intensity Normalization, and HE	RFMiD, EyePACS, Messidor	Adaptive feature enhancement improves visibility of retinal structures	Introduces smoothing effect which makes small lesions invisible
[33]	Edge enhancement by utilizing Gaussian Blur	EyePACS	Enhances contrast and improves visibility of retinal lesions	Risk of over-enhancement leading to loss of fine details
[34]	Survey of enhancement (HE, CLAHE, filter-based)	CHASE and DRIVE	Comprehensive comparison for improving lesion visibility	Lacks unified experimental validation across all techniques
[35]	Novel contrast enhancement technique	IDRiD and DIARETDB	Improves contrast between lesions and background regions	Performance dependent on parameter tuning and image quality
[36]	Analysis of CNN-based pipeline (HE, CLAHE, ESIHE)	DRIVE	Evaluates impact of preprocessing on feature learning	Limited generalization across multi-disease or heterogeneous datasets

2.5 Research Gap

The current literature focuses on disease-specific data, where the images of the retina typically have a single class (either DR or normal), so the lesion patterns are fairly distinct for deep learning models. However, real-world patients may suffer

from various retinal diseases, such as hypertension, glaucoma, central retinal vein occlusion (CRVO), and branch retinal vein occlusion (BRVO), which share similar patterns in lesion distribution to DR [8]. This similarity in the lesions of DR and other retinal pathologies create diagnostic uncertainty and affects the sensitivity of models, leading to a false negative rate for DR lesions.

Although there are a few studies that have done the classification of DR by using a multi-disease dataset, but there sensitivity is not higher than 87% in terms of CNN based models and upto 90.20% in terms of hybrid approaches depicting that the existing models struggle to detect DR accurately in the presence of other similar retinal pathologies. This is due to the fact that lesions are misclassified as noise or other anatomical structures. Hence, as a result, there is an increase in false negatives, subsequently lowering the sensitivity of the diagnostic system. There is a need for efficient lesion extraction from images containing multiple retinal pathologies to reduce the false negatives of DR and increase the sensitivity.

This chapter discussed current deep learning methods for diabetic retinopathy detection, such as CNN-based, Transformer-based and hybrid models, as well as image preprocessing strategies to enhance retinal images. The review shows that while there has been significant advancement, the methods are still struggling to separate DR from other visually similar retinal diseases, especially in the presence of multiple diseases.

Chapter 3

Research Methodology and Proposed Framework

3.1 Overview of CALF-Net Methodology

This chapter focuses on the methodology that was adopted for the detection of DR by proposing CALF-Net (CNN Attention-Leveraged Transformer Network) using fundus images of the retina.

The chapter covers the rationale behind the selection of the architectural design, the learning mechanism, and the overall flow of the data within the proposed system by explaining the presence of each component and how they contribute towards the final classification of DR.

CALF-Net combines data preparation with representation learning and a classification mechanism to automate the detection of DR. The proposed pipeline integrates Convolutional Neural Networks (CNNs) with Vision Transformers (ViT) in order to address variability in lesions and limitations seen in standalone systems.

The CALF-Net consists of data acquisition, along with preprocessing steps to address the challenges of the fundus images. The next step is to mitigate any existing class imbalance by employing data augmentation techniques that are specific to fundus imaging and retinal diseases. Following this step is the feature extraction

module of CALF-Net, which is responsible for local-level feature extraction using a CNN backbone that is carefully selected after performing a series of experiments that are discussed later.

The next step is to understand the global relationship among the extracted features by integrating them with Vision Transformers. This aids in providing information about the presence of the disease indicators on various anatomical spots of the retina. The final representation is passed through a multi-layer perceptron (MLP) for the classification of the disease.

The CNNs, along with ViT, are specifically chosen for the proposed methodology for two reasons. The first one is that the CNNs are known for their efficiency in extracting the features and learning them, thus saving time over a large dataset compared to manual analysis of fundus images, as stated by the literature as well.

The choice of ViT, on the other hand, is to reduce the false negatives in the system and improve the sensitivity of the system.

As the literature has reported it that the use of the transformers alone has even able to outperform the CNNs due to their inherent nature to understand the overall context of the fundus image and the presence of all disease indicators with other anatomical structures.

But as transformers need a large amount of data to understand the features, we cannot employ them standalone in the field of medical imaging, where the data is not present in huge amounts, especially in clinical settings where other pathologies co-exist with DR.

The comprehensive explanation of the proposed methodology is discussed in the subsequent sections of this chapter. Figure 3.1 represents the complete flow of the proposed model.

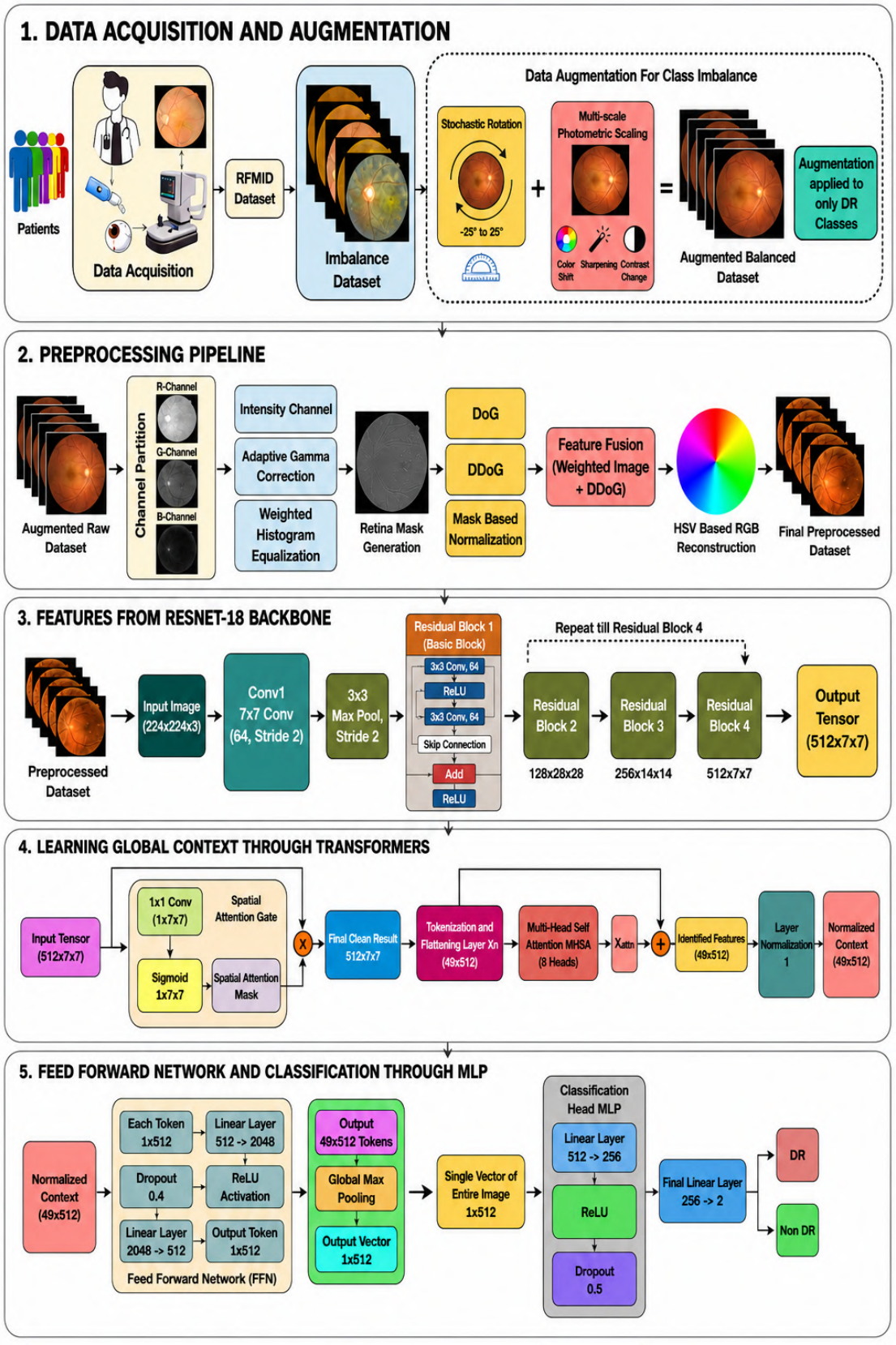

FIGURE 3.1: Proposed Architecture of CALF-Net

3.2 Research Design

The effectiveness of the proposed CALF-Net is evaluated with the help of experimental, quantitative, and model-based research design to detect DR. The experimental nature of the research is defined by the development and controlled evaluation of the deep learning architecture. Evaluation Metrics are used to assess the proposed methodology in a quantitative manner.

3.2.1 Binary Detection of DR in Multi-Disease Setting

The goal is to detect the DR even in the presence of other similar retinal diseases, which contain similar indicators as those of DR, including lesions, patterns resembling hemorrhage, and other vascular abnormalities. The setting is chosen explicitly to replicate the real-world environment and to make the model capable of detecting DR more accurately. In this way, the proposed pipeline is trained and tested, unlike conventional binary DR detection scenarios, by posing a better reflection of real-world clinical settings where numerous ocular pathologies may exhibit those visual characteristics that are overlapping and present simultaneously with DR.

3.2.2 Problem Formulation

The binary classification of DR in a multi-disease clinical setting is the problem addressed in this research. The main objective of this research is to predict the presence of DR in the retina or the absence of DR, even though there are other retinal diseases that may exist and have similarities with DR visually.

Mathematically, the problem statement can be defined with the help of a colored retinal fundus image represented with I such that $I \in \mathbb{R}^{H \times W \times C}$ and height and width are represented by H and W, respectively. $C = 3$ defines the three RGB channels of an image. CALF-Net works on a mapping function defined as

$$f : I \rightarrow y \tag{3.1}$$

where $y \in \{0, 1\}$, in which non-DR class is represented by 0 and DR classes are represented by 1.

3.2.3 Input-Output Specification

3.2.3.1 Input

The fundus images will act as the input of CALF-Net, which can contain both DR and other retinal conditions, as well as varying imaging conditions.

3.2.3.2 Output

The output of the model is the classification of DR presented by a binary label of 0 or 1. The output explains the ability of the model to distinguish between DR and other retinal abnormalities.

3.2.4 Experimental Assumption and Operational Constraints

3.2.4.1 Label Reliability

The ground truth, which is presented in the dataset, is considered to be reliable and clinically validated. There still exists ambiguity regarding the label due to the presence of multiple diseases that can coexist in one image.

3.2.4.2 Variations in Image Quality

The fundus images, as explained earlier, may possess camera artifacts, poor lighting conditions, low contrast, and different demographics of patients. Therefore,

the proposed model should be able to work perfectly even in such varied conditions as well.

3.3 Class Distribution Characteristics of Dataset

The RFMiD dataset is used in this study, which contains a total of 46 different pathologies, in which DR samples are not uniformly distributed. It is noted that only a total of 376 images out of 1920 in the training sets are related to DR. This means DR constitutes only approximately 19.5% of the training dataset. Therefore, a severe class imbalance is present.

Class imbalance, in which the majority class (usually normal retinal images) outnumbers the minority class (disease cases), results in a model that is more likely to learn the characteristics of the majority class.

This biases it towards more frequently predicting the majority class, resulting in a greater number of true negatives, thus increasing specificity. But the model doesn't learn enough about the minority class (disease cases), leading to an increased number of false negatives and thus decreased sensitivity. There are a variety of methods to handle class imbalance in medical imaging data, such as data-level and algorithm-level methods. Data-level approaches include resampling techniques, such as oversampling the underrepresented class, undersampling the overrepresented class, or using data augmentation to artificially boost the class.

Algorithm-level methods can use class-weighted cross-entropy, focal loss, or cost-sensitive learning to modify the learning process in order to weigh the minority class samples more heavily.

In this study, we employ both data-level and algorithm-level approaches to overcome this problem. At the data level, geometric augmentation and photometric augmentation are used on minority class samples. Such data augmentation methods enhance the representation and variability of disease images, allowing the model to effectively learn the features of lesions.

At the algorithm level, a weighted cross-entropy loss function is used, with increased weights for minority class samples. This allows the model to give a higher penalty to the misclassification of disease cases, thus reducing the class imbalance.

These strategies, collectively, help improve sensitivity by better recognizing disease cases, thus reducing the number of false negatives. Although there may be a minor decrease in specificity, this is acceptable in certain medical applications where sensitivity is of utmost importance.

3.4 Data Augmentation Techniques to address class imbalance

Data augmentation is applied on the minority class to remove class imbalance. The training on the balanced dataset enables the model to focus on the patterns that are related to the minority class in order to obtain better generalization of the disease class samples.

In clinical practice, data are not balanced; the number of non-disease cases is much larger than that of disease cases. While this work uses data augmentation to balance the training data, the suggested model is not restricted to balanced data conditions. The geometric and photometric augmentation enhances the representation of the minority class features, and the weighted cross-entropy loss prevents the model from favoring the majority class. This enables the model to learn discriminative features that are resilient to class imbalance and can adapt to real-world non-balanced data.

Two augmentation techniques were used in the proposed methodology as shown in the Figure 3.2 .

i. Geometric Augmentation

ii. Photometric Augmentation

Both of these techniques are separately applied on the raw dataset to address class imbalance but the results have shown that geometric transformation outperforms the latter one in terms of sensitivity, whereas on the other hand, the photometric augmentation gives better AUC values, indicating that the model has a strong sense of separability in terms of positive and negative disease classes.

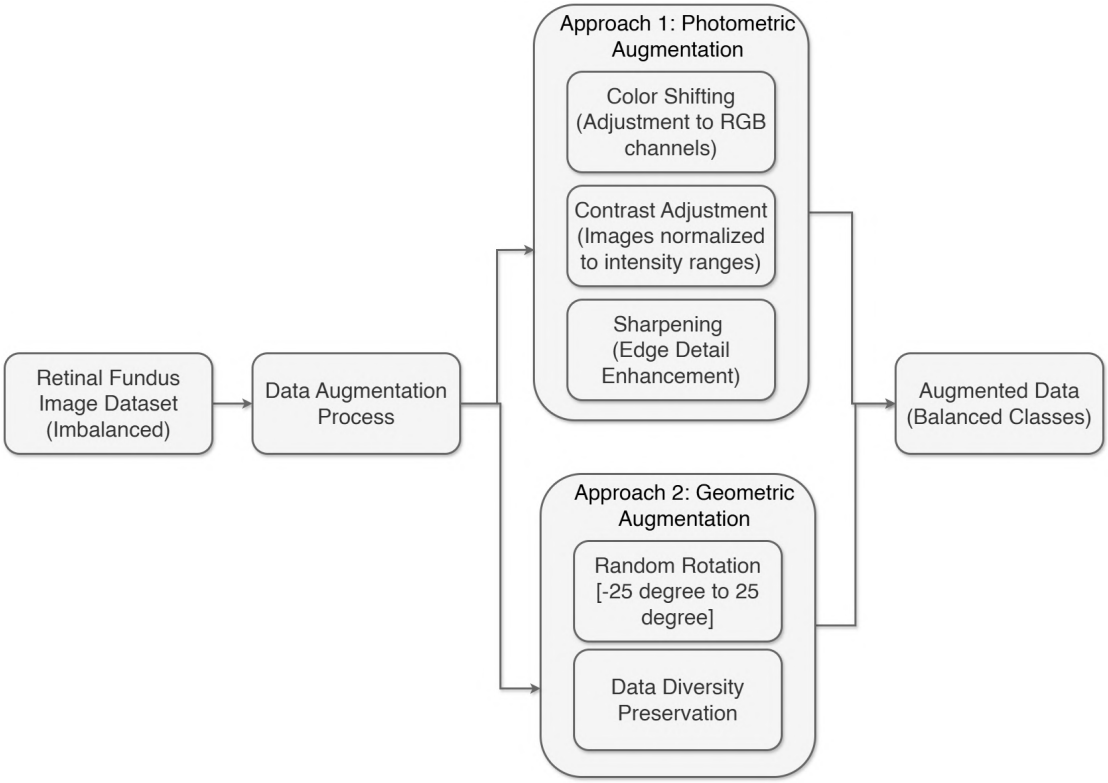

FIGURE 3.2: Data Augmentation Techniques used in the Proposed Research

3.4.1 Geometric Augmentation

The rotation augmentation applied to the dataset was inspired by a study conducted particularly on fundus images [37]. The rotation angle was randomly chosen in a limited range $[-25°, 25°]$. Each original image was rotated 10 times clockwise with unique values of angles in this range. The reason behind the selection of the aforementioned range was to replicate the minor differences in the head tilt and eye position variability that occur in clinical fundus photography without causing deformation of the entire global anatomy of the retina. The figure 3.3 represents the images before rotation and after rotation.

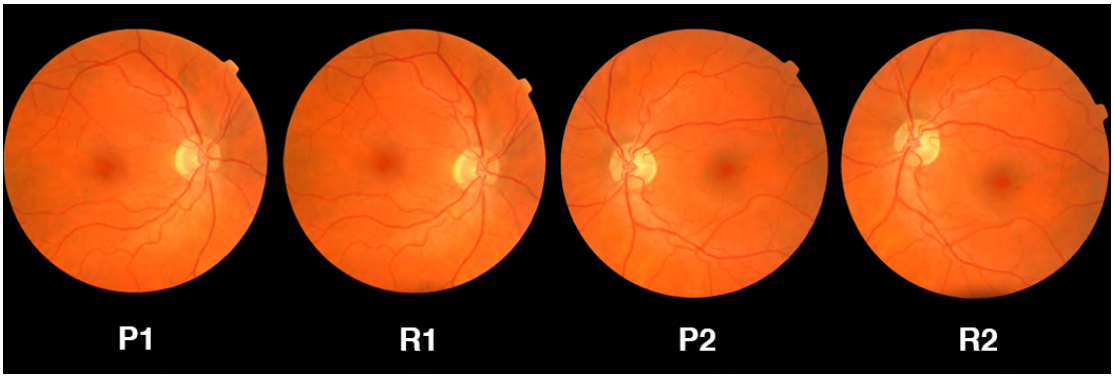

FIGURE 3.3: P1 and P2 are images before rotation, R1 and R2 are images after rotation

A rotation angle *theta* is taken from a uniform distribution for every augmentation instance:

$$\theta \sim \mathcal{U}(-25^\circ,\, 25^\circ) \tag{3.2}$$

Each image is rotated around its geometric center $(Cx,\, Cy)$, where:

$$C_x = \frac{W}{2}, \quad C_y = \frac{H}{2} \tag{3.3}$$

The dimensions of the image is maintained while applying the rotation.

The augmentation is only applied to DR cases in order to generate the synthetic images of DR for better generalization of the model. Let the DR samples be denoted by N_{DR} and the desired number of the DR samples after augmentation be denoted by N_{target}. The number of images that need to be augmented can be calculated using the following equation.

$$N_{\text{aug}} = N_{\text{target}} - N_{\text{DR}} \tag{3.4}$$

DR image IDs are randomly sampled with replacement during augmentation, and rotation is applied repeatedly until N_{AUG}. They create artificial visuals. By avoiding the duplication of identical samples, this regulated oversampling technique guarantees class balance.

Real-world variability is replicated with the help of rotation augmentation without distorting the spatial relations of the anatomical structures and the shapes as

well. The various changes in the orientation do not affect the model's training due to such a kind of augmentation, hence reducing the chance of over-fitting, which is important for automated DR screening systems used in diverse imaging contexts. The dataset distribution before augmentation and after rotation-based augmentation is shown in Table 3.1

TABLE 3.1: Dataset Distribution Before and After Rotation Augmentation

Stage	**DR Samples**	**Non-DR Samples**	**Total Images**
Before Augmentation	376	1544	1920
After Rotation Augmentation	1544	1544	3088

3.4.2 Photometric Augmentation

Differences in camera sensors, lighting settings, retinal pigmentation, and acquisition techniques all contribute to problems associated with fundus imaging.

It is prevalent from the literature that if various techniques are combined and applied to the fundus images may yield promising results.

Therefore, a compound strategy comprising color shifting, sharpening, and contrast alteration was implemented on the dataset [37]. Figure 3.4 illustrates the effect of the augmentation technique mentioned above on the fundus images.

3.4.2.1 Color Shifting

By separately disturbing each color channel, color shifting mimics color variations between devices and between patients. The transformed image I^{cs} is calculated as follows for a given shift tuple $(\Delta R, \Delta G, \Delta B)$:

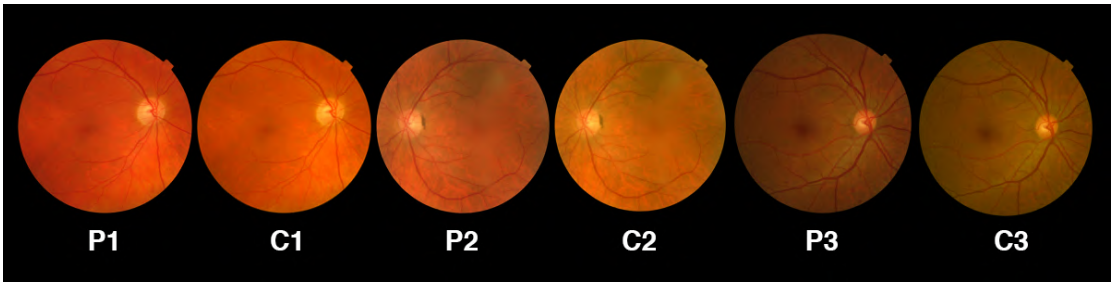

FIGURE 3.4: P1, P2, P3 are images before photometric augmentation, C1, C2, C3 are images after photometric augmentation

$$\begin{aligned} I_R^{cs}(x,y) &= \text{clip}\big(I_R(x,y) + \Delta R, 0, 255\big), \\ I_G^{cs}(x,y) &= \text{clip}\big(I_G(x,y) + \Delta G, 0, 255\big), \\ I_B^{cs}(x,y) &= \text{clip}\big(I_B(x,y) + \Delta B, 0, 255\big) \end{aligned} \tag{3.5}$$

There are three pre-established color shift settings.

$$(\Delta R, \Delta G, \Delta B) \in \{(20, -10, 5), (10, 15, -20), (-15, 5, 10)\} \tag{3.6}$$

These changes add realistic chromatic diversity without changing the features of structural lesions.

3.4.2.2 Image Sharpening

A sharpening technique utilizing unsharp masking is used to improve edge definition and highlight fine retinal structures like vessel borders and microaneurysms. First, an image that has been Gaussian-blurred is calculated:

$$I_{\text{blur}} = G_\sigma(I), \quad \sigma = 1 \tag{3.7}$$

where $G_\sigma(\cdot)$ denotes Gaussian smoothing with a 5×5 kernel. Next, the sharpened image I^{sh} is obtained as:

$$I^{sh} = \alpha I + \beta I_{\text{blur}} \tag{3.8}$$

Where $\alpha = 1.5$ and $\beta = -0.5$ [37]. By taking the difference of the Gaussian-blurred image from the original image, we are keeping only high-frequency components, and low-frequency components are eliminated due to the subtraction operator.

In simple words, the Gaussian-blurred image does not contain the details, and the original image does contain them, so the difference removes the background information and only focuses on the main details that are present inside the image.

3.4.2.3 Contrast Enhancement

Controlled intensity remapping is used to introduce contrast variation. Let $[i_1, i_2]$ indicate the range of the desired intensity. The normalized image is as follows:

$$I_{\text{norm}} = \text{Normalize}(I; i_1, i_2) \tag{3.9}$$

Where the minimum and maximum intensities of pixels are mapped by normalization of I to i_1 and i_2. The contrast ranges used are as follows:

$$(i_1, i_2) \in \{(30, 220), (40, 210), (20, 200)\} \tag{3.10}$$

Threshold is used to guarantee appropriate intensity bounds:

$$I_{\text{con}}(x, y) = \begin{cases} 255, & I_{\text{norm}}(x, y) > i_2, \\ 0, & I_{\text{norm}}(x, y) < i_1, \\ I_{\text{norm}}(x, y), & \text{otherwise.} \end{cases} \tag{3.11}$$

This method exposes the network to a broad range of realistic imaging situations by simultaneously modeling color variability, edge enhancement, and contrast fluctuations.

Therefore, the above augmentation techniques not only maintained the diagnostic integrity of the model but also enhanced the generalization by not altering the lesion geometry.

The proposed augmentation is particularly suitable for settings where clinical heterogeneity in medical imaging exists. The dataset distribution before and after applying photometric augmentation is shown below in Table 3.3.

TABLE 3.3: Dataset Distribution Before and After Photometric Augmentation

Stage	**DR Samples**	**Non-DR Samples**	**Total Images**
Before Augmentation	376	1544	1920
After Photometric Augmentation	1544	1544	3088

3.5 Data Preprocessing

The proposed preprocessing contains a number of steps, which are discussed in detail below and represented in Figure 3.5.

3.5.1 Channel Partitioning and Color Domain Conversion

Once the dataset is balanced by applying the above data augmentation techniques, the retinal fundus image is partitioned on the basis of its three channels, which are Red (R), Green (G), and Blue(B).

The reason behind this step is that we want to retain the color information of the pixels because it contains information about the disease. For instance, yellow color shows excudates, and similarly, red color shows the hemorrhages.

Following this step, the image is converted from RGB to HSI, which means Hue (the shade of the color), Saturation (the purity of that shade), and Intensity (the numeric pixel value which is associated with each pixel) are calculated. All the enhancement techniques will impact the intensity of the pixels.

3.5.2 Intensity Channel

The next step is intensity channel calculation. For this purpose, we have used the standard weighted equation of intensity, which is used to sum up the values of the

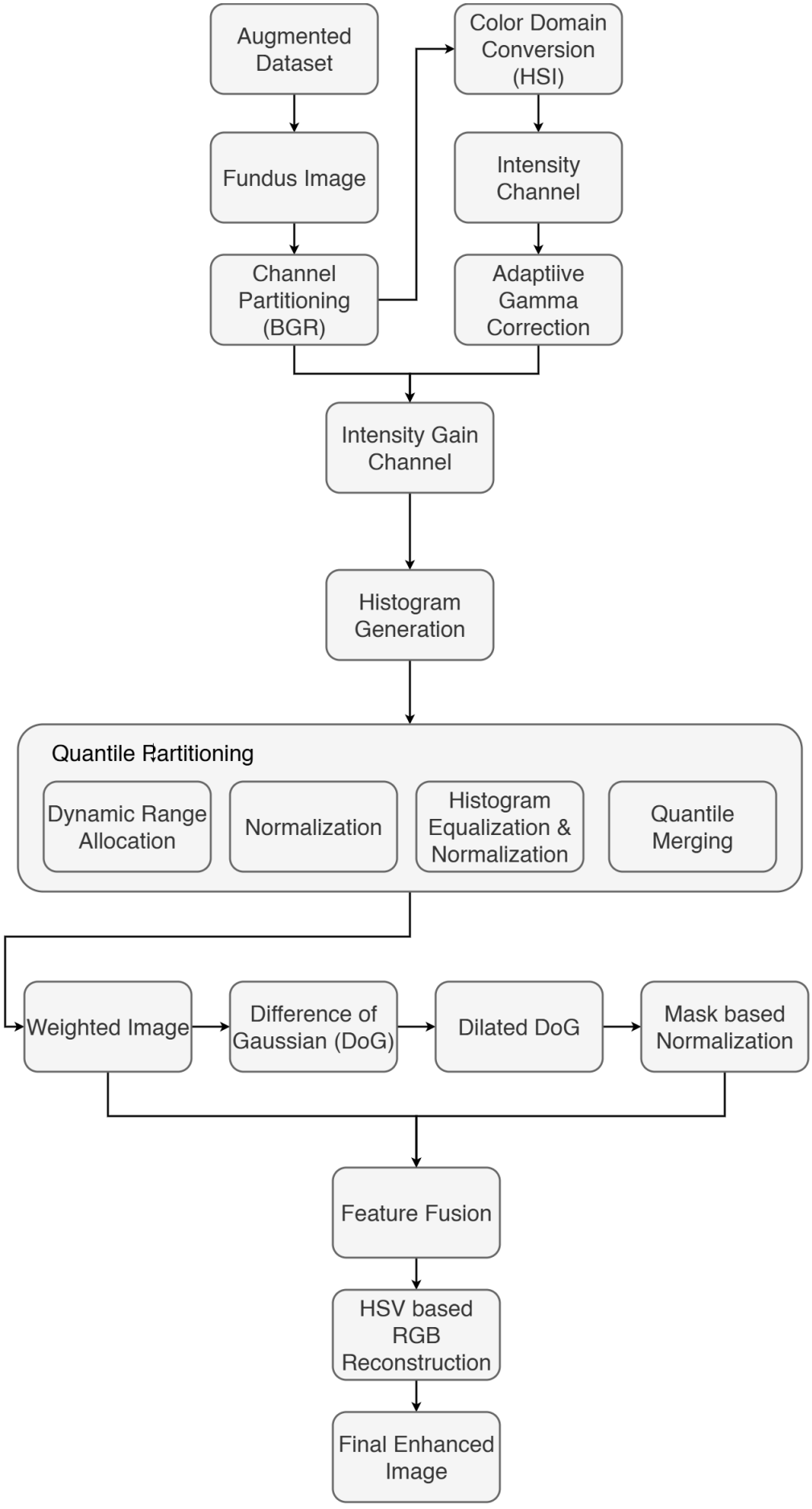

FIGURE 3.5: Proposed Preprocessing Pipeline

R, G, and B channels, where each channel is multiplied by its weights

$$\text{Intensity} = 0.299 \times R + 0.587 \times G + 0.114 \times B \tag{3.12}$$

The coefficients of the red, green and blue channels used in the intensity formula are based on the human visual model, which accounts for the different sensitivity of the human eye to different light wavelengths. Our eyes are most sensitive to green, followed by red, and least sensitive to blue. As a result, the weight of the green channel (0.587) is higher than the red channel (0.299) and lower than the blue channel (0.114).

These weights are not random; they are fixed constants corresponding to luminance models used in many color spaces, such as YUV, YCbCr and are standardized in international standards such as ITU-R BT.601 [38]. Their correctness is supported by psychovisual experiments that measure the brightness of different color channels in the human visual system. Thus, this perceptually weighted intensity representation is more accurate than the intensity derived from equally weighted RGB components.

The intensity formulation based on perceptually weighted RGB components better captures the perceived brightness of the image. This improved intensity is used in the preprocessing phase to enhance retinal features and fine-grained lesion patterns. Given the subtle nature of diabetic retinopathy, which includes the presence of microaneurysms, hemorrhages and exudates, it is important to maintain and enhance these features.

3.5.3 Adaptive Gamma Correction

The next step is to preprocess retinal images based on an adaptive gamma correction step to enable the enhancement of diagnostic information in the retinal images. The images of the fundus are either too dark or have uneven lighting that may hide minute signs of diseases such as microaneurysms and hemorrhages.

Our first step is to calculate a histogram of the intensity of the pixels in the image to view the distribution of the intensity along pixels.

As opposed to the traditional histogram equalization, which may make the image excessively bright, thus forming some noise, we modified the default histogram with a flat (uniform) histogram. This preserves the natural appearance of the image and diffuses the values of intensity more evenly.

$$H_m(i) = \alpha\, H(i) + (1 - \alpha)\, H_u \tag{3.13}$$

where H_u is the uniform histogram and $H(i)$ is the original histogram and α is the weighting factor usually taken as 0.7 [1]. The uniform histogram is calculated below.

$$H_u = \frac{MN}{256} \tag{3.14}$$

The uniform histogram is calculated to distribute an equal number of pixels across all the gray intensities. M is the number of rows and N is the number of columns. That is why the image pixels are divided by 256, because in grayscale, the intensity range varies from 0 to 255.

The cumulative distribution function (CDF) was calculated from the modified histogram. This CDF was used for the calculation of adaptive gamma correction (AGC). Instead of applying a constant gamma value, AGC varies the level of enhancement based on the brightness of the pixels of the image.

Darker regions become brighter and brighter, whereas the light areas receive minor changes, providing an equal balance. For darker regions, the value of gamma is small, and for already brighter regions, the value of gamma is large, so that a very minute change appears in the pixel.

$$\gamma(x, y) = \frac{1 - \mathrm{CDF}\big(I(x, y)\big)}{0.7} \tag{3.15}$$

where γ is the value calculated for each pixel, CDF is cumulative distribution function, $I(x, y)$ is the original intensity of the pixel. As a result of this step, the

low-contrast fundus images become clearer. The vascular and lesion features are more prominent.

3.5.4 Intensity Gain Channel and Histogram Generation

The next step is to evaluate the enhancement, which is done by adaptive gamma correction. This step calculates the total intensity gain by dividing the enhanced image by the original image using the following formula.

$$L(i,j) = \frac{\alpha'(i,j)}{\alpha(i,j)} \tag{3.16}$$

where $\alpha'(i,j)$ is the intensities after applying AGC and $\alpha(i,j)$ of original image.

Adaptive gamma correction is characterized by the gain function, which indicates the relative intensity. Because gamma correction utilizes non-linear changes of intensity values, the distribution of intensities may become unbalanced or skewed. Hence, the gain histogram is recomputed to observe the new distribution of pixel intensities and ensure that this process does not produce unwanted contrast distortions and/or enhance noise.

This allows verification of whether the transformation has successfully enhanced contrast without introducing inconsistencies in the retinal images. By recalculating the histogram, the preprocessing stage confirms that the enhancement of lesions' pertinent parts is carefully controlled, facilitating proper feature extraction and improved model training.

3.5.5 Quantile Partitioning

Quantile Partitioning refers to the division of the image into equal portions. Firstly, the image is converted into a 1D vector and is sorted in ascending order. Three quantiles have been created. The first one consists of low pixel intensities, the second one contains mid-range intensities, and the third one contains high intensities.

3.5.6 Dynamic range allocation with Normalized Power

After the successful Quantile partitioning, the images further went through dynamic range allocation of pixels, in which pixel-wise range is allocated to each pixel.

But this step sometimes over-enhances the image, so in order to make the contrast more realistic, the normalized function is applied, which tones down the intensities of the pixels by taking the log function.

3.5.7 Histogram Equalization

After all these extensive steps, the histograms of the image are created to analyze the intensity distribution over the pixels again. After this, Histogram Equalization (HE) is applied to the images, which makes the contrast of the image better.

The image obtained after this is named the weighted image. Now the visual appearance of this image has greatly enhanced, but there is a smoothing effect that is seen.

This smoothing effect can still hinder the detection of the minute lesions of the DR as shown in Figures 3.6, and 3.7 in which P2 and R2 represents the weighted images in which the smoothing effect is visually noticeable.

3.5.8 Retina Mask Generation

A retina mask is created from the original fundus image to isolate the retinal region and suppress non-informative background features. The black peripheral background that is usually seen in fundus images is separated from the foreground retinal tissue by this mask.

3.5.9 Difference of Gaussian

The Weighted-enhanced image from the previous preprocessing step is subjected to a Difference of Gaussian (DoG)-based enhancement approach in order to further improve lesion-level structures while reducing low-frequency background textures.

The purpose of this phase is to highlight minute pathological characteristics like hemorrhages and microaneurysms, which usually appear as localized intensity fluctuations in fundus images.

First, a 9×9 kernel with a fixed standard deviation of $\alpha = 2.0$ is used to execute a Gaussian smoothing operation [3]. This parameter selection allows for the efficient suppression of progressive illumination fluctuations while maintaining clinically significant high-frequency information, in line with the approximate spatial scale of tiny retinal lesions. The low-frequency backdrop of the retinal image is represented by the smoothed image.

Next, the Gaussian-blurred image is subtracted from the Weighted-enhanced image to calculate the DoG mask. This subtraction minimizes background roughness and uneven light effects while isolating high-frequency components that correspond to lesion margins, vessel edges, and other tiny retinal structures.

$$\text{DoG} = I_{\text{weighted}} - I_{\text{GaussianBlur}} \tag{3.17}$$

where $I_{weighted}$ is the weighted image and $I_{GaussianBlur}$ is the image after applying Gaussian Blur.

3.5.10 Dilated Difference of Gaussian

Morphological dilatation is performed to the DoG mask using a 2×2 structuring element in a single iteration to enhance the visibility of these retrieved features. This technique improves the continuity and prominence of lesion regions by gently expanding the observed feature boundaries. The selection of the 2 x 2 structuring element is because we do not want to over-dilate the lesions.

3.5.11 Feature Fusion

To match the dimensions of the augmented fundus image, the dilated DoG (D-DoG) mask is then duplicated over three color channels, which were extracted in the first step of preprocessing. The Weighted-enhanced image is then additively merged with this three-channel D-DoG representation, enabling the integration of localized structural information with global contrast augmentation.

$$I_{\text{fused}} = I_{\text{weighted}} + \alpha \cdot I_{\text{D-DoG,norm}} \tag{3.18}$$

where α is a weighting factor that regulates the D-DoG feature contribution and $I_{D-DoG,norm}$ is the normalized image after applying D-DoG. This combination guarantees the preservation of both local and global data.

Thus, this stage's final output reinforces lesion-specific features while maintaining the general intensity normalization attained by adaptive gamma correction and weighted histogram equalization. In automated DR screening tasks, where minor pathological indicators are crucial, this composite representation offers a more discriminative input for downstream learning.

3.5.12 HSV-Based RGB Reconstruction

Ultimately, HSV-based reconstruction is utilized to merge the fused representation back into the original color space. The Hue and Saturation channels of the HSV representation are kept from the original image, but the enhanced intensity information is added to the Value (V) channel. This method guarantees:

i. The appearance of natural color is preserved,

ii. Brightness, as opposed to chromatic distortion, reflects structural improvements.

iii. The finished product is still clinically realistic and visually comprehensible.

The final improved output used for downstream categorization is the reconstructed RGB image. The Figure 3.6, 3.7 shows the comparison of the raw image data with our progressive preprocessing steps.

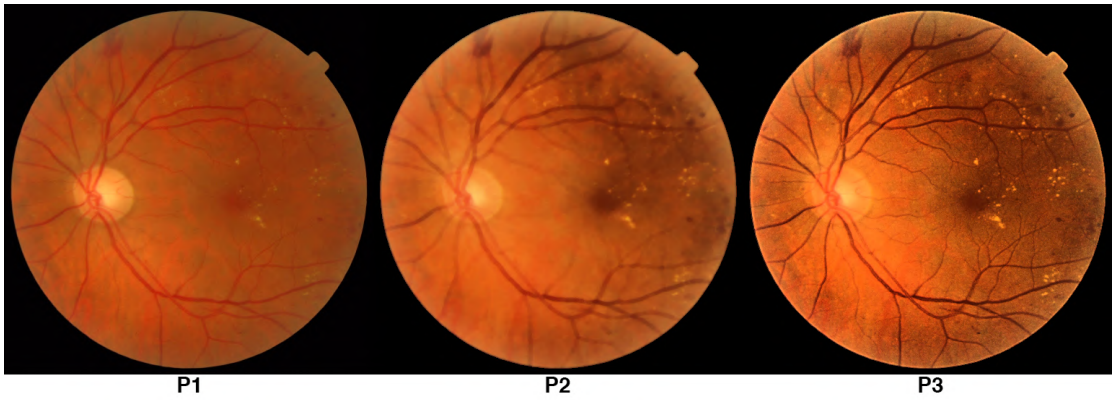

FIGURE 3.6: P1 is original Image, P2 is weighted image, P3 is final preprocessed image

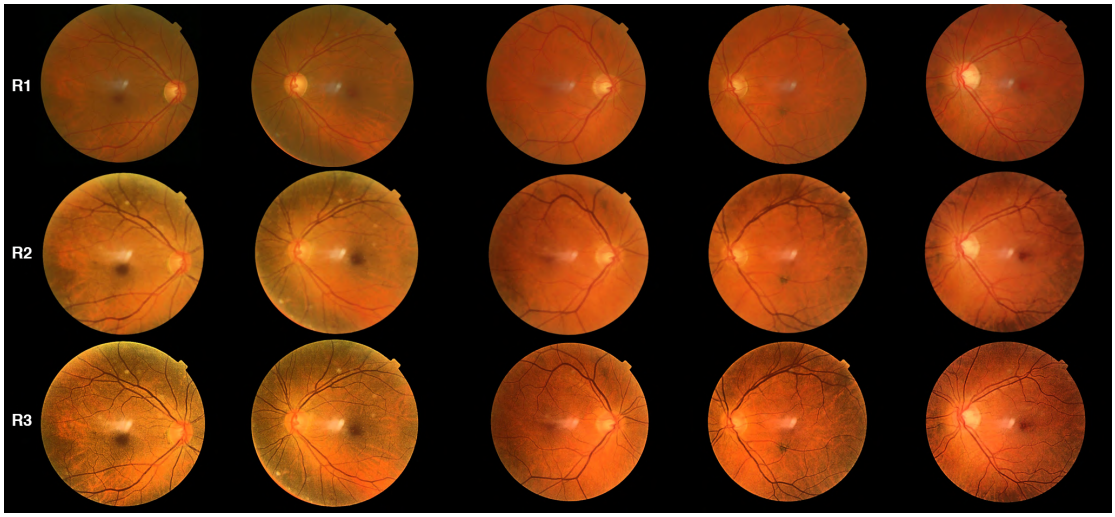

FIGURE 3.7: R1: Original Images, R2: Weighted Images, R3: Final Preprocessed images

According to the proposed preprocessing, the weighted images extracted after the histogram normalization step of the pipeline reveal that the images were enhanced with a smoothing effect.

So, in order to make the lesions appear sharper, there was a need to apply DoG and DDoG to indicate the lesions more properly.

3.6 Local Feature Representation Using Convolutional Neural Networks

The proposed CALF-Net framework uses a convolutional neural network (CNN) backbone to extract local lesion-level information from retinal fundus images. In order to accurately diagnose DR, the CNN must learn hierarchical feature representations that capture minute pathological patterns, including microaneurysms, hemorrhages, and exudates.

3.6.1 CNN Backbone Selection

The selection of ResNet-18 as a CNN backbone was a crucial step of the methodology. It was chosen for a number of reasons. The first and foremost reason behind the selection of ResNet compared to other CNNs is its capability to handle the issue of vanishing gradients. This problem is seen in many Convolutional neural network models because of the depth of the Convolutional layers. As the number of layers increases, during back-propagation, the loss function becomes extremely small. This results in a very minute or negligible updation in the early layers of CNNs, resulting in no learning in the initial layers. Subsequently, it leads to slow convergence and poor learning of the overall model. The second reason of choosing ResNet-18 specifically is due to its lightweight architecture, resulting in computational efficiency.

ResNet (Residual Networks) allows the convolutional layers to learn better without the issue of vanishing gradients. This is handled by the model with the help of skip connections, known as residual connections. The propagation of the gradient directly through the network is possible by adding these connections.

3.6.2 Overall Network Structure

The total number of layers present in ResNet-18 is eighteen, which includes the convolutional layers along with the final fully connected (FC) layer. Residual

blocks are present in ResNet, which are made up from convolutional layers and residual connections, also termed as shortcut connect allowing the model to learn through residual mapping, resulting in optimized convergence. The basic architecture of ResNet-18 is shown in figure 3.8 below.

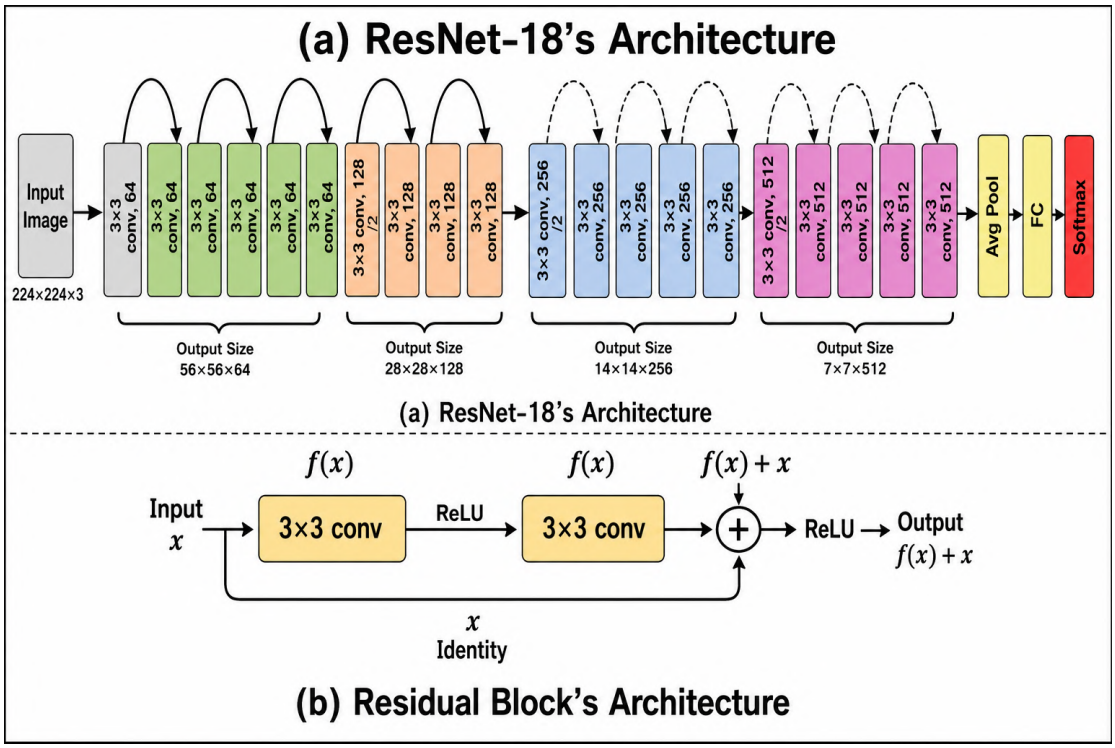

FIGURE 3.8: ResNet-18 Architecture [39]

3.6.3 Layer-wise Composition of ResNet-18

The layers of the ResNet-18 are described below.

i. First Convolutional and Pooling Layer: The initial convolutional layer is of 7x7 with a total of 64 filters.

   The stride is 2 in order to understand the basic features of the image, including edges and intensity variations. The resultant is batch normalized, and a max pooling layer is applied with a 3x3 filter size.

ii. Residual Stage 1 (Conv2$_x$): This stage comprises two residual blocks, where each block consists of a 3x3 convolutional layer with 64 total feature maps.

Due to the same dimensions that are used in the input and output of this block, residual connections are enabled to reuse direct features.

iii. Residual Stage 2 (Conv2_x): The number of feature maps is increased to 128 channels to understand more features. In this layer, a convolutional layer of 1x1 is applied. More abstract and mid-level representations are the output of this layer.

iv. Residual Stage 3 (Conv4_x): A total of 256 channels are obtained with the help of down-sampling in the first block.

Semantic features that are associated with the disease patterns are learned in this stage.

v. Residual Stage 4 (Conv5_x): The 4th block is the final block in which the number of channels is increased to 512 in total, where all the discriminative and lesion-specific features are obtained from the fundus images of retina.

3.6.4 Feature Aggregation and Output Representation

The global average pooling and the fully connected layers are eliminated from the architecture. The reason is that the global average pooling converts the image into a single 1D array, and all the locality information which are required for the transformer will be lost in this phase.

The fully connected layer of ResNet-18 is used for classification, and since we are using ResNet-18 for only the extraction of the local features and not for classification, it is also eliminated from the system.

3.6.5 Transfer Learning Strategy

The pretrained version of the ResNet-18 on ImageNet is used in this research. It helps the model to learn visual representation from the natural image dataset. As medical imaging datasets are limited to acquire.

Hence, the transfer learning is beneficial in the mentioned scenario. The pre-trained layers of the model allowed the feature extractors at low and mid-level range (including edges, textures, and simple shapes). These are fine-tuned for understanding the structures that are specifically related to the retina.

The preprocessed images obtained from the previous stage are given as input to the ResNet-18 backbone to extract the local features and generate representations that are discriminative in nature, allowing us to understand DR.

3.7 Spatial Gate for Noise Suppression and Lesion Enhancement

The features that are extracted from the convolutional neural networks are not always useful. Some features are presenting noise, variations in background, and sometimes imaging artifacts instead of true lesions.

The preprocessing steps where Difference of Gaussian (DoG) and Dilated difference of Gaussian (DDoG), are successfully able to enhance the retinal lesions; the noise also accidentally got enhanced. This happened specifically in the regions where there was low uneven illumination, vessel edges, or the presence of background textures.

Hence, there is a need for an additional mechanism that can isolate the dilated noise from the original lesions.

3.7.1 Definition of Noise in Retinal Images

The image patterns that are not according to the pathological retinal lesions are termed as noise. They can be created due to

i. Uneven variation in illumination: Due to variation in lightning conditions, some areas can appear to be bright or dark, which can be falsely considered as microaneurysms or hemorrhages.

ii. Background Texture: The natural pattern of the retina might produce reflections and pigmentation during enhancement that resemble small lesions.

iii. Imaging Artifacts: These include blur caused by motion or dust particles, which may appear on the lens of the camera.

Their appearance is irregular and varies in sizes which sometimes looks like lesions.

iv. False emphasis on vessel boundaries by filters: Vessel edges can be exaggerated after the application of DoG or DDoG filters, resulting in bright lines which are mistakenly considered as lesions.

The bright or dark spots that are diffuse in different regions are termed as noise, which have nothing to do with the disease. It does not have a pathological pattern, which is present in a disease. Usually, it appears to be irregular in shape and inconsistent across the image.

It is to be noted that missing pixels are not considered as noise in this research. This type of noise appears in corrupted data, whereas our proposed methodology removes the noise that comes from enhancement artifacts or variations that are found in the retinal background naturally.

If the above-mentioned noise is not treated, it may results in false negatives and can lead to a reduction in the accuracy of the proposed system.

3.7.2 Spatial Gate Design and Function

The feature representation from the CNN is of size 512×7×7, with 512 being the number of feature maps. This feature map is then fed into the spatial attention gate, which is a 1×1 convolution.

Rather than squashing all feature maps to a single value, the 1×1 convolution learns a linear combination of all feature maps to produce a spatial attention map as shown in the Figure 3.9.

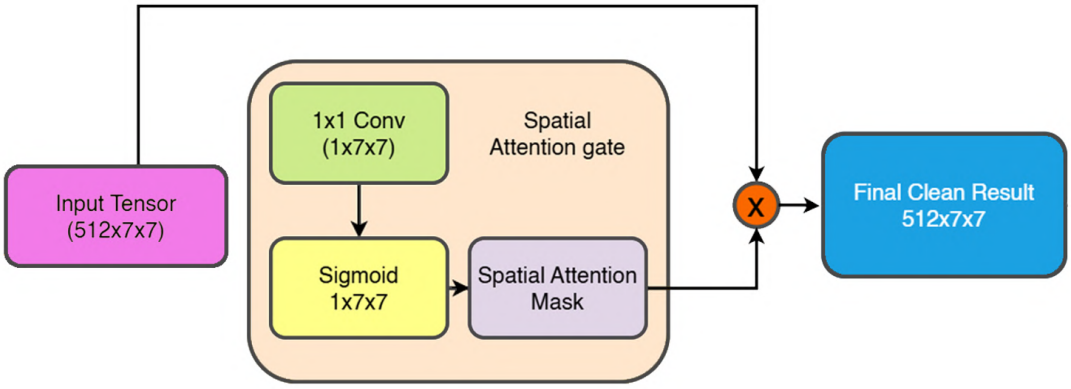

FIGURE 3.9: Spatial Attention Gate

The output is a 1-channel feature map (1×7×7) that represents the spatial importance. A sigmoid activation function is then used to scale these values to the range 0-1, resulting in the spatial attention mask. Regions closer to 1 are considered more important (e.g., lesion sites) and those closer to 0 are less important or noisy.

This mask is then multiplied with the original feature tensor, enabling the network to focus on the relevant spatial features and downplay less relevant or noisy features. Critically, this is a learnable and soft suppression, with less relevant features being down-weighted instead of being hard-removed. The resulting tensor is of the same size 512×7×7, however, with improved lesion-focused feature maps and suppressed background noise.

3.8 Global Context Modeling using ViTs

There is a need to understand the global context of the image when it comes to DR detection due to disease progression. The presence of an abnormal lesion is related to its location on the retina, whether it is located near the macula or the optic disc. This helps in understanding the severity of the disease. To address this requirement, transformers are employed in the proposed CALF-Net. The strengths of CNN are complemented by integrating the vision transformers, helping the system to obtain not only the features but their global representations as well.

3.8.1 Motivation for Transformer-Based Modeling

The long-range dependencies are not modeled by convolutional neural networks because of the presence of local receptive fields. They are bound to obtain the hierarchical information, which is locality bias, although the effective receptive field is expanded in deeper CNNs. This results in the inability of the CNN to obtain the relationship present between the distant retinal regions. Additionally, the severity of the disease is not determined by the presence of lesions only, but also by how they are distributed or arranged on the retina.

3.8.2 Tokenization and Feature Flattening

The output tensor, which is obtained from the spatial gate, is of the dimensions 7x7x512. However, the transformers work with tokens. So the first step is to create a flattened feature map consisting of a series of tokens. Each spatial location will be converted into a token, so for 7x7, we will get a total of 49 tokens with 512 dimensions. The resultant matrix will be of the size 49x512. The channel-wise information related to the features is retained while the processing of each spatial location is performed through transformers, as shown in figure 3.10.

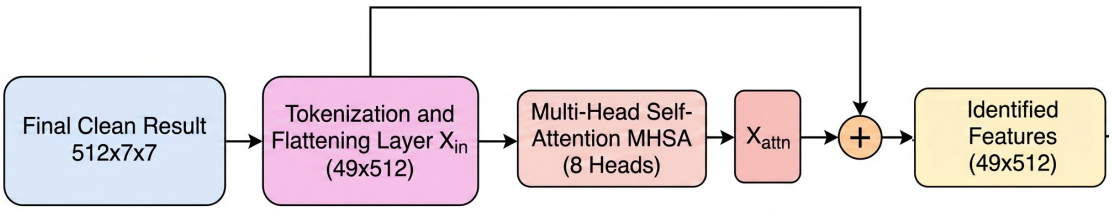

FIGURE 3.10: Tokenization of Features

3.8.3 Multi-Head Self-Attention

The long-range dependencies and global context are captured with the help of multi-head self-attention (MHSA), which is a core part of vision transformers.

The total number of MHSA used in the proposed method is eight, so that simultaneously a large amount of the retina should be covered and dedicated attention should be given. Different weights for importance are applied to different regions of the network to make sure that the lesions get extra attention compared to the background. The output is a series of tokens that are attention-enhanced, containing both the local feature information and their dependencies across the retina.

3.8.4 Residual Connection and Normalization

A residual connection is established inside the transformers by pixel-wise addition of the input tokens to the attention-enhanced tokens. In this manner, the original local feature information is not lost and is added to the attention context. For ensuring stability, the result of the residual connection is normalized for the subsequent layers.

3.8.5 Feed-Forward Network with Linear Layers

Now, the above-mentioned tokens are processed through the feed-forward network containing

i. Linear Layer: $512 \rightarrow 2048$

ii. ReLU Activation

iii. Dropout = 0.4

iv. Linear Layer ($2048 \rightarrow 512$)

Each token is handled side by side simultaneously in the FFN, and the final output is 49x512. The purpose of the first linear layer, by expanding the size to 2048, is to understand the non-linear, complex patterns that are associated with fundus images. The Feed-Forward Network (FFN) in the Transformer consists of dense layers with excessive parameters, and is therefore very likely to overfit, particularly

when learning retinal features from small medical datasets. A dropout rate of 0.4 is used for regularization by randomly switching off 40% of neurons during training. This helps the network learn more general feature representations, as opposed to relying on particular neurons. A higher dropout rate is used in the FFN because this component is responsible for deep feature transformation, where neurons tend to co-adapt, and regularization is needed to avoid over-reliance on features. The dimensions are again compressed back to 512 from where they underpass the global aggregation, as shown in figure 3.11. .

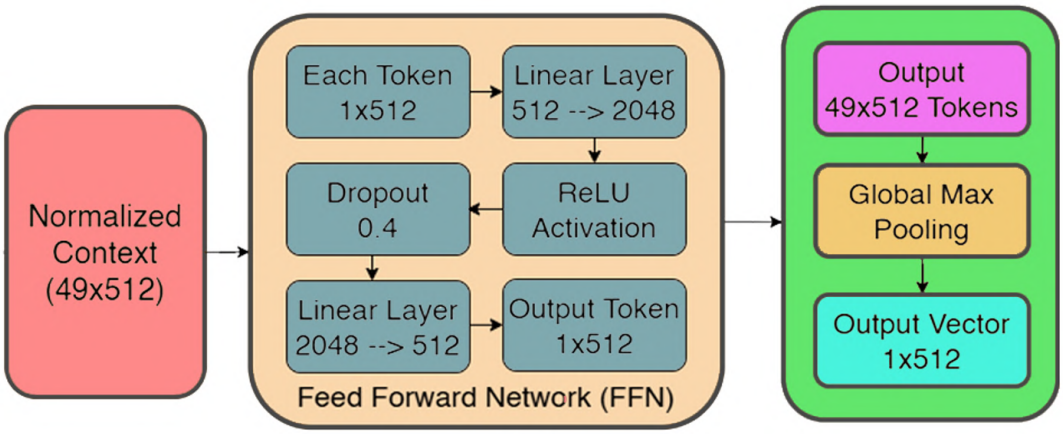

FIGURE 3.11: Feed Forward Network

3.8.6 Global Max Pooling

Global max pooling is applied to the matrix of 49x512, resulting in a single 1D vector containing dimensions of 1x512. The selection of global max pooling over global average pooling is due to its ability to emphasize the most salient features for preserving the signals that are associated with lesions. The 1D vector represents the entire retina.

3.8.7 MLP Classification Head

The final classification is performed with the help of a Multi-layer Perceptron (MLP) containing a single 1D vector 1x512 as an input. The components contain

i. Linear Layer: $512 \rightarrow 256$

ii. ReLU Activation

iii. Dropout =0.5

iv. Final Linear Layer: $256 \rightarrow 2$

The dimensions are even compressed to 256 before applying ReLU activation. The MLP classification head uses a higher dropout rate (0.5) because it is the final layer for the diabetic retinopathy classification.

Given that this layer translates high-level features to the class labels, it is more vulnerable to overfitting on the training data.

Increasing the regularization in this layer ensures that the decision-making does not overfit to the training data and can effectively identify diabetic retinopathy in novel retinal images.

Hence we use a dropout rate of 50% to increase the stochasticity and make the final classification more robust. The final classification layer is reduced to 2 because of the binary classification of DR vs Non DR cases.

The MLP is responsible for assigning the predicted class scores to the global feature representation. This is the last step of the pipeline for the DR classification, as shown in figure 3.12.

3.9 Training of CALF-Net

The section covers the configurations that are used in training, the loss function, and strategies for regularization. The training configurations are selected keeping in mind the objective of improving the sensitivity of the system.

3.9.1 Training Configuration

The training configurations are listed below.

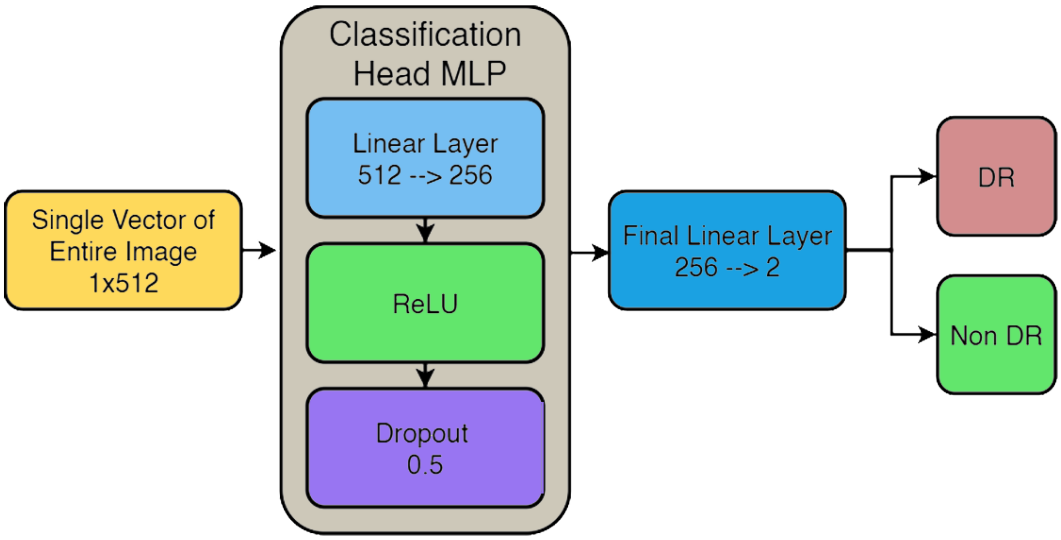

FIGURE 3.12: MLP Classification

i. Optimizer: AdamW optimizer is incorporated for its known adaptive learning rates and regularization in weight decays.

ii. Learning Rate and Scheduling: The initial learning rate was set to 2×10^{-5}. To escape the local minima and to stabilize the convergence, cosine annealing with warm restarts is used as a scheduler.

iii. Batch Size: The batch size was set to 16.

iv. Number of Epochs: The maximum number of epochs was 15, with the early stopping criteria to remove the overfitting from training.

3.9.2 Loss Function Definition

Cross-entropy loss was used in the training of CALF-Net. To make sure that the results are not affected by any class imbalance that still exists after augmentation, weighted cross-entropy loss was employed. A weight of 4.5 was given to the DR class, whereas 1.2 weight was assigned to the Non-DR class. The overconfident predictions were prevented by applying a smoothing factor of 0.05, which helps to improve generalization as well.

The proposed system was created as an end-to-end pipeline that started with preprocessing contrast-enhanced fundus images before moving on to local feature extraction with a ResNet-18 backbone. In order to minimize non-pathological noise that was enhanced during augmentation and allow the network to concentrate on clinically important lesion locations, a spatial gating mechanism was implemented. In order to capture long-range spatial dependencies and global retinal context, the enhanced feature maps were then converted into token sequences and processed using a Vision Transformer. Lastly, an MLP-based classification head and global feature aggregation were used to get the final predictions. Taking into consideration the detailed approach of CALF-Net, it creates a systematic and expandable basis for reliable DR identification even in the presence of other retinal disorders.

Chapter 4

Results and Discussions

The experimental evaluation of the proposed CALF-Net is presented in this chapter. The main goals are to evaluate CALF-Net's overall performance, compare it to current baseline techniques, and confirm particular design decisions using a thorough ablation study.

In order to guarantee robustness and generalizability, experiments are carried out on a publicly accessible dataset, RFMiD. The evaluation provides a thorough assessment of the model's efficacy in identifying and categorizing retinal lesions by taking into account a number of quantitative parameters, including accuracy, sensitivity, specificity, and AUC-ROC.

The results shown here demonstrate the advantages of the suggested model and provide information on how it compares to other approaches.

4.1 Dataset Description

The choice of selecting the dataset in order to train and test the proposed CALF-Net, along with its attributes, is discussed in this section. To replicate the clinical settings where different ocular diseases can coexist with each other, leads to the motivation behind selecting the dataset.

4.1.1 Dataset Source

Retinal Fundus Multi-Disease Image Dataset (RFMiD) is used in this research to evaluate the performance of the proposed model[8]. RFMiD is considered to be a benchmark dataset for the evaluation of multiple retinal diseases. Thus, it makes it a suitable candidate for mimicking the real-world clinical settings.

RFMiD was released during the international challenge of Retinal Image Analysis. It consisted of fundus images obtained from a variety of clinical sources.

Standard fundus cameras were used to perform the image acquisition by creating varying imaging conditions in order to obtain variations found in the anatomy of the retina, making the dataset more diverse. Differences in the illumination, along with pathological manifestations, were seen in RFMiD.

4.1.2 Dataset Composition and Size

The dataset composition and size can be summed up in the following points.

i. Diversity: The dataset contains 46 different retinal pathologies that appear in different sizes, color and shapes. These include Glaucoma, Age-related Macular Degeneration (AMD/ARMD), Cataracts, Macular Edema, DR, and many others.

   Some of the diseases include Branch Retinal Vein Occlusion (BRVO) and Central Retinal Vein Occlusion (CRVO), resembles with DR visually, making the dataset more challenging.

ii. Total Images: A total of 3200 images are present in RFMiD

iii. Training Set: The dataset comes with a predefined splitting of Training, Validation, and Testing folders. 1920 images are present in the training set.

iv. Validation Set: During training of the dataset, the model is validated after each epoch on unseen data to test the training procedure. For this purpose, 640 images are dedicated to the validation folder.

v. Testing Set: The testing set also contains a total of 640 images, upon which the actual performance of the model is evaluated.

For the problem of binary classification, all the DR-related images are grouped and labeled as 1, and all the other fundus images that do not contain DR and possibly can contain other ocular diseases, or can be healthy, are labeled as 0.

4.1.3 Image Characteristics

High-resolution images in the RGB color space are captured and presented in the RFMiD dataset. The high quality of the images allows the various anatomical structures to appear clearly, providing a detailed analysis of macula, optic disc, blood vessels, and other parts of the retina.

As images are acquired by various devices, a difference in the field of view, along with resolution fluctuations, is seen.

The images are stored in Portable Network Graphics (png) format that allows lossless compression of the images in order to retain their resolutions.

The images are resized according to the input format of CALF-Net so that compatibility can be achieved without losing the diagnostically relevant features.

4.2 Experimental Setup

In order to train and evaluate the model efficiently, the experiments were conducted on a computing environment with high performance. A NVIDIA GPU was used for the acceleration required for training deep learning models with a multi-core processor and RAM with enough capacity to handle a large dataset.

Python based implementation was conducted by using PyTorch libraries for data visualization and handling. CUDA was used for GPU acceleration.

The predefined division of training, validation, and testing folders as mentioned in section 4.1.2 is used for model development and to ensure unbiased evaluation. Extensive preprocessing techniques were applied to address the issues related to fundus images.

Augmentation techniques were also applied to mitigate the impact of class imbalance present in the dataset and to produce unbiased results without overfitting. The details of the experimental setup are summed up in the following Table 4.1

TABLE 4.1: Complete Experimental Setup: Hardware and Software Configuration

Category	**Specification**
Hardware Configuration	
GPU	Kaggle Notebook Environment: NVIDIA Tesla T4 (Dual GPU support - T4 x2)
GPU Memory	16 GB VRAM per T4 GPU (shared cloud allocation)
CPU	Intel Xeon Processor (Cloud-based Kaggle CPU instance)
RAM	Approximately 32 GB System RAM (Kaggle runtime dependent)
Storage	Temporary SSD-based Kaggle cloud storage
GPU Acceleration	CUDA-parallel processing with cuDNN support
Software Configuration	
Operating Env	Kaggle Notebooks (Cloud-based Linux environment)
Language	Python 3.10+
DL Framework	PyTorch (latest Kaggle-supported version)
CUDA Version	CUDA 11.x (preconfigured Kaggle environment)
DL Libraries	torchvision, torch, numpy, pandas, matplotlib, scikit-learn
Image Libs	OpenCV, PIL (Python Imaging Library)
Visualization	Matplotlib, Seaborn
Training Configuration	
Optimizer	AdamW Optimizer
Loss Function	Weighted Cross-Entropy Loss
Batch Size	Based on GPU memory constraints (typically 16–32)
Epochs	Trained until convergence with early stopping
Acceleration	CUDA + cuDNN GPU acceleration enabled

4.3 Evaluation Metrics

CALF-Net performance was evaluated using multiple evaluation metrics, which depend on a number of classification outcomes. These are True Positives, True Negatives, False Positives, and False Negatives.

True Positives (TN) are defined as the disease images, and the model also predicted them likewise. On the other hand, True Negatives (TN) are those in which the model predicts the healthy cases to be non-disease cases.

False Positives (FP) mean the model predicts a healthy case to be a disease case, and False Negatives (FN) are those particular disease cases that the model classifies as healthy cases or non-disease cases. All four classification outcomes are mathematically represented below.

$$\text{True Positives (TP)} = \sum_{i=1}^{N} \mathbb{I}(y_i = 1 \wedge \hat{y}_i = 1) \tag{4.1}$$

$$\text{True Negatives (TN)} = \sum_{i=1}^{N} \mathbb{I}(y_i = 0 \wedge \hat{y}_i = 0) \tag{4.2}$$

$$\text{False Positives (FP)} = \sum_{i=1}^{N} \mathbb{I}(y_i = 0 \wedge \hat{y}_i = 1) \tag{4.3}$$

$$\text{False Negatives (FN)} = \sum_{i=1}^{N} \mathbb{I}(y_i = 1 \wedge \hat{y}_i = 0) \tag{4.4}$$

4.3.1 Accuracy

The correctness of the model is evaluated by measuring the proportion of correctly classified samples among the total number of samples.

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{4.5}$$

where TP: True Positives TN: True Negatives FP: False Positives FN: False Negatives

4.3.2 Sensitivity

The recall is measured by the number of correctly identified disease cases with respect to the samples that are not classified as disease. It is calculated by the following formula.

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

4.3.3 Specificity

The ability of the model to evaluate the non-DR classes accurately is measured by the specificity calculated with the help of the following formula.

$$\text{Specificity} = \frac{TN}{TN + FP} \tag{4.7}$$

4.3.4 Area Under the ROC Curve

The model's discriminative ability to distinguish between the DR classes and non-DR classes is identified by using Area Under the Receiver Operating Characteristics Curve (AUC-ROC). It explains the probability of the higher rank given to the model's ability to detect DR compared to Non-DR samples. A strong separability among classes is noted with the higher values of AUC.

4.3.5 Image Quality Assessment

The preprocessing of the fundus images is usually evaluated by two known metrics, named as Peak Signal to Noise Ratio (PSNR) and Structural Similarity Index

(SSIM)[37]. PSNR is calculated using the following formula.

$$\text{PSNR} = 10 \log_{10} \left(\frac{MAX_I^2}{\text{MSE}} \right) \tag{4.8}$$

MAX_I represents the highest pixel intensity, usually 255 for a grayscale image. The mean square error is defined by

$$\text{MSE} = \frac{1}{HW} \sum_{i=1}^{H} \sum_{j=1}^{W} \left(I(i,j) - \hat{I}(i,j) \right)^2 \tag{4.9}$$

I and $\hat{I}$ represent the original image vs the processed image. H and W represent the height and width of the image.

By comparing the original image with the preprocessed image, the ratio of the maximum power signal to the noise signal is calculated using the above-mentioned formula 4.8. The results are measured in decibels (dB). Higher values of PSNR indicate a low presence of noise. The values usually lower than 30 dB depict the presence of a high noise ratio, and the preprocessing technique resulting in lower PSNR is not considered to be good[1].

Another metric for evaluating the preprocessing technique is the Structural Similarity Index (SSIM), which is commonly used to assess the similarity of the structure between the processed image and the original image[1]. It is calculated using the following formula.

$$\text{SSIM}(x,y) = \frac{(2\mu_x\mu_y + C_1)(2\sigma_{xy} + C_2)}{(\mu_x^2 + \mu_y^2 + C_1)(\sigma_x^2 + \sigma_y^2 + C_2)} \tag{4.10}$$

where x and y are the original image vs the processed image. The mean intensities of both images are represented by μ_x and μ_y. The variances of images are denoted by σ_x^2 and σ_x^2. Covariance σ_{xy} is between x and y, whereas C_1 and C_2 are the constants employed for stabilized division.

Unlike PSNR, which is responsible for measuring the absolute pixel difference, SSIM measures the degradation of the image after applying preprocessing steps

on it. The values of SSIM usually lie between 0 and 1. 0 means no structural similarity at all, and 1 means both images are identical to each other.

4.4 Impact of Class Imbalance and Data Balancing

To investigate the impact of class imbalance on the performance, six pretrained CNN models including AlexNet, ResNet-152, VggNet-s, VggNet-16, VggNet-19 and GoogleNet are chosen from the baseline methodology [1]. We have performed experiments under the following three settings:

i. Models trained using the original, imbalanced RFMiD dataset,

ii. Models trained with geometric augmentation and the proposed preprocessing, and

iii. Models trained using photometric augmentation along with the proposed preprocessing pipeline.

Here we compare these settings to assess the impact of class imbalance on the classification performance metrics of sensitivity and specificity.

The RFMiD dataset is class-imbalanced, as it has a large number of non-disease samples compared to DR samples. This class imbalance influences the learning process of deep learning models, leading to a bias towards the dominant class.

4.4.1 Baseline Performance on Imbalanced RFMiD Dataset

As shown in the baseline results in Table 4.2, the performance of various CNN architectures on the original (unbalanced) RFMiD dataset is evaluated. It is noted that model learning is affected by the imbalance in the training data, with

a greater proportion of non-disease samples than DR samples. As a result, the majority of the models suffer from bias towards the majority class, leading to relatively high specificity across all model architectures. Although the specificity

TABLE 4.2: Baseline Results on RFMiD Dataset (Imbalanced)

Model	**Accuracy**	**Sensitivity**	**Specificity**	**AUC**
AlexNet	84.38%	83.06%	84.69%	0.9157
ResNet-152	91.56%	85.48%	93.02%	0.9655
VGGNet-S	90.94%	82.26%	93.02%	0.9497
VGGNet-16	86.56%	91.13%	85.47%	0.9491
VGGNet-19	90.62%	81.45%	92.83%	0.9531
GoogleNet	91.25%	90.32%	91.47%	0.9716

values are relatively high, suggesting that the models are correctly identifying non-disease cases, the sensitivity values of some of the models are relatively low. This implies that the number of false negatives has increased, meaning some of the diseased cases are not being identified. This is especially important in clinical diagnostic applications where the failure to detect DR can lead to delayed or lost treatment opportunities with a consequent loss of vision.

GoogleNet is the most balanced model among the baseline models, with an accuracy of 91.25% and the highest value for the area under the curve AUC (0.9716), and reasonably balanced sensitivity (90.32%) and specificity (91.47%). This suggests that while GoogleNet outperforms other models, its performance is still affected by the class imbalance in the data.

4.4.2 Effect of Geometric Augmentation and Proposed Preprocessing

The results achieved after performing geometric augmentation and the proposed preprocessing steps are shown in Table 4.3. It can be seen that the addition of geometric transformations along with improved preprocessing, helps the model learn more discriminative features for DR.

TABLE 4.3: Results using Geometric Augmentation and Proposed Preprocessing

Model	**Accuracy**	**Sensitivity**	**Specificity**	**AUC**
AlexNet	89.53%	86.29%	90.31%	0.9379
ResNet-152	91.72%	91.94%	91.67%	0.9522
VGGNet-S	90.00%	83.06%	91.67%	0.9430
VGGNet-16	91.72%	88.71%	92.44%	0.9605
VGGNet-19	88.59%	95.16%	87.02%	0.9596
GoogleNet	89.06%	91.94%	88.37%	0.9583

In particular, there is a significant increase in sensitivity for all models, implying that the models are better at detecting DR cases.

This suggests that geometric augmentation helps to reduce class-imbalance bias in the model by introducing more variation of retinal lesions, which improves the model's generalization ability.

For example, VGGNet-19 has the highest sensitivity of 95.16%, suggesting that it is highly effective at detecting positive DR cases after geometric augmentation. Likewise, ResNet-152 and GoogleNet also exhibit an increase in sensitivity but with little change in specificity. This suggests that the boost in true positives is not accompanied by a high false positive rate.

Ultimately, geometric augmentation allows the model to better learn the distinctive features of a lesion, hence boosting recall-oriented performance, which is important for medical screening tasks where false negatives are unacceptable.

4.4.3 Effect of Photometric Augmentation and Proposed Preprocessing

Table 4.4 shows the results of photometric augmentation with the proposed preprocessing pipeline. While geometric augmentation changes spatial features, photometric augmentation changes photometric features such as brightness, contrast

and illumination.

These changes improve the conspicuity of subtle retinal lesions, which can be challenging to discern under different lighting conditions. Our findings suggest

TABLE 4.4: Results using Photometric Augmentation and Proposed Preprocessing

Model	**Accuracy**	**Sensitivity**	**Specificity**	**AUC**
AlexNet	85.62%	83.87%	86.05%	0.9155
ResNet-152	86.56%	89.52%	85.85%	0.9395
VGGNet-S	89.06%	87.90%	89.34%	0.9475
VGGNet-16	86.25%	91.94%	84.88%	0.9548
VGGNet-19	91.41%	90.32%	91.67%	0.9646
GoogleNet	90.62%	90.32%	90.70%	0.9548

that photometric augmentation has a more balanced performance in terms of sensitivity and specificity for most models.

The models ResNet-152 and VGGNet-19, for instance, exhibit robust and consistent performance with relatively high sensitivity and specificity.

ResNet-152 shows a sensitivity of 89.52% and a specificity of 85.85%, suggesting enhanced lesion detection while keeping a good balance between true and false positives.

Meanwhile, VGGNet-19 achieves a balanced performance with an accuracy of 91.41%, sensitivity of 90.32%, and specificity of 91.67%, and a high AUC of 0.9646, which indicates strong discrimination ability.

In summary, photometric augmentation enhances the stability of the models by making contrast-based lesions more visible, which results in more consistent classification performance across various models.

But, the gains in sensitivity are relatively limited when compared with geometric augmentation, albeit it plays an important role in the performance stability.

4.5 CALF-Net: Proposed Pipeline Results

The proposed CALF-Net model results were evaluated on the RFMiD dataset to evaluate the performance in a multi-disease environment.

Table 4.5 compares the accuracy, sensitivity, specificity, and AUC with the existing literature methodologies on the RFMiD dataset including CNN based and hybrid approaches.

The results offer a detailed performance comparison of the proposed CALF-Net versus the existing CNN-based and hybrid models on RFMiD dataset. The results are also depicted visually in Figure 4.1 that shows the performance of all models across different evaluation metrics.

The proposed CALF-Net has an overall accuracy of 91.88%, which is on par with the state of the art and matches the highest performing baseline models.

Although some CNN-based approaches and hybrid models achieve comparable accuracy, accuracy is not the ideal metric to evaluate the diagnostic ability of models in medical image classification, especially in the presence of class imbalance.

A more important metric for the detection of DR is sensitivity, which measures the model's capacity to detect positive cases. Our CALF-Net model demonstrates a much higher sensitivity of 96.77% than other models.

By comparison, state-of-the-art CNN-based models, such as the Adaptive Deep CNN, and hybrid models, are not able to achieve this level of sensitivity.

This significant gain shows that CALF-Net successfully decreases the number of false negatives, thereby minimizing the number of missed cases.

In regard to specificity, the proposed model presents 90.89%, which is in line with existing approaches, although slightly lower than some CNN-based approaches presenting higher specificity.

TABLE 4.5: Comparative analysis of Proposed CALF-Net with existing CNN and Hybrid approaches

Model	**Acc (%)**	**Sen (%)**	**Spe (%)**	**AUC**
Adaptive Deep CNN [1]	91.56	85.48	93.02	0.9655
Multi-label Deep Learning Framework [2]	90.00	87.14	91.00	0.9542
CNN–Transformer Hybrid [30]	92.10	90.20	91.80	0.9680
CNN–Transformer RFMiD Model 3DECNN [31]	90.85	89.70	90.10	0.9615
ViT–CNN Hybrid [32]	88.40	86.00	89.10	0.9520
Proposed CALF-Net	**91.88**	**96.77**	**90.89**	**0.9750**

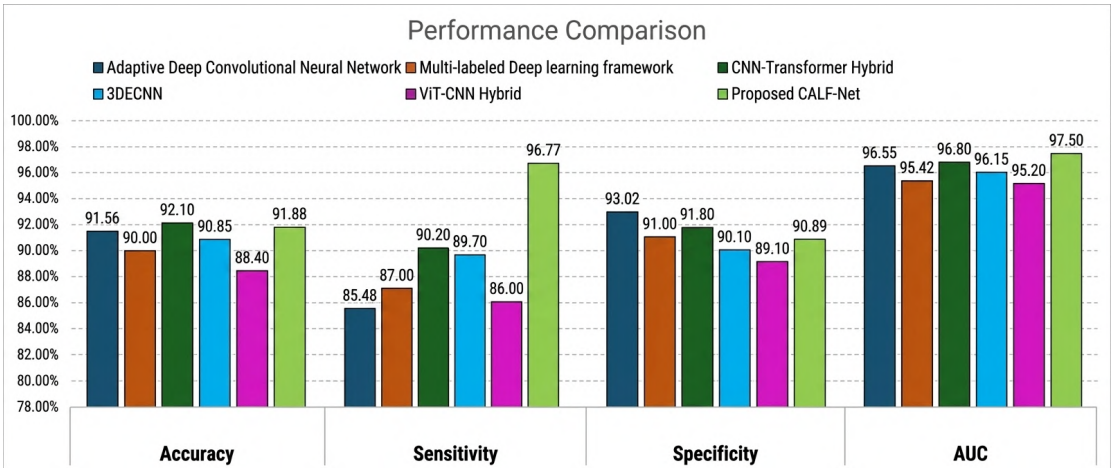

FIGURE 4.1: Performance Comparison of Proposed CALF-Net compared to existing approaches

This suggests that the proposed model exhibits a reasonable trade-off between sensitivity and specificity. In particular, the gain in sensitivity is accompanied by a slight loss in specificity (false positives), which is less costly in a medical diagnosis than false negatives [18].

This sensitivity vs. specificity trade-off is evident in this model. CALF-Net aims at high sensitivity, which means that most DR cases are detected.

This trade-off is in line with the key goal of this study, which is to boost sensitivity and reduce false negatives in the presence of similar retinal diseases.

The experimental results verify that the proposed strategy achieves this goal and is therefore more fit for clinical screening.

In addition, the proposed model has the highest AUC of 0.9750, suggesting that it is more capable of discriminating between diseased and non-diseased classes. This shows that CALF-Net retains high classification performance while enhancing key diagnostic measures.

To gain a deeper understanding of the model's performance, the confusion matrix is shown in Figure 4.2. The confusion matrix offers a comprehensive overview of true positives, true negatives, false positives and false negatives.

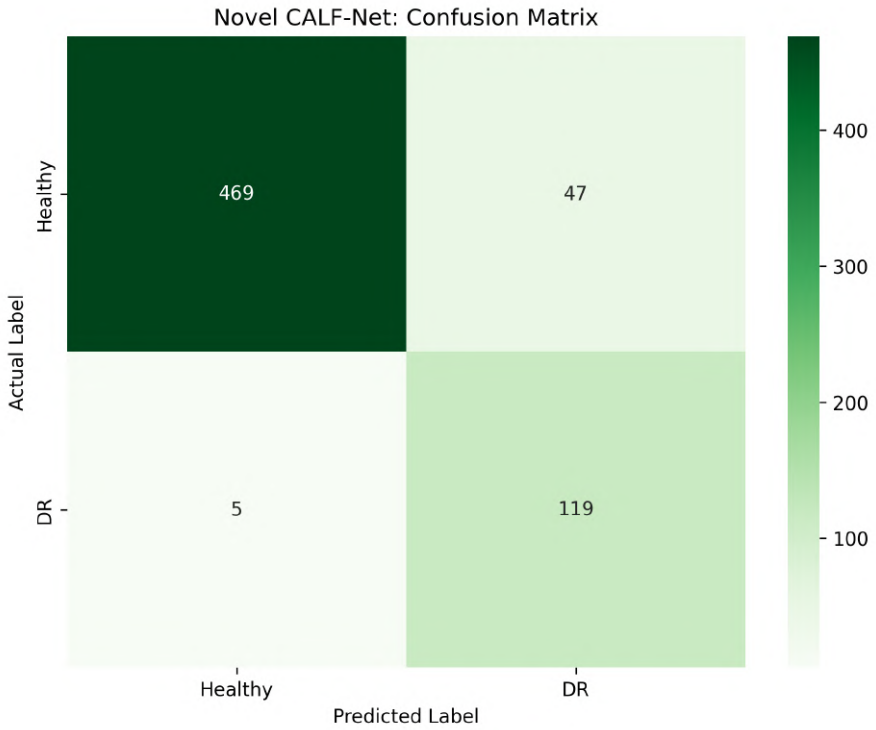

FIGURE 4.2: Confusion Matrix of Proposed CALF-Net

It can be seen that there are a relatively low number of false negatives (around 4.03%), suggesting that the model seldom misses a diagnosis of disease.

This highlights the success of the proposed technique to prevent the risk of missed diagnosis in DR detection.

The effectiveness of CALF-Net can be explained by its hybrid nature, which combines CNN to learn local features of the lesion, and transformers to learn global features of the retina.

Moreover, the proposed preprocessing helps to make the lesion more visible and the model can effectively discriminate between DR and other retinal diseases with similar appearance.

In conclusion, the findings indicate that CALF-Net offers a clinically feasible and effective approach for DR detection by striking a balance between accuracy and specificity while significantly increasing sensitivity and decreasing false negatives, which was the aim of this study.

The proposed CALF-Net is also compared with baseline 6 CNN models and the results outline how well the suggested CALF-Net performed on the RFMiD dataset.

The suggested CALF-Net performs better than all current models, especially in terms of sensitivity (96.77%) and AUC (0.9750), suggesting that it is more capable of accurately identifying cases of DR.

This performance shows how well the CALF-Net design captures retinal features on a local and global level while retaining competitive specificity.

4.6 Computational Cost Analysis

The efficiency of the proposed CALF-Net is compared with the baseline CNN models in terms of number of model parameters, training time, memory requirements (GPU memory) and the overall model complexity, as shown in Table 4.7. This exploration sheds light on the relationship between computational efficiency and diagnostic accuracy.

It can be seen that the earlier CNN models (AlexNet and GoogleNet) have a relatively low computational cost owing to their simpler and well-optimized architecture. On the other hand, deep CNN architectures like VGG-16, VGG-19 and

ResNet-152 have higher computational cost with increased memory and longer training time due to increased depth and number of parameters.

TABLE 4.7: Comparative analysis of Computational Resource Usage between Baseline CNN Models and Proposed CALF-Net

Model	**Params (M)**	**Time (epoch)**	**Mem (GB)**	**Model Complexity**
AlexNet	60M	8–10 min	2.5	Low complexity; shallow architecture with limited feature depth.
ResNet-152	60M	10–14 min	5.0	Deep residual network with skip connections and high demand.
VGG-16	138M	12–16 min	6.5	Very deep CNN architecture with large parameter size.
VGG-19	144M	14–18 min	7.0	Highest parameter complexity among VGG variants.
GoogleNet	7M	9–12 min	3.5	Efficient inception-based architecture with reduced parameters.
Proposed CALF-Net	**75M**	**15–20 min**	**7.5–8.0 GB**	**Hybrid CNN–Transformer with attention-based learning.**

The proposed CALF-Net has relatively higher computational requirements with a training time of around 15-20 minutes per epoch and higher GPU utilization.

This is largely because of its hybrid design, which combines convolutional neural networks for local feature learning with attention mechanisms of transformer-based networks for capturing global contextual information. The addition of attention blocks and extra pre-processing steps further adds to the computational burden.

It is worth noting that the training times mentioned are only rough estimates based on experience from experimentation in a Kaggle environment with an NVIDIA

Tesla T4 GPU, where, for example, training 15 epochs of a model took about 2-3 hours depending on the complexity of the model and the availability of resources.

Although this comes at the cost of increased computational cost, the improvement in performance offered by CALF-Net, especially in sensitivity and reduction in false negatives, is an acceptable compromise.

For medical imaging tasks like the detection of DR, sensitivity (avoiding false negatives) is more important than computational efficiency. Thus, the proposed model strikes a balance between computational cost and performance, and can be successfully applied to real-world clinical decision-making systems.

4.7 Image Quality Assessment using PSNR and SSIM

Image quality metrics were calculated for the proposed preprocessing pipeline to evaluate its efficacy. The evaluation's main goal is to determine whether contrast enhancement and noise reduction can be accomplished without compromising anatomical structures, which is crucial for accurate DR detection.

The table 4.9 shows the comparison of baseline preprocessing with the proposed preprocessing.

The proposed preprocessing pipeline clearly improved the image quality, as shown by the quantitative results in the above table.

More efficient noise suppression and less distortion are indicated by the PSNR value, which rises from 28.72 dB for the baseline method to 31.26 dB.

Comparably, the SSIM score increases dramatically from 0.72 to 0.88, indicating improved perceptual and structural information preservation in the retinal images.

TABLE 4.9: PSNR and SSIM comparison between existing and proposed preprocessing methods

Preprocessing Method	**PSNR (dB)**	**SSIM**
Adaptive Deep Convolutional Preprocessing [1]	28.72	0.72
Proposed Preprocessing	**31.26**	**0.88**

These improvements demonstrate that the proposed preprocessing technique preserves important anatomical characteristics like blood vessels and lesion borders while improving contrast and reducing noise.

Because it enables more dependable feature extraction and helps to improve classification performance seen in later experiments, this structural preservation is especially important for detecting DR accurately.

The PSNR and SSIM analysis confirms the proposed preprocessing pipeline's efficacy and defends its incorporation into the CALF-Net architecture.

4.8 Ablation Study

An ablation study was carried out to assess the effects of various design decisions made within the CALF-Net architecture, with a primary focus on the positioning of the Transformer after the CNN backbone's blocks.

Finding the optimal trade-off between feature representation, model generalization, and classification performance on the RFMiD dataset was the aim of this work.

Several experiments were conducted, which are given below.

i. Transformer placement after 1st Conv layer: The transformers were tested after the first convolutional layer and the pooling layer of ResNet-18, to understand what happens if the high-level features are fed to the transformer.

ii. Transformer placement after each ResNet-18 Residual Block: The placement of the transformer was tested after each residual block of the ResNet-18 architecture.

    By doing so, the transformer was tested by learning global context from a range of feature hierarchies, showcasing the contextual information that is associated with the intermediate feature maps.

iii. Transformer after global average pooling: The transformer in this experiment was placed after the global average pooling of ResNet-18 before the Fully Connected (FC) layer.

    It yields the impact of converting the image feature maps to a 1D vector and provides comprehensive image-level representations to the transformers.

iv. Standalone Transformers: As the existing literature revealed the increased performance of transformers as compared to CNNs on bigger datasets like EyePACS, the results were replicated on RFMiD dataset with standalone transformers to reveal their capacity in a smaller multi-disease dataset.

    Along with this, it was conducted to analyze how the integration of CNNs with transformers impacts the performance.

The above configurations were tested under the evaluation metrics of accuracy, sensitivity, and specificity as shown in table 4.11. The ablation study results clearly show the impact of Transformer location within the ResNet-18 backbone. With an accuracy of 74.69% and specificity of 73.45%, the Transformer model performs relatively poor when it is included right after the first convolutional layer.

This suggests that very early feature maps have low discriminative power because they lack the semantic richness necessary for efficient attention modelling. When the Transformer is positioned after the first residual block, there is a significant gain in accuracy (85.16%) and sensitivity (86.29%), indicating that mid-level characteristics offer more significant contextual information for attention mechanisms.

Long-range dependencies between retinal structures can be efficiently modeled by the Transformer due to this configuration's appropriate blend of semantic generalization and spatial detail.

Placing the Transformer after the third residual block yields the best accuracy of any individual placement, with 93.75% accuracy and 94.96% specificity. However, this configuration's lower sensitivity (88.71%) suggests a tendency for more false negatives, which is undesirable when it comes to the diagnosis of DR.

TABLE 4.11: Results of Ablation Study

Model Design	**Accuracy (%)**	**Sensitivity (%)**	**Specificity (%)**
ResNet-18: 1st Convolution Layer + Transformer	74.69	79.84	73.45
ResNet-18: 1st Residual Block + Transformer	85.16	86.29	84.88
ResNet-18: 2nd Residual Block + Transformer	90.00	87.09	89.34
ResNet-18: 3rd Residual Block + Transformer	93.75	88.71	94.96
Transformers after the Global Average Pooling Layer	54.06	76.61	48.64
Standalone Transformers	89.22	90.97	87.60
Proposed CALF-Net	**91.88**	**96.77**	**90.89**

Transformers perform substantially low when applied after the global average pooling layer; accuracy drops to 54.06% and specificity to 48.64%.

This demonstrates that focusing only on global features misses fine-grained spatial correlations and localized lesion patterns that are essential for analyzing retinal diseases.

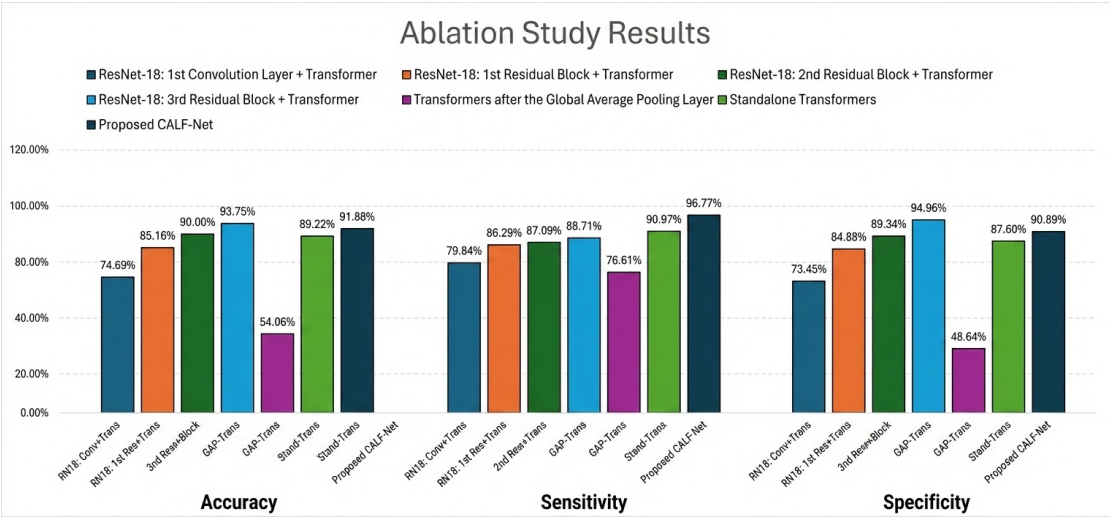

FIGURE 4.3: Results of Ablation Study

With 89.22% accuracy and 90.97% sensitivity, the results from standalone performers were quite competitive. However, it still falls short of hybrid CNN–Transformer configurations, emphasizing the significance of Convolutional feature extraction before attention modeling.

The accuracy of 91.88%, sensitivity of 96.77%, and specificity of 90.89% are the best balanced and clinically meaningful results obtained by the suggested CALF-Net architecture.

Although CALF-Net's accuracy is slightly reduced from that of the third residual block configuration, its sensitivity greatly exceeds all other designs, which is consistent with the study's main goal of lowering false negatives. Additionally, this enhancement is made without significantly sacrificing specificity, guaranteeing accurate differentiation between cases that are ill and those that are healthy.

The results of the ablation study support the final CALF-Net architectural design, showing that careful integration of Transformers at various feature levels permits efficient contextual learning while maintaining local lesion details.

These results demonstrate that the suggested pipeline provides the best possible balance between sensitivity, specificity, and accuracy, which makes it especially appropriate for the identification of DR in a multi-disease environment.

Using the RFMiD dataset, this chapter provided a thorough experimental assessment of the suggested CALF-Net for DR detection. According to the results, CALF-Net continuously outperforms existing techniques in a number of evaluation parameters, such as accuracy, sensitivity, and AUC-ROC. Specifically, the suggested approach significantly increased sensitivity, successfully decreasing false negative cases, a crucial prerequisite for early disease diagnosis in clinical screening situations.

The design decisions of CALF-Net were confirmed through extensive experiments comprising an ablation study and a preprocessing quality assessment using PSNR and SSIM. The suggested pipeline's discriminative quality was further validated using the confusion matrix and ROC analysis. The results validate CALF-Net as an efficient system for automated study of DR.

Chapter 5

Conclusion and Future Work

5.1 Conclusion

This study represents a hybrid model that integrates the lesion aware preprocessing with features extractions using CNN and attention mechanisms of vision transformers. The main goal of this study was to improve the detection of DR by reducing the false negatives and increasing the sensitivity in the presence of other retinal pathologies that shares the visually similar lesion patterns. This leads to the need of effective lesion extraction of DR from the fundus images so that they can be distinguished from the similar retinal pathologies.

The first objective of this study was to implement a lesion aware preprocessing technique to specifically enhance the lesions of the DR. For this purpose, an extensive preprocessing technique was adopted including the adaptive gamma correction followed by the contrast enhancement using histogram equalization and dilation of difference of Gaussian to further enhance the lesions. The proposed preprocessing was able to successfully highlight the lesions of DR without compromising the quality of image and without introducing irrelevant noise. The 0.8 SSIM values depicts that the structural similarity was maintained whereas the PSNR of 31.26 dB presents that the noise was reduced compared to the baseline preprocessing. Therefore the first objective of the study was achieved.

The second objective of this study was to develop an automated system for DR in a multi-disease environment by focusing on reducing the false negatives. In order to increase sensitivity and reliability, the study concentrated on integrating local features obtained from the CNN backbone of ResNet-18 and global contextual information using transformers. A spatial attention gate was in-cooperated to further suppress the noise and enhance the DR lesions by assigning weights based on the loss calculation during back propogation. Large weights were assigned to those features that were contributing in decreasing the loss and small weights were assigned to the features that were increasing the loss. This enables the models to suppress the noise and to enhance the lesions even further.

The results of the experiment conducted on a multi-disease dataset RFMiD reveal that the proposed CALF-Net outperforms the baseline architectures by showcasing improvement in accuracy, sensitivity (reducing the number of false negatives), and AUC (Area under the ROC). The model achieves an accuracy of 91.88%, sensitivity of 96.77%, specificity of 90.89%, and AUC of 0.9750. The false negatives were significantly reduced only showing 4.03% cases of DR that were misclassified. The comparison of the CALF-Net with the baseline six CNN models also reveals that the suggested model outperforms in terms of sensitivity. Therefore CALF-Net was able to improve the sensitivity while detecting DR in the presence of other retinal pathologies that are visually similar. The detailed ablation study reveals the optimal placement of the transformer after the 4th residual block of ResNet-18 architecture.

To conclude, the proposed CALF-Net architecture confirms that the integration of lesion aware preprocessing with CNN feature extraction, spatial attention gate and transformer multi-head self attention improves the detection of DR significantly. The study reveals that lesion aware preprocessing and attention mechanisms followed by spatial attention gate can results in more reliable and meaningful results in terms of clinical predictions. Therefore, the proposed CALF-Net provides application in real-world health-care scenarios where it could be deployed for effective detection of DR.

5.2 Future Work

The proposed CALF-Net framework can be improved in a number of ways in future research. Expanding the dataset is a crucial step in improving generalization across various populations by testing the model on bigger and more varied retinal imaging datasets. Furthermore, investigating different Transformer architectures, like hybrid attention processes or Swin-Transformers, may improve feature modeling and capture intricate interactions in retinal images. Diagnostic robustness and accuracy may also be improved by using multi-modal data, such as OCT scans or patient metadata. Clinical edge-device applications could be made possible by optimizing CALF-Net for real-time inference by lightweight designs, model compression, or pruning.

Enhancing explainability is still a crucial area, and attention mapping and learnt feature visualization may help increase therapeutic acceptance and trust. Performance can also be further improved by methodological improvements like as automatic hyperparameter optimization, dynamic feature fusion algorithms, and sophisticated augmentation approaches. Lastly, future studies can concentrate on improving specificity, which is essential for trustworthy clinical screening since it guarantees that the system not only efficiently detects DR but also reduces false positives.

Bibliography

[1] R. Abbasi, F. Amin, A. Alabrah, G. S. Choi, S. Khan, M. B. Bin Heyat, M. S. Iqbal, and H. Chen, "Diabetic retinopathy detection using adaptive deep convolutional neural networks on fundus images," *Scientific Reports*, vol. 15, no. 1, p. 24647, 2025.

[2] R. Kumar, V. Kohli, R. K. Singh, and R. K. Ratnesh, "Multi-label deep learning framework for early detection of diabetic retinopathy diseases," in *2025 International Conference on Next Generation Information System Engineering (NGISE)*, vol. 1, pp. 1–7, IEEE, 2025.

[3] M. H. Ashraf, M. N. Mehmood, M. Ahmed, D. Hussain, J. Khan, Y. Jung, M. Zakariah, and D. M. AlSekait, "Hird-net: An explainable cnn-based framework with attention mechanism for diabetic retinopathy diagnosis using clahe-d-dog enhanced fundus images," *Life*, vol. 15, no. 9, p. 1411, 2025.

[4] C. Suedumrong, S. Phongmoo, T. Akarajaka, and K. Leksakul, "Diabetic retinopathy detection using convolutional neural networks with background removal, and data augmentation," *Applied Sciences*, vol. 14, no. 19, p. 8823, 2024.

[5] I. Y. Abushawish, S. Modak, E. Abdel-Raheem, S. A. Mahmoud, and A. J. Hussain, "Deep learning in automatic diabetic retinopathy detection and grading systems: a comprehensive survey and comparison of methods," *IEEE Access*, vol. 12, pp. 84785–84802, 2024.

[6] Z. Khan, F. G. Khan, A. Khan, Z. U. Rehman, S. Shah, S. Qummar, F. Ali, and S. Pack, "Diabetic retinopathy detection using vgg-nin a deep learning architecture," *IEEE Access*, vol. 9, pp. 61408–61416, 2021.

[7] F. J. M. Shamrat, R. Shakil, B. Akter, M. Z. Ahmed, K. Ahmed, F. M. Bui, M. A. Moni, *et al.*, "An advanced deep neural network for fundus image analysis and enhancing diabetic retinopathy detection," *Healthcare Analytics*, vol. 5, p. 100303, 2024.

[8] S. Pachade, P. Porwal, D. Thulkar, M. Kokare, G. Deshmukh, V. Sahasrabuddhe, L. Giancardo, G. Quellec, and F. Mériaudeau, "Retinal fundus multi-disease image dataset (rfmid): A dataset for multi-disease detection research," *Data*, vol. 6, no. 2, p. 14, 2021.

[9] C.-L. Lin and K.-C. Wu, "Development of revised resnet-50 for diabetic retinopathy detection," *BMC bioinformatics*, vol. 24, no. 1, p. 157, 2023.

[10] D. Das, S. K. Biswas, and S. Bandyopadhyay, "Detection of diabetic retinopathy using convolutional neural networks for feature extraction and classification (drfec)," *Multimedia tools and applications*, vol. 82, no. 19, pp. 29943–30001, 2023.

[11] S. Das, A. Lasker, M. Ghosh, S. M. Obaidullah, and K. Roy, "A deep learning-based approach for detecting diabetic retinopathy in retina images," in *Internet of things-based machine learning in healthcare*, pp. 85–95, Chapman and Hall/CRC, 2024.

[12] T. Karkera, C. Adak, S. Chattopadhyay, and M. Saqib, "Detecting severity of diabetic retinopathy from fundus images: A transformer network-based review," *Neurocomputing*, vol. 597, p. 127991, 2024.

[13] N. J. Mohan, R. Murugan, T. Goel, and P. Roy, "Vit-dr: Vision transformers in diabetic retinopathy grading using fundus images," in *2022 IEEE 10th region 10 humanitarian technology conference (R10-HTC)*, pp. 167–172, IEEE, 2022.

[14] Z. Gu, Y. Li, Z. Wang, J. Kan, J. Shu, and Q. Wang, "Classification of diabetic retinopathy severity in fundus images using the vision transformer and residual attention," *Computational Intelligence and Neuroscience*, vol. 2023, no. 1, p. 1305583, 2023.

[15] J. Wu, R. Hu, Z. Xiao, J. Chen, and J. Liu, "Vision transformer-based recognition of diabetic retinopathy grade," *Medical Physics*, vol. 48, no. 12, pp. 7850–7863, 2021.

[16] R. Bala, A. Sharma, and N. Goel, "Ctnet: convolutional transformer network for diabetic retinopathy classification," *Neural Computing and Applications*, vol. 36, no. 9, pp. 4787–4809, 2024.

[17] P. Sekar, R. Bhoopalan, N. Nagaprasad, T. R. Mamo, S. Dhanabal, and R. Krishnaraj, "D-tnet: a hybrid dense net-transformer model for robust diabetic retinopathy detection," *Scientific Reports*, vol. 15, no. 1, p. 39594, 2025.

[18] c. m. darapaneni, b. s. babu, s. satyanarayana, l. madupu, v. sujay, n. janardhan, g. kishore, k. k. kaveti, and r. c. s. sairam, "hybrid vision transformer-cnn framework for automated retinal disease detection from fundus images," *journal of theoretical and applied information technology*, vol. 103, no. 24, 2025.

[19] J. Fan, N. Xiao, Y. Zhang, R. Zhai, and Y. Chu, "A hybrid model merging convolutional neural network and differential vision transformer for diabetic retinopathy identification," *Biomedical Signal Processing and Control*, vol. 115, p. 109435, 2026.

[20] A. Ikram and A. Imran, "Resvit fusionnet model: An explainable ai-driven approach for automated grading of diabetic retinopathy in retinal images," *Computers in Biology and Medicine*, vol. 186, p. 109656, 2025.

[21] G. Alwakid, W. Gouda, and M. Humayun, "Enhancement of diabetic retinopathy prognostication using deep learning, clahe, and esrgan," *Diagnostics*, vol. 13, no. 14, p. 2375, 2023.

[22] A. Jabbar, H. B. Liaqat, A. Akram, M. U. Sana, I. D. Azpíroz, I. D. L. T. Diez, and I. Ashraf, "A lesion-based diabetic retinopathy detection through hybrid deep learning model," *IEEE Access*, vol. 12, pp. 40019–40036, 2024.

[23] S. Llam, "Understanding the layers of convolutional neural networks (cnns)," *Medium*, 2020. Blog post.

[24] 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. 12116–12128, 2021.

[25] V. Awasthi, N. Awasthi, H. Kumar, S. Singh, P. P. Singh, P. Dixit, and R. Agarwal, "Vit-hho: Optimized vision transformer for diabetic retinopathy detection using harris hawk optimization," *MethodsX*, vol. 13, p. 103018, 2024.

[26] S. Alayón, J. Hernández, F. J. Fumero, J. F. Sigut, and T. Díaz-Alemán, "Comparison of the performance of convolutional neural networks and vision transformer-based systems for automated glaucoma detection with eye fundus images," *Applied Sciences*, vol. 13, no. 23, p. 12722, 2023.

[27] Z. Qu, M. Li, B. Yuan, and G. Mu, "A method of hybrid dilated and global convolution networks for pavement crack detection," *Multimedia Systems*, vol. 30, no. 4, p. 210, 2024.

[28] A. Zafar, M. Aamir, N. Mohd Nawi, A. Arshad, S. Riaz, A. Alruban, A. K. Dutta, and S. Almotairi, "A comparison of pooling methods for convolutional neural networks," *Applied Sciences*, vol. 12, no. 17, p. 8643, 2022.

[29] R. Nirthika, S. Manivannan, A. Ramanan, and R. Wang, "Pooling in convolutional neural networks for medical image analysis: a survey and an empirical study," *Neural Computing and Applications*, vol. 34, no. 7, pp. 5321–5347, 2022.

[30] A. Yamuna, D. Selvakumar, and R. Suresh, "Towards accurate diabetic retinal disease detection using advanced deep metric learning," *Biomedical Signal Processing and Control*, vol. 113, p. 109127, 2026.

[31] C. Kuruba and N. Gopalan, "3decnn: a novel method for segmentation of the diabetic retinopathy in retinal fundus images using 3d-edge cnn," *Neural Computing and Applications*, vol. 37, no. 21, pp. 16187–16201, 2025.

[32] D. Singh, S. Agarwal, and S. Mishra, "Retinal fundus multi-disease image classification using hybrid cnn-transformer-ensemble architectures," in *International Health Informatics Conference*, pp. 103–120, Springer, 2023.

[33] S. H. Abbood, H. N. A. Hamed, M. S. M. Rahim, A. Rehman, T. Saba, and S. A. Bahaj, "Hybrid retinal image enhancement algorithm for diabetic retinopathy diagnostic using deep learning model," *IEEE Access*, vol. 10, pp. 73079–73086, 2022.

[34] J. R. Balashunmugam, M. M. R. Sindha, A. Makkie, and U. M. Pandiyan, "Image enhancement techniques for fundus images-a review," in *AIP Conference Proceedings*, vol. 2857, p. 020072, AIP Publishing LLC, 2023.

[35] H. Naz and N. J. Ahuja, "A novel contrast enhancement technique for diabetic retinal image pre-processing and classification," *International Ophthalmology*, vol. 45, no. 1, p. 11, 2024.

[36] R. Alaguselvi and K. Murugan, "Quantitative analysis of fundus image enhancement in the detection of diabetic retinopathy using deep convolutional neural network," *IETE Journal of Research*, vol. 69, no. 9, pp. 6315–6325, 2023.

[37] E. Goceri, "Medical image data augmentation: techniques, comparisons and interpretations," *Artificial intelligence review*, vol. 56, no. 11, pp. 12561–12605, 2023.

[38] International Telecommunication Union, "Recommendation itu-r bt.601: Studio encoding parameters of digital television for standard 4:3 and wide-screen 16:9 aspect ratios," 2011.

[39] F. Ramzan, M. U. G. Khan, A. Rehmat, S. Iqbal, T. Saba, A. Rehman, and Z. Mehmood, "A deep learning approach for automated diagnosis and multi-class classification of alzheimer's disease stages using resting-state fmri and residual neural networks," *Journal of medical systems*, vol. 44, no. 2, p. 37, 2020.

[40] H.-C. Shin, N. A. Tenenholtz, J. K. Rogers, C. G. Schwarz, M. L. Senjem, J. L. Gunter, K. P. Andriole, and M. Michalski, "Medical image synthesis for data augmentation and anonymization using generative adversarial networks," in *International workshop on simulation and synthesis in medical imaging*, pp. 1–11, Springer, 2018.

[41] D. Muthusamy and P. Palani, "Deep learning model using classification for diabetic retinopathy detection: an overview," *Artificial Intelligence Review*, vol. 57, no. 7, p. 185, 2024.

[42] R. Vij and S. Arora, "A systematic review on diabetic retinopathy detection using deep learning techniques," *Archives of Computational Methods in Engineering*, vol. 30, no. 3, pp. 2211–2256, 2023.

[43] C. Wangweera and P. Zanini, "Comparison review of image classification techniques for early diagnosis of diabetic retinopathy," *Biomedical Physics & Engineering Express*, vol. 10, no. 6, p. 062001, 2024.

[44] C. Lam, D. Yi, M. Guo, and T. Lindsey, "Automated detection of diabetic retinopathy using deep learning," *AMIA summits on translational science proceedings*, vol. 2018, p. 147, 2018.

[45] D. C. R. Novitasari, F. Fatmawati, R. Hendradi, H. Rohayani, R. Nariswari, A. Arnita, M. I. Hadi, R. A. Saputra, and A. Primadewi, "Image fundus classification system for diabetic retinopathy stage detection using hybrid cnn-delm," *Big Data and Cognitive Computing*, vol. 6, no. 4, p. 146, 2022.

[46] K. D. K. Wardhani, S. Kasim, R. Hassan, R. Hidayat, and K. M. Sujon, "Deep learning for diabetic retinopathy detection: A review of multimodal data fusion approaches," 2025.

[47] C. Shorten and T. M. Khoshgoftaar, "A survey on image data augmentation for deep learning," *Journal of big data*, vol. 6, no. 1, pp. 1–48, 2019.

[48] A. Srinivas, T.-Y. Lin, N. Parmar, J. Shlens, P. Abbeel, and A. Vaswani, "Bottleneck transformers for visual recognition," in *Proceedings of the IEEE/CVF*

conference on computer vision and pattern recognition, pp. 16519–16529, 2021.

[49] J. Maurício, I. Domingues, and J. Bernardino, "Comparing vision transformers and convolutional neural networks for image classification: A literature review," *Applied Sciences*, vol. 13, no. 9, p. 5521, 2023.

[50] Y. Liu, G. Sun, Y. Qiu, L. Zhang, A. Chhatkuli, and L. Van Gool, "Transformer in convolutional neural networks," *arXiv preprint arXiv:2106.03180*, vol. 3, 2021.

[51] G. Hemalakshmi, M. Murugappan, M. Y. Sikkandar, S. S. Begum, and N. Prakash, "Automated retinal disease classification using hybrid transformer model (svit) using optical coherence tomography images," *Neural Computing and Applications*, vol. 36, no. 16, pp. 9171–9188, 2024.

[52] A. Khan, Z. Rauf, A. Sohail, A. R. Khan, H. Asif, A. Asif, and U. Farooq, "A survey of the vision transformers and their cnn-transformer based variants," *Artificial Intelligence Review*, vol. 56, no. Suppl 3, pp. 2917–2970, 2023.

[53] S.-H. Wang, P. Phillips, Y. Sui, B. Liu, M. Yang, and H. Cheng, "Classification of alzheimer's disease based on eight-layer convolutional neural network with leaky rectified linear unit and max pooling," *Journal of medical systems*, vol. 42, no. 5, p. 85, 2018.

[54] R. Paul, M. S.-u. Hassan, E. G. Moros, R. J. Gillies, L. O. Hall, and D. B. Goldgof, "Deep feature stability analysis using ct images of a physical phantom across scanner manufacturers, cartridges, pixel sizes, and slice thickness," *Tomography*, vol. 6, no. 2, p. 250, 2020.

[55] W. Nazih, A. O. Aseeri, O. Y. Atallah, and S. El-Sappagh, "Vision transformer model for predicting the severity of diabetic retinopathy in fundus photography-based retina images," *IEEE Access*, vol. 11, pp. 117546–117561, 2023.

[56] A. G. Persada, I. Ardiyanto, M. B. Sasongko, and H. A. Nugroho, "A review of cam-based visual explanation on diabetic retinopathy," *IEEE Access*, 2026.

[57] T.-E. Tan and T. Y. Wong, "Diabetic retinopathy: Looking forward to 2030," *Frontiers in Endocrinology*, vol. 13, p. 1077669, 2023.