Cross-Attention Fusion Vision Transformer for Explainable and Efficient Multi-Class Eye Disease Detection

by

Yasin Rahman
23341115
Md. Mozahedul Hoque
22101379
Moriyum Chowdhury
21301210
Mustakim Al Mahmud
24241323
Mahidul Islam Mahi
21301542

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

Department of Computer Science and Engineering
BRAC University
January 2026.

Declaration

It is hereby declared that

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

Student's Full Name & Signature:

Yasin Rahman
23341115

Md. Mozahedul Hoque
22101379

Mustakim Al Mahmud
24241323

Mahidul Islam Mahi
21301542

Moriyum Chowdhury
21301210

Approval

The thesis titled " Cross-Attention Fusion Vision Transformer for Explainable and Efficient Multi-Class Eye Disease Detection " submitted by

1. Yasin Rahman (23341115)
2. Md. Mozahedul Hoque (22101379)
3. Mustakim Al Mahmud (24241323)
4. Mahidul Islam Mahi (21301542)
5. Moriyum Chowdhury (21301210)

of Fall, 2025 has been accepted as satisfactory in partial fulfillment of the requirement for the degree of B.Sc. in Computer Science in 31 January, 2026 Year.

Examining Committee:

Supervisor:
(Member)

Annajiat Alim Rasel

Senior Lecturer
Department of Computer Science and Engineering
BRAC University

Co-Supervisor:
(Member)

Ahmed Mayeesha Reza Agomoni

Lecturer
Department of Computer Science and Engineering
BRAC University

Thesis Coordinator:
(Member)

Md. Golam Rabiul Alam, PhD

Professor
Department of Computer Science and Engineering
BRAC University

Head of Department:
(Chair)

Sadia Hamid Kazi, PhD

Chairperson and Associate Professor
Department of Computer Science and Engineering
BRAC University

Abstract

Early and accurate detection of retinal fundus diseases is critical for preventing irreversible vision loss and supporting effective clinical decision-making. Retinal fundus imaging is widely used for large-scale screening due to its non-invasive nature; however, the diversity and structural complexity of retinal pathologies pose significant challenges for automated analysis. Convolutional Neural Networks (CNNs) have been extensively employed for fundus image classification owing to their strong local feature extraction capabilities, however, their limited ability to model long-range contextual dependencies constraints performance in complex multi-class disease scenarios. Vision Transformers (ViTs), on the other hand, leverage self-attention mechanisms to capture global contextual information but often suffer from high computational costs and reduced effectiveness in limited-data medical imaging settings. The study proposed a lightweight hybrid Cross-Attention Fusion Vision Transformer architecture can be used to classify multi-class retinal diseases through fundus images. The suggested model combines CNN-based local feature extraction and transformer-based global contextual modelling with the cross-attention fusion mechanism, which allows both fine-grained pathological features and holistic retinal structure to interact and at the same time keep computational efficiency. The hybrid model is specifically designed with small-scale medical data in mind and uses attention-based interpretability, which helps to include explainable AI in the future to improve clinical transparency. Experimental evaluation on publicly available fundus datasets demonstrates that the proposed hybrid approach achieves a favorable balance between accuracy, robustness, and efficiency compared to standalone CNN and Vision Transformer models, highlighting its suitability for automated retinal disease screening applications.

Keywords: Retinal fundus imaging; Eye disease classification; Convolutional Neural Networks (CNN); Vision Transformer (ViT); Hybrid CNN–Transformer; Cross-attention fusion; Lightweight deep learning

Acknowledgement

Firstly, all praise to the Great Allah for whom our thesis have been completed without any major interruption.

Secondly, we are grateful to our thesis Supervisor Mr. Annajiat Alim Rasel sir for his kind support and advice in our work. He helped us whenever we needed help.

Thirdly, to our Co-supervisor Ahmed Mayeesha Reza Agomoni for her kind support and advice in our work. Her continuous support, guidance, and insightful feedback have greatly contributed to the depth and quality of our research.

And finally to our parents without their throughout support it may not be possible. With their kind support and prayer we are now on the verge of our graduation.

Table of Contents

List of Figures

Chapter 1

Introduction

1.1 Background

With the widespread adoption of digital devices such as computers and smartphones, prolonged screen exposure has become increasingly common in modern lifestyles. Extended use of electronic screens is associated with visual fatigue, dryness, and ocular discomfort, primarily due to reduced blinking frequency and inadequate ocular lubrication. In the long-term, these factors lead to diminished ocular health and poor quality of life, as well as to more demanding health care systems [4].

Fundus associate eye disease are one of the major causes of irreversible vision loss and blindness in the world. Specifically, glaucoma, cataract, AMD and diabetic retinopathy cause the significant share of the worldwide visual impairment and represent dangerous socioeconomic threats [2]. The conditions usually exhibit early pathologic features in fundus images, such as optic nerve head deformity, abnormality of vascular, and retinal lesions. However, initial symptoms are often asymptomatic, which leads to delayed diagnosis and treatment.

Fundus imaging is a non-invasive and objective examination, which can visualize retinal structures, morphology of optic nerves, and vascular patterns. This technique is a vital component in terms of mass screening and monitoring of diseases. Although, the interpretation of fundus images relies heavily on experienced ophthalmologist and the vast amount of the imaging data or information has made manually screening time-consuming and resource intensive [3].

Recent developments surrounding deep learning, specifically Convolutional Neural Networks (CNNs) and Vision Transformers (ViTs) show a high potential in the automation of fundus image analysis and glaucoma, cataract, AMD, and diabetic retinopathy diagnostic quality [5]. Yet, the black-box character of these models brings about several concerns with transparency and clinical trust, underscoring the importance of developing efficient and interpretable diagnostic frameworks for real-world deployment.

1.2 Research Objective

The principal objectives of this research are as follows:

1. To design and implement a hybrid Convolutional Neural Network (CNN) and Vision Transformer (ViT) architecture for multi-class eye disease classification using fundus images.

2. To investigate how combining CNN-based local feature extraction with transformer-based global contextual modeling improves the discrimination of visually similar retinal diseases such as glaucoma, cataract, AMD and diabetic retinopathy.

3. To conduct a comparative evaluation between CNN-only, ViT-only, and hybrid CNN–ViT models in terms of classification performance, robustness, and computational efficiency.

4. To develop a lightweight hybrid architecture that balances diagnostic accuracy with reduced parameter count and computational complexity, making it suitable for deployment in resource-constrained clinical settings.

5. To incorporate explainable artificial intelligence (XAI) techniques, including Grad-CAM for CNN components and attention rollout for Vision Transformer components, in order to provide visual interpretability of model predictions.

6. To analyze the strengths and limitations of hybrid CNN–Transformer architectures for fundus image analysis and identify directions for future improvements in automated ophthalmic diagnosis.

1.3 Problem Statement

Despite the widespread use of fundus imaging for the screening and diagnosis of ocular diseases, accurate and scalable interpretation of fundus images remains a significant challenge. Manual assessment by ophthalmologists is time-consuming, subject to inter-observer variability, and increasingly impractical due to the growing volume of retinal imaging data and limited clinical resources. Automated deep learning–based diagnostic systems have therefore gained considerable attention as a means to support large-scale screening and early disease detection.

Convolutional Neural Network (CNN) based models have proven to be very effective in fundus image analysis by effectively capturing both local spatial and lesion-level patterns. However, their locality bias limits their ability to model long-range relationships and global anatomical context, both of which are often critical to discriminate between visually similar retinal diseases. Conversely, Vision Transformer (ViT) models use self-attention mechanisms to capture global contextual relationship, but typically require large datasets and heavy computational resources, which make them unsuitable for deployment in resource-limited clinical settings.

Additionally, many existing deep-learning methods function as black-box models, and therefore, offers limited interpretability of diagnostic outputs. The lack of transparency is a hindrance to clinical trust and adoption, especially in safety critical medical applications. Although, hybrid CNN-Transformer architectures have already been suggested to combine local and global feature representations, many of the existing solutions rely on the heavy backbones, multi-stage pipelines or the

ensemble strategies that increase computational complexity and hinders real world application.

Therefore, an effective, lightweight and interpretable hybrid deep-learning framework that can effectively merge CNN-based local features extraction with transformer-based global contextual modeling to robustly classify a multi-class of fundus image is required. Such a framework shall be balanced between diagnostic performance, computational performance and explainability so that it can be practical to implement in automated ophthalmic screening systems.

1.4 Scopes and Challenges

However, there are significant challenges. Public datasets with high-quality and diverse images are limited which hinders model training and adaptation. Additionally, the computational expense of ViT may be impractical in clinics with limited resources, and ensuring model consistency across varying imaging conditions and disease severities is a significant challenge.

Chapter 2

Literature Review

2.1 Preliminaries

2.1.1 Fundus Imaging

Fundus imaging is a widely used, non-invasive ophthalmologic imaging technique. It captures high-resolution images of the interior parts of the eye such as the retina, optic disc, macula, and the retinal vascular structure. These anatomical structures provide significant visual biomarkers which aid to the diagnosis and longitudinal changes over a wide range of both ocular and systemic pathologies like glaucoma, cataract and diabetic retinopathy. Owing to its ability to provide objective, reproducible, and clinically interpretable visual information, fundus imaging has become a cornerstone of large-scale eye disease screening programs and an essential data source for automated computer-aided diagnostic systems [3].

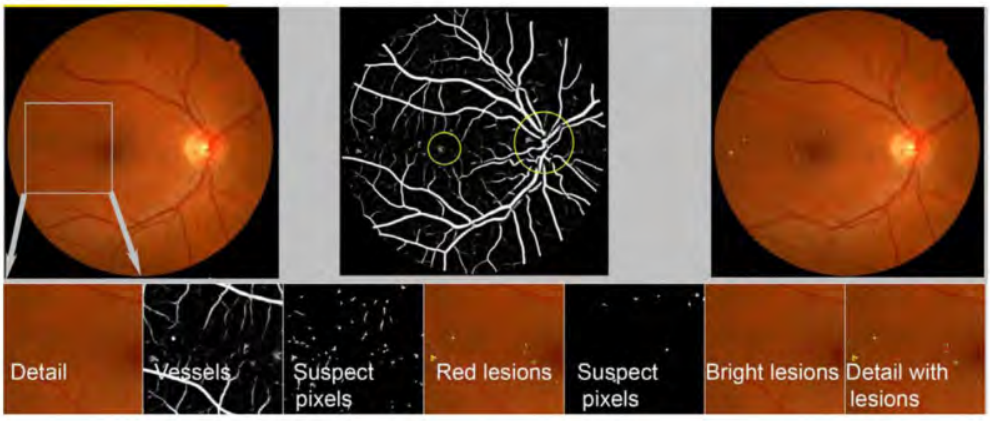


Figure 2.1: Typical steps necessary for analysis of fundus images, in this case for early diabetic retinopathy. Adapted from Abràmoff et al. [3].

2.1.2 Convolutional Neural Networks (CNNs)

Convolutional Neural Networks (CNNs) are deep learning architectures particularly designed to work with visual data such as images. It is commonly used in computer vision, image recognition and natural language processing (NLP). There are three

main components in CNN: the convolution layer, pooling layer and fully connected layer 2.2. In the convolution layer, CNN applies filters to input regions, to extract features while reducing parameters through weight sharing. The ReLU activation function is used to capture complex patterns and address vanishing gradient problems. The pooling layer mainly reduces the spatial dimension by decreasing the amount of parameter; it also decreases computational complexity within the network, hence down sampling the feature maps. Max-pooling is the most common pooling method used. Finally, for result prediction, the fully connected layer flattens the data by connecting every neuron to the next layer. Additionally, strides and padding are also used to make the architecture more efficient and preserve boundary information for spatial and flexible processing [25].

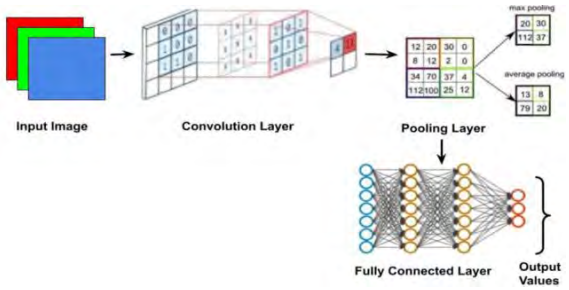


Figure 2.2: Architecture of A Typical Deep Convolutional Neural Network (reprinted from [42], CC BY-SA 4.0)

2.1.3 Vision Transformer (ViT)

The Vision Transformer (ViT) [30] utilizes transformers, initially developed for natural language processing, to analyze images. Vision Transformer generates a grid of square patches from an image. According to Kameswari et al. [30], each patch is converted to a single vector by concatenating the channels of all its pixels and then linearly projecting it to the chosen input dimension. A learnable token is added to consolidate global features for tasks such as categorization. The patch embeddings are fed as input into a transformer encoder which includes alternating multi-head self-attention and feed-forward layers. Self-attention enables the Vision Transformer (ViT) to capture global interactions across patches, in contrast to the local filters utilized by Convolutional Neural Networks (CNNs). ViT discards convolutions, depending solely on attention mechanisms. In addition to classification, ViT is adaptable for tasks such as detection and segmentation. The scalability and global reasoning of transformers underscore their adaptability in vision, although with increased processing expenses. According to Dosovitskiy et al. [14], Vision Transformer matches or exceeds the state of the art on many image classification datasets, whilst being relatively cheap to pre-train.

2.1.4 Hybrid models of ViT-CNN

The two models Convolutional Neural Networks (CNNs) and Vision Transformers (ViTs) exhibit complementary strengths for medical image analysis. CNNs are inherently effective at extracting local spatial features making them ideal for identifying fine-grained retinal abnormalities, such as subtle vascular changes, texture

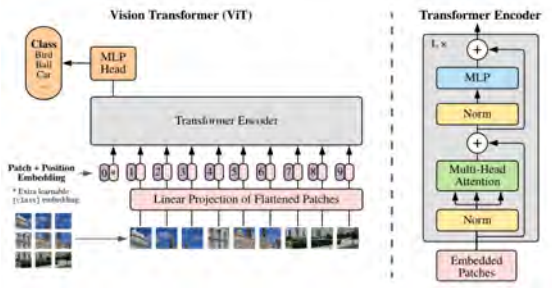


Figure 2.3: Overview of the Vision Transformer (ViT) architecture, reprinted from [14].

variations, and localized lesions. In contrast, Vision Transformers are designed to model long-range dependencies through self-attention mechanisms, enabling a global understanding of spatial relationships across the entire image. This is essential as the manifestations of the disease may depend not only on localized features.

A hybrid CNN–ViT architecture combines these strengths by allowing local feature representations learned by convolutional layers to be enriched with global contextual information through transformer-based attention mechanisms. This unified representation enables simultaneous analysis at both micro- and macro-levels, improving the model's ability to discriminate between visually similar disease classes. Such an approach is particularly well suited to fundus image classification, where pathological differences are often subtle and distributed across multiple spatial scales. By jointly leveraging local structural detail and global contextual awareness, hybrid architectures offer a balanced and robust framework for automated fundus-based eye disease detection, while addressing the inherent limitations of standalone CNN or transformer models.

2.1.5 Explainable Artificial Intelligence in Fundus Image Analysis

In medical image analysis in ophthalmology, the interpretability of a model is significant for clinical acceptance and trust. Explainable artificial intelligence (XAI) methods focus on giving either visual or quantitative explanations that point to areas of an image that a model has used to make a prediction, and therefore help a clinician determine whether a diagnostic decision-making process relies on features of medical interest.

In the case of the convolutional neural network (CNN)-based models, Gradient-weighted Class Activation Mapping (Grad-CAM) has also been one of the most popular XAI methods. Grad-CAM based on the use of gradient information of the final convolutional layers to produce class-discriminative heatmaps; these maps can be used to visualize the local retinal features, including vessels, lesions, or optic disc areas, or hemorrhages, that affect disease classification decision-making. [7]. This localization facility is especially useful in fundus image analysis, where fine spatial details are usually used to detect the disease.

Vision Transformers (ViTs), on the contrary, employ self-attention mechanisms itself, that model global contextual relationships across image patches. An attention-based explanation methods, e.g. attention rollout, combine the attention weights of transformer layers to detect which image regions contribute the most to the end prediction. [12]. Unlike CNN-based methods, attention rollout provides insight into how global contextual dependencies influence classification decisions, making it well suited for analyzing distributed retinal patterns.

By combining CNN-based Grad-CAM visualizations with transformer-based attention rollout, a complementary explanation framework can be achieved that captures both localized feature importance and global contextual reasoning. Such a dual-view interpretability strategy enhances transparency and supports more reliable analysis of automated fundus-based disease diagnosis systems.

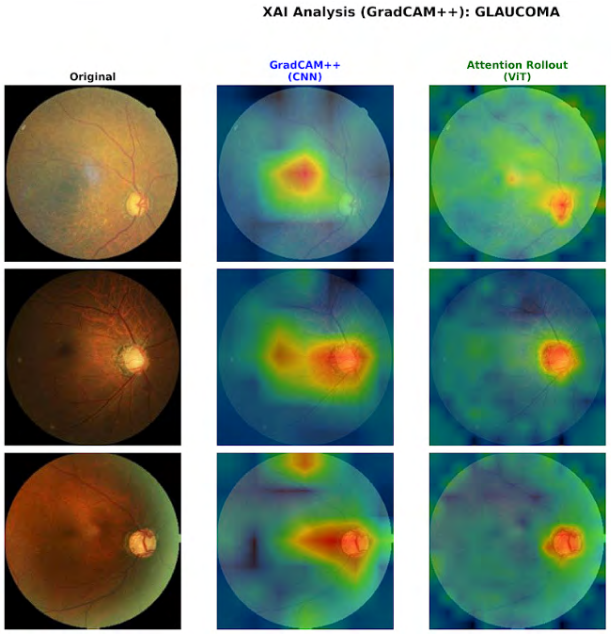


Figure 2.4: Visualization of explainable AI outputs for fundus image classification. Grad-CAM highlights discriminative local regions used by CNN-based models, while attention rollout illustrates global contextual focus learned by the Vision Transformer.

2.2 Research Methodology

A comprehensive review of existing literature was conducted to understand the evolving role of deep learning architectures in ophthalmic image classification . This review discusses how various model families deal with the inherent challenges of fundus based eye disease detection including detection of subtle lesion patterns, high interclass similarity, and variation of image quality. Particular attention was given to the comparative strengths and limitations of CNN and ViT as both architectures have been widely adopted in recent ophthalmological studies. The surveyed works were systematically categorized into three primary groups. The first group comprises studies that employ CNN based architectures for eye disease classification,

leveraging strong local feature extraction and hierarchical representation learning.

The second group includes studies that explore ViT based approaches which utilize self-attention mechanisms to capture global contextual relationships within retinal images. This structured categorization enabled a critical assessment of how each architecture contributes to performance,robustness and generalization. The insights derived from these studies provide the methodological foundation for examining hybrid CNN Transformer models which aim to combine the complementary advantages of both architectures. The reviewed studies within each category are discussed in the following sections.

While the third group particularly focused on ensemble models utilizing various CNN or Transformer backbones , it also highlighted the strengths and limitations of hybrid architectures and their unique implementation to tackle both local and global dependencies. Each of these models contributed into shaping our model and identify existing gaps in such architecture and ultimately improving them .

2.2.1 Convolutional Neural Network–Based Methods for Fundus Image Analysis

In this article [46], the authors propose RSG-Net, which is a custom convolutional neural network consisting of four convolutional layers, max-pooling, batch normalization, dropout, and fully connected layers. The model is trained to perform both four-stage (multi-class) and two-stage (binary) diabetic retinopathy grading on fundus images. The authors trained and evaluated the model on the Messidor-1 benchmark dataset, which is a collection of 1,200 color fundus images classified into four grades of diabetic retinopathy (0–3). Moreover, RSG-Net demonstrates strong performance, achieving a four-stage test accuracy of 99.36%, specificity of 99.79%, sensitivity of 99.41%, very low false positive and false negative rates, and a near-optimal AUC of 99.98%. The model performs better in four-stage classification than a number of state-of-the-art CNN-based, transfer learning, hybrid, and ensemble methods reported in previous studies. However, the study also recognizes its limitations, particularly evidence of overfitting, as the applicability of the model to other datasets and varying imaging conditions has not yet been validated.

Chakraborty et al.[38] proposed a transfer learning–based CNN model using DenseNet-121 aided by two attention mechanisms for the binary classification of glaucoma from fundus images. The authors trained their architecture on two standard datasets, namely RIM-ONE and ACRIMA, where features extracted from DenseNet-121 were further refined using the Convolutional Block Attention Module (CBAM) and the Channel Recalibration Module (CRM). In the RIM-ONE ablation study, DenseNet-121 demonstrated higher performance compared to other backbone architectures, including MobileNetV2, ResNet-50, and InceptionV3. The progressive addition of CBAM and CRM improved performance, achieving approximately 93.81% accuracy, 93.40% precision, 93.59% recall, and a 93.49% F1-score. However, the authors also noted the need to explore more lightweight models in future work to better address data-constrained environments.

The authors [44] presented a deep learning architecture, Fundus-DANet, which was applied to the OIA-ODIR multilabel dataset consisting of a total of 9,476 retinal fundus images. The model is based on a convolutional neural network backbone with dilated convolution layers to broaden the receptive field and capture long-range contextual information without introducing excessive additional parameters. Inception-ResNet-v2 serves as the core backbone of the model for deep feature extraction. A Dilated Convolutional Block Attention Module (DA-CBAM) is incorporated on top of this backbone to enable the capture of multiscale spatial features without reducing feature map resolution. This module extends the conventional CBAM by integrating dilated convolution along with channel and spatial attention mechanisms, allowing the network to focus on disease-relevant regions while suppressing unnecessary background information. The architecture is also designed for multilabel prediction, enabling the detection of multiple retinal diseases from a single fundus image.

This paper [40] suggests the use of Deep Convolutional Ensemble (DCE), which is an ensemble of five InceptionV3 CNN models pretrained on ImageNet and further trained on 43,055 fundus images collected from 12 datasets using bootstrap aggregation. The objective of the model is to distinguish between normal eyes, diabetic retinopathy (DR), glaucoma, and age-related macular degeneration (AMD). The proposed DCE achieved an accuracy of 79.2%, an F1-score of 79.9%, a positive predictive value (PPV) of 85.0%, sensitivity of 79.2%, and specificity of 93.1%, with particularly strong performance in detecting diabetic retinopathy. Notably, the ensemble successfully identified DR cases that were misclassified as normal by ophthalmologists. To ensure a single class prediction per image, the final decision was concluded through majority voting across a set of five networks. However, the study highlights the limitation connected to the use of publicly available data, where label noise may be present as it is based on lack of standardized clinical annotations.

The authors [54] introduce RetinaDNet, a two-branch artificial intelligence architecture for retinal disease classification that incorporates both fundus images and vessel segmentation maps generated using a U-Net model. These inputs are further enhanced through transfer learning using backbone networks such as ResNet50, and classification is performed using multiple classifiers, including Support Vector Machines (SVM), Multi-Layer Perceptrons (MLP), and Extreme Gradient Boosting (XGB), whose outputs are combined through soft voting. The framework applies preprocessing techniques such as Gaussian blur, image flipping, and Contrast Limited Adaptive Histogram Equalization (CLAHE). Vessel segmentation is optimized using a Dice–Binary Cross-Entropy loss, while cross-entropy loss is employed for classification. Model interpretability is achieved using Grad-CAM visualizations. The proposed RetinaDNet achieved 99.2% accuracy for diabetic retinopathy classification and 98.8% multi-class accuracy on a merged MuReD and RFMiD dataset consisting of 6,000 images. In addition, the model reported accuracies of 97.5% and 91% on the MESSIDOR2 and EyePACS-1 datasets, respectively. Performance improvements were attributed to the transfer learning strategy and the fusion of vessel segmentation maps with original fundus images, enabling the model to learn more discriminative retinal features and improve precision. The authors indicate that future work will focus on further optimization of the architecture and validation on

Table 2.1: Comparison of CNN-Based Fundus Image Analysis Methods

Study	**Dataset(s)**	**Methodology**	**Key Metrics**	**Main Limitations**	**Explainable AI(XAI)**
RSG-Net [46]	Messidor-1	Custom CNN with four convolutional layers for DR grading	Acc 99.36%, AUC 99.98% (4-class)	Evaluated on a single dataset; generalization not demonstrated	No
DADGC [38]	RIM-ONE, ACRIMA	DenseNet-121 with channel and spatial attention modules	Acc ∼93%, F1 ∼93%	Binary-only task; backbone is computationally heavy	No
Fundus-DANet [44]	OIA-ODIR	Inception-ResNet-v2 with dilated attention modules for multilabel prediction	Acc ∼93%	Complex architecture; no cross-dataset validation	No
DCE Ensemble [40]	12 public datasets	Ensemble of five InceptionV3 CNNs with majority voting	Acc 79.2%, F1 79.9%	High inference cost; affected by label noise	No
RetinaDNet [54]	MuReD, RFMiD, Messidor2	CNN features fused with vessel maps and ML classifiers	Acc 98.8% (multi-class)	Multi-stage pipeline increases complexity	Yes
U-Net + ResNet101 [41]	IDRiD, DDR, FGADR	CNN-based retinal lesion segmentation with ensembles	Dice up to 0.66	Segmentation-focused; not a classification model	No

additional, newly available datasets.

In this paper [41], Playout and Cheriet describe a tuned U-Net architecture with a ResNet-101 encoder and Squeeze-and-Excitation blocks, pretrained on ImageNet, as their baseline retinal lesion segmentation model. The baseline architecture is further improved through the incorporation of generalization strategies such as model ensembles, Stochastic Weight Averaging (SWA), and model soups to enhance cross-dataset performance. The evaluation was conducted across five fundus datasets, namely IDRiD (81 images), DDR (757 images), FGADR (1,842 images), Retinal Lesions (1,593 images), and Messidor-MAPLES-DR (200 images). The proposed baseline achieved competitive performance, reporting mean AUPR scores of 0.652 on IDRiD, 0.429 on DDR, and 0.571 on the Retinal Lesions dataset. In addition, cross-dataset Dice scores reached 0.661 on IDRiD test sets under both fine and coarse training combinations. However, the study also highlights several limitations, including reliance on pre-existing architectures rather than introducing novel designs, restriction of experiments to only five datasets, and reduced effectiveness of SWA-based model soups compared to classification tasks, particularly in scenarios with limited data and high label variance.

2.2.2 Vision Transformer–Based Approaches for Retinal Fundus Image Analysis

The model Retinal ViT [43] proposed a Vision Transformer architecture for the classification of multi label retinal disease. It addresses the limitation of single label classification as a single eye can have multiple diseases simultaneously. Through

patch embedding and self attention mechanism the model ensures global features are extracted explicitly and long-range lesion dependencies are taken in context which CNN struggles to identify. The model was evaluated on the ODIR-2019 dataset across eight disease categories including AMD, Glaucoma, Cataract, Diabetic Retinopathy etc. They used weighted binary cross-entropy loss and a sigmoid-based head for treating each label independently. The proposed non-convolutional approach showed better results compared to the baseline CNN models, i.e., ResNet, VGG, DenseNet, as the models achieved a higher Kappa score of 0.645 and an increased F1 scores. In spite of these findings, the model is constrained with the imbalance in the classes within the ODIR-2019 dataset and needs additional testing on more diverse datasets so that they can be more generalized.

The main purpose of the paper [37] is to design an end-to-end pipeline to enhance the transparency of automated glaucoma diagnosis. The proposed framework uses YOLOv11 object detector for localizing the optic disc followed by MaskFormer segmentation model with Swin Transformer backbone to accurately indicate both functions, optic disc and optic cup. The final diagnosis is based on the vertical cup-to-disc ratio (vCDR), which is a standard clinical biomarker, where a ratio higher than 0.5 is an indicator of the presence of glaucoma. The framework has shown better performance than the traditional CNN-based U-Net models, because of ViTs ability to capture global relationships across an entire image effectively. The model obtained segmentation Dice Similarity coefficient (DSC) of 93.83% and 93.71% for optic disc segmentation and segmentation of the optic cup, using the REFUGE and the ORIGA and G1020 datasets with a classification accuracy of 84.03%. A notable limitation includes the frame's sensitivity to image rotation; a slight rotation of 10 degrees can have a significant impact on the accuracy of vCDR estimation due to changes in vertical dimensions of the optic disc and optic cup.

The Control-guided and Augmented Vision Transformer (CA-VisT) [39], is proposed in order to cope with some of the drawbacks in the detection of early-stage glaucoma, such as the limited data availability, the heterogeneity of the features distribution as well as the similarity in structure inside the fundus images. The architecture uses a Vision Transformer backbone instead of traditional convolutional neural networks to enable better feature extraction and better interpretability through self-attention mechanisms. Furthermore, the framework integrated a Conditional Variational Generative Adversarial Network (CVGAN) to augment the dataset using the synthesis of diverse samples and a contour-guided module. This aids in robust extraction of salient edge information of optic disc and cup, thereby enhancing object localization. The study was evaluated on the SMDG dataset with 24,127 images after augmentation. Experimental results show that CA-ViT outperformed state-of-the-art models, and achieved 93% of accuracy, precision and recall, which is a significant improvement over ResNet50 (72.81%) and ViT16 (73.31%). The model is able to achieve this performance while being a lightweight model with parameters 41.1M. Nonetheless, the robustness of the models across all stages of the disease is limited by the class imbalance that exists in the data.

The Vision Transformer with Masked Autoencoders for large size Retinal Images (VLMRI) model [45] used a Vision transformer (ViT) architecture which is pre-

trained directly on large-scale retinal images using Masked Autoencoders (MAE) for supervised learning, to classify referable Diabetic Retinopathy (rDR). The model overcomes the limitations of CNNs and ViT on extracting global features and data limitations, proving that domain-specific pretraining with 100k unlabeled retinal images is superior to traditional ImageNet with 1M+ images. The experimental results on the APTOS dataset shows state-of-the art performance by achieving 93.42% accuracy, 0.9825 AUC and 0.9539 specificity, significantly outperforming VGG16 and InceptionV3 and AmInceptionV3 in clinical sensitivity, 0.9662. While the model overcomes the limitation of capturing global dependencies with fewer data, it still has its own drawbacks. The model includes increased computational costs due to high-resolution inputs and lacks lesion localization for better explainability.

STMF-DRNet [50] uses a multi-branch Swin Transformer V2 to formulate high-precision and fine-grained diabetic retinopathy (DR) grading of high-resolution fundus images. The model also incorporates a global feature branch where Attention-Based Object Localization (AOLM) and Patch-Processing modules are merged, allowing the process to capture large retinal structures and at the same time isolate sparse and subtle lesions that one branch models do not capture. The hierarchical structure of Swin Transformer, complemented with shifted-window attention and category-specific feature fusion, achieves higher results on various metrics with multiple datasets, such as DDR, EyePACS, and APTOS-2019. It suppresses the background noise and improves stage separations. However, the method has certain limitations due to its multi-branch architecture high dependence on preprocessing methods like Contrast-Limited Adaptive Histogram Equalization (CLAHE) and Gaussian blur to maintain discriminatory ability.

In the study [52], AI-based diagnostic system was developed to classify the severity of diabetic retinopathy by combining local and global image analysis. The hybrid model is based on a ResNet-18 backbone for hierarchical local features extraction with a Transformer encoder that considers long-range dependencies and global contextual correlations across the retina. Thereby overcoming the drawback of traditional CNNs that tend to poorly represent long-range correlations between the retinal regions that are distant and signify disease progression. The model also includes a Region-of-Interest (ROI) alignment, and a bounding-box mechanism, which focuses specifically on lesions like microaneurysms and hemorrhages,, ensuring they are accurately integrated with global structural data. The model was trained on the augmented APTOS 2019 dataset and validated on the IDRID dataset with a high accuracy of 94.28 percent and assured its generalizability through constant 95.23 percent accuracy on external data. While highly effective, the limitation of the study remains the computational complexity of the transformer mechanism, and therefore may become difficult to implement in real-world clinical environments with limited hardware reseources.

Table 2.2: Comparison of Vision Transformer Fundus Image Analysis Methods

Study	Dataset(s)	Methodology	Key Metrics	Main Limitations	Explainable AI(XAI)
Retinal ViT [43]	ODIR-2019	Pure ViT model with patch embedding and self-attention for multi-label classification	Kappa 0.645; improved F1 over CNN baselines	Severe class imbalance; evaluated on a single dataset	No
Explainable ViT-based Glaucoma Framework [37]	REFUGE, ORIGA, G1020	YOLO-based optic disc localization followed by MaskFormer (Swin Transformer) segmentation and vCDR-based diagnosis	DSC 93.83% (OD), 93.71% (OC); Acc 84.03%	Sensitive to image rotation; diagnosis depends on geometric ratios	Yes
CA-ViT [39]	SMDG (multi-source fundus datasets)	ViT backbone with CVGAN-based data augmentation and contour-guided feature fusion	Acc 93%, Precision 93%, Recall 93%	Performance affected by class imbalance; relies on synthetic data	Partial
ViT with Masked Autoencoders (VLMRI) [45]	APTOS-2019	ViT pretrained on large retinal images using MAE for referable DR classification	Acc 93.42%, AUC 0.985, Sens 0.973	High computational cost; no lesion-level localization	No
STMF-DRNet [50]	DDR, EyePACS, APTOS-2019	Multi-branch Swin Transformer with global and local attention-based lesion modeling	High accuracy and Kappa across datasets	Complex multi-branch design; heavy reliance on preprocessing	No
CNN–Transformer Hybrid DR Model [52]	APTOS-2019, IDRiD	ResNet-18 for local features fused with Transformer encoder for global context	Acc 94.28% (APTOS), 95.23% (IDRiD)	Transformer increases inference cost; hardware-intensive	No

2.2.3 Hybrid CNN–Transformer Architectures for Ophthalmic Image Analysis

Jibon et al. [21] proposed TL-MED, an ensemble-based transfer learning framework for multiclass eye disease classification using fundus images. The model integrates multiple pretrained CNN architectures, including VGG16, ResNet50, EfficientNetB3, DenseNet121, and NASNetMobile, where final predictions are obtained through softmax output averaging. The proposed ensemble achieved classification accuracies of up to 97% on the Eye Disease Classification (EDC) dataset, demonstrating the effectiveness of transfer learning and ensemble strategies for ophthalmic image analysis. Moreover, the use of several deep CNN backbones greatly adds to the increase in computational complexity, memory patterning, and inference latency. In addition, there is no explicit model of global contextual modeling, which restricts the range of the long-range dependencies of retinal structures captured by the framework. Following these restrictions, lightweight hybrid systems that combine local

feature extraction and global contextual representation in a single end-to-end system are required, as it is opposed to using computationally-intensive ensemble-based systems.

The cross dataset analysis shown by Pham et al., [31] trained their models on the ODIR dataset and tested them on the RFMiD dataset to analyze how the models perform in generalization when the data distribution changes. Performance was measured using per-class Area Under the ROC Curve (AUC), which is a standard and robust metric for multi-label medical image classification.The CoAtNet model achieved the highest overall performance, reporting an average AUC of 90.81%, outperforming standalone CNN-based models (ResNet50, DenseNet121) and pure transformer models (ViT and Swin Transformer). Specifically, CoAtNet obtained the best AUC scores for diabetic retinopathy, glaucoma, cataract, and myopia, while the Vision Transformer marginally outperformed CoAtNet only in the case of age-related macular degeneration (AMD). These results indicate that hybrid CNN–attention architectures exhibit superior cross-dataset generalization for most fundus disease categories.

A recent hybrid approach [34] integrating lightweight transformers with convolutional architectures was proposed through the MobileViT-Plus framework for diabetic retinopathy (DR) severity grading. The study introduced a parallel hybrid model combining a ResNet-101 backbone for local feature extraction with a MobileViT-Plus transformer branch to capture global contextual information from fundus images. The model was evaluated on a five-class DR dataset derived from the Kaggle DR repository, where class imbalance was addressed through extensive data augmentation. Performance was assessed using accuracy, F1-score, and AUC, with the hybrid architecture achieving an accuracy of 93.67%, an F1-score of 0.937, and a notably high AUC of 0.994, outperforming both the standalone CNN and transformer branches. These results demonstrate the effectiveness of combining local lesion-level features with global contextual reasoning for DR grading. However, the study is limited to a single disease and a single dataset, with no evaluation on other fundus pathologies such as glaucoma or cataract .

Rezaee and Farnami [33] proposed a CNN–Transformer fusion framework for diabetic retinopathy severity classification, demonstrating that the integration of convolutional feature extraction with transformer-based global contextual modeling improves generalization across multiple datasets, including APTOS-2019 and IDRiD. The model achieved over 94% classification accuracy, 91% sensitivity, a 92% F1-score, and an AUC of 0.9638, indicating strong discriminative capability.However, although the proposed approach performs highly, it does not put much focus on computational efficiency and heavily depends on the aggressive data augmentation tactics to achieve robust results. The resulting computational cost limits its applicability to use in resource-limited clinical settings. Moreover, the model can only be tested on diabetic retinopathy severity grading and not on the task of multi-disease fundus classification, which limits its applicability to wider ophthalmic screening.

Liu et al. [22] introduced CNN-transnet, a parallel dual-branch hybrid system where a CNN branch that includes the Xception, Coordinate Attention (CoordAtt) and

LSTM blocks derives local retina characteristics, and a transformer branch derives long-range dependencies on fundus images. The characteristics of the two branches are then combined to be finally classified. The accuracy of the results was reported as 94.72 percent on a set of four categories of eye disease classification, with an average accuracy of 80.68 percent in a set of seven categories. Nonetheless, the model is characterized by a great level of computational complexity, about 85 million parameters, and the price of computation is 2.84 x 106 FLOPs. The authors of their discussion admit that they do not see significant performance gains after providing the LSTM component, which implies that substituting it with a more effective option may be a valid solution. Also, although feature fusion is demonstrated to decrease the need in single-feature representations and provide higher feature diversity, the authors mention that feature fusion increases the risk of overfitting, especially in small-data settings.

The model [26] consists of EfficientNet and ResNet backbones and a transformer encoder to remove locality bias of CNNs and retain the powerful inductive priors of the CNNs, which leads to an overall balanced design, consisting of 32.75 million parameters. The accuracy (92.54), precision (92.60), recall (92.54), F1-score (92.55), specificity (97.52) and AUC (98.82) of the proposed model on retinal image classification benchmarks are all significant absolute improvements compared to baseline Vision Transformer models. Irrespective of these great results, the design relies on a number of heavyweight back bones and concatenating features plans that exponentially increase the number of parameters, memory traffic, and inference cost. This poses a serious problem in implementation in the constrained resource clinical environments even though it did better in classification.

Transformer [32] based feature extraction and feature fusion with ensemble learning are used in the multi-stage framework of MST-EDS approach with the aim of improving the performance of eye disease classification. The suggested pipeline combines several transformer backbones, such as ViT, DeiT, and Swin Transformer, and a post-processing step, Principal Component Analysis (PCA), to reduce the dimensions and a stacking scheme with a meta-learner to aggregate the predictions of individual models. MST-EDS reported an accuracy of 97.163, precision of 97.163, recall of 97.168 and a F1-score of 97.164 on the eye disease classification dataset, surpassing a number of independent transformer models and intermediate hybrid versions on the same dataset. The authors clearly recognize that the multi-stage ensemble design is incredibly complex with regard to computation, restricting its use in resource-bound and real-time systems, although it is performing well. Also, the framework fails to implement explainability mechanisms, which the authors consider one of the essential conditions of clinical trust, interpretability, and intensive implementation.

Badar et al. [27] proposed MSCAS-Net, a transformer-based hybrid architecture to fine-grained diabetic retinopathy classification, which is aimed at overcoming such issues as the small appearance of lesions, an imbalance of the classes and image quality variability. The proposed model uses Swin Transformer backbone to generate multi-scale feature representations as well as combines self-attention and cross-attention processes to effectively align and combine information at varying spatial resolutions.

It was tested on several retinal fundus datasets, such as APTOS, DDR, and IDRiD and the classification accuracy was 93.8, 89.8, and 86.7, respectively. Although the model is doing well, it has a computed cost of about 9.2 GFLOPs and has about 31 million parameters. Besides, it can be characterized by a narrow scope with only one type of disease and the lack of clear explainability functions, which restricts its generalizability and clinical readability in a multi-disease ophthalmic screening environment.

[28] In this paper they focused on the classification of retinal disease as automated based on the OCT images using a deep learning hybrid model known as Conv-ViT. The suggested model is composed of Inception-V3, ResNet-50 and a Vision Transformer (ViT) that learns to use local texture markers and global structural representations together. Outputs of all the three branches are integrated through feature-level fusion followed by a DNN classifier. The model is tested on a publicly available large dataset of OCT that has four categories namely: CNV, DME, DRUSEN and NORMAL. Conv-ViT is compared with single CNN models, transformer-only models, and the current ensemble solutions. Experimental findings demonstrate better accuracy, precision, recall, and F1-score, approximately the 94 percent compared to the methods. The research proves that the integration of CNNs and ViTs increases the strength and the ability to discriminate. However, the model is of high computational complexity, uses only one dataset and is not cross-dataset validated. Its applicability in real-world scenarios is also limited by the fact that multimodal clinical data were not available.

The complete framework of a multi-task deep learning that integrated convolutional neural networks, Vision Transformer, and Particle Swarm Optimization (PSO) [29] was suggested to streamline the accurate diagnosis of diabetic retinopathy and lesions localization .The framework was tested against the Diabetic Retinopathy Two-field Image Dataset (DRTiD) which consists of macula-centered and optic disc-centered fundus images with five severity ratings and lesion bounding boxes. The experimental findings had high performance and attained about 98.9% classification accuracy and the ablation experiments proved that cross-attention-based fusion worked significantly better than simple feature concatenation. Although such encouraging findings have been achieved, the suggested method is based on the inputs of the multi-view funduses and accurate lesionlevel annotations, which restricts its application in the single-view screening context that is widely adopted in large-scale clinical applications. Moreover, the combination of PSO and several task-specific heads is both more complex in architecture and demands a fine-tuning of hyperparameters. Although the parameter count of the model is roughly 34.2 million and its computational cost is only 22.6 GFLOPs per inference pass of two-view images achievable on a workstation it would need pruning or optimization to be used in mobile or embedded clinical systems.

Table 2.3: Comparison of Hybrid and Ensemble-Based Fundus Image Classification Methods

Study	**Dataset(s)**	**Methodology (Brief)**	**Key Metrics**	**Main Limitations**	**Explainable AI(XAI)**
TL-MED [21]	EDC (Normal, DR, Glaucoma, Cataract)	Ensemble of pre-trained CNNs with softmax averaging	Accuracy: 97%	High computational cost; no global contextual modeling	No
MST-EDS [32]	EDC (compiled fundus datasets)	Multi-stage ViT/DeiT/Swin features + PCA + stacking meta-learner ensemble	Accuracy: 97.163%; Precision: 97.163%; Recall: 97.168%; F1: 97.164%	Not end-to-end; very high compute; no explainability module	No
Multi-task CNN–ViT–PSO Framework [29]	DRTiD (two-field fundus images)	Multi-view CNN+ViT with cross-attention fusion; PSO optimizes fusion; multi-task (grading + localization)	Accuracy: ∼98.9%; IoU: ∼88.7%	Requires paired views + lesion annotations; PSO adds complexity; heavy for mobile deployment	No
MSCAS-Net [27]	APTOS, DDR, IDRiD	Swin Transformer with multi-scale attention fusion (self + cross attention)	Accuracy: 93.8% (APTOS), 89.8% (DDR), 86.7% (IDRiD)	Transformer-heavy; high compute (9.2 GFLOPs) and params (∼31M); no explainability	No
Cross-Dataset Fundus Analysis [31]	ODIR → RFMiD	Hybrid CNN–attention models evaluated for cross-dataset generalization	Average AUC: 90.81% (CoAtNet)	No end-to-end training; limited disease coverage	No
MobileViT-Plus DR Framework [34]	Kaggle DR (5-class)	Parallel CNN and MobileViT branches for DR grading	Accuracy: 93.67%, AUC: 0.994	Single disease; single dataset evaluation	No
CNN–Transformer Fusion [33]	APTOS-2019, IDRiD	CNN backbone fused with Transformer encoder	Accuracy: 94.28%, AUC: 0.9638	High computational overhead; DR-only task	No
CNN–TransNet [22]	Fundus datasets (4(EDC)- and 7-class)	Parallel CNN and Transformer branches with feature concatenation	Accuracy: 94.72% (4-class), 80.68% (7-class)	Very high parameter count; overfitting risk	No
EFFResNet-ViT [26]	EDC	CNN backbones fused with Transformer encoder	Accuracy: 92.54%, AUC: 98.82%	Heavy backbone fusion; high inference cost	Yes
Conv-ViT [28]	OCT dataset (CNV, DME, DRUSEN, NORMAL)	Triple-stream hybrid: Inception-V3 + ResNet-50 + ViT with feature-level fusion	Accuracy: ∼94%; Precision/Recall/F1: ∼94%	High computational complexity; single-dataset evaluation; no cross-dataset validation; lacks multimodal clinical data	No

Chapter 3

Requirements, Impacts and Constraints

3.1 Final Specifications and Requirements

In the following subsection, we detail the experimental environment and computing system that was employed to train and test the proposed model.

The early model development and prototyping were done on a local workstation with NVIDIA GeForce RTX 3060 graphics card and 16 GB of system RAM. The system comes with a 12 th Generation Intel core i5-12600K processor with a 64-bit, x64 operating system. But all the end experiments, such as hyperparameter optimization, ablation, and the entire training of the proposed MuninFormer model, were performed on Kaggle cloud computing platform to benefit access to more powerful computer graphics and guarantee the ability to be reproducible in the community of researchers.

Kaggle platform offered the availability of NVIDIA Tesla T4 GPUs that had 16 GB of. VRAM providing higher computational stability and efficiency with the large amount of training needed by the hybrid CNN-Transformer architecture. All the models were interpreted with the help of the PyTorch deep learning platform and Python 3.10.

Table the overall specifications of the development environment on both platforms, and presents the general development environment specifications applicable to both platforms, while Table 3.2 contains the specifics of the Kaggle cloud infrastructure that was used to train the final model and evaluate it.

Table 3.1: Development Environment Specifications

Component	**Specification**
Local Development Hardware	
GPU	NVIDIA GeForce RTX 3060
GPU Memory	12 GB GDDR6
System RAM	16 GB DDR4
CPU	Intel Core i5-12600K (12th Gen)
Operating System	Windows 11 / Ubuntu 22.04 (64-bit)
Software Environment	
Programming Language	Python 3.10
Deep Learning Framework	PyTorch 2.0+
Model Library	timm (PyTorch Image Models)
Image Processing	Albumentations, OpenCV, PIL
Scientific Computing	NumPy, Pandas, Scikit-learn
Visualization	Matplotlib, Seaborn
Development Environment	Jupyter Notebook / VS Code

Table 3.2: Kaggle Cloud Computing Infrastructure Specifications

Component	**Specification**
Platform Details	
Computing Platform	Kaggle Cloud Computing
Platform Type	Free Cloud-based GPU Service
Access Method	Web-based Jupyter Notebooks
GPU Hardware	
GPU Model	NVIDIA Tesla T4
GPU Architecture	Turing Architecture
GPU Memory (VRAM)	16 GB GDDR6
Tensor Cores	Yes (320 Turing Gen 2)
CUDA Cores	2,560
GPU Thermal Design Power	70W

3.2 Societal Impact

The social impact of MuninFormer architecture focuses on improving rural and accessible healthcare. It effectively helps classify globally common diseases in resource constrained regions with limited clinical capacity. Therefore, helps to improve the socioeconomic burden on global healthcare systems. Through its enhanced features of early and accurate classification of asymptomatic diseases, the model plays a significant role in preventing and detecting irreversible vision loss. Delayed diagnosis or treatment often leads to global visual impairment. Furthermore, it acts a secondary means of making clinical decisions, to help reduce human error and inter-observer variability in manual assesment which maybe time intensive. It is specifically proposed for resource-constrained environments, thus the model's lightweight design lowers the difficulty for diagnosis while still maintaining better accuracy to enhance clinical decision-making.

3.3 Environmental Impact

The environmental impact of the research primarily stems from the energy expenditure during model training and evaluation which requires heavy computational resources. Our model was designed as such to mitigate this through efficient architectural design. It uses an exceptionally lightweight model to maintain sustainability by utilizing only 7.36 million parameters. This results in low computational demand compared to standard Transformer models, requiring only 1.18 GFLOPs per inference at a standard 224 x 224 input resolution. Due to such compact footprint the model makes a better energy-efficient alternative for both the development and long term clinical use. Furthermore, the research was strategically managed through model design and ablation studies were carefully carried out to avoid redundant number of experiments. As a result, total carbon footprint was reduced which comes with large-scale hardware utilization making sure the research produces an environmentally responsible solution for global healthcare technology.

3.4 Ethical Issues

The research accounts to the importance of ethical standards in medical AI. The research uses a publicly available, de-identified dataset handled in compliance with medical regulations. By integrating Explainable AI techniques, the model produces visual heatmaps highlighting the parts of retinal structure resulting in a diagnosis. This helps in transparency and acts as a support for better interpretability for building clinical trust. It allows ophthalmologists to verify from the result to make informed decisions based on lesions, deformity of optic disc etc. It only acts as a secondary clinical support tool and not a replacement, therefore medical professionals has the final diagnostic authority. Furthermore, the study is supported by a thorough and transparent documentation of methodology to reinforces accountability and ethical integrity. This includes detailed process of data cleaning, pre-processing, splitting, comparison of other models, mitigating biases through minimizing overfitting and ensuring the reliability of diagnostic output. It also ensured fairness through internal and external cross validation of data on the proposed architecture as well as a comparative analysis with different hybrid models.

3.5 Project Timeline and Management

The project followed a structured eight-phase timeline spanning seven months, as summarized below:

- **Phase 1–2**: Literature review, problem formulation, and research design (Month 1–3)
- **Phase 3**: Dataset acquisition, preprocessing, and augmentation pipeline development (Month 4–6)
- **Phase 4**: Baseline model implementation and initial benchmarking (Month 7)

- **Phase 5**: MuninFormer architecture design and inspired component development (Month 8–9)
- **Phase 7**: Systematic ablation studies with 7 model variants (Month 10–11)
- **Phase 8**: Comparative evaluation, explainability analysis, and error case studies (Month 11)
- **Phase 9**: Final validation, thesis documentation, and code repository preparation (Month 12)

3.6 Risk Management

The MuninFormer model is designed specifically as a secondary means of support in clinical decision rather than a standalone diagnostic system. Therefore, it should not be used to drive treatment decisions from and should always require the authority of experienced ophthalmologists for final diagnosis. Furthermore, key risks also lies on the misclassification of early stage disease like Glaucoma-Normal confusion, which could delay critical interventions if used without expert oversight.

Chapter 4

Methodology

4.1 Methodology Overview

The current research is built on the concept of a supervised deep artificial intelligence approach of automated eye disease classification in retinal fundus images. In order to meet the objectives of high classification and explainability of models, the proposed methodology is based on the integration of the data preprocessing, the hybrid CNN Vision Transformer architecture, and the explainable artificial intelligence technologies. It is through this proposed architecture that classification metrics have been proposed to provide a solution along with qualitative insights that are internalized in the reasoning of the model. An architectural workflow of the proposed methodology has been shown in figure 4.1.

The dataset that has been used in the given research is retrieved by a publicly available Kaggle repository of labeled retinal fundus images in four categories of eye diseases. The data set is stratified into training, validation and test set to balance the classes to achieve unbiased analysis.

Before the model training, data cleaning and quality control process were carried out to enhance the reliability of the dataset. Perceptual hashing and image validation checks were used to detect duplicate and corrupt images and remove them. Potential label noise was identified in a confident learning-based method, where the high confidence score difference in prediction was brought to attention. To determine the balance of the data set, a class distribution analysis was done and the clean dataset obtained was used in all the further experiments.

After cleaning the datasets, image-level preprocessing was used. Enhancing local contrast and emphasizing the clinically of interest retinal structures was performed by Contrast Limited Adaptive Histogram Equalization (CLAHE). Images were all downscaled to the same fixed resolution and standard ImageNet statistic was used to normalize them. In their training, random cropping, flipping, rotation, color perturbation, noise injection and coarse dropout were used as the data augmentation methods in order to enhance the robustness and minimize the overfitting. Deterministic transformations were applied to validation and test images so as to have fair evaluation.

A number of special components have also been added in order to improve discriminative ability. The gates such as Pathology Aware Attention Pooling are selective in highlighting disease relevant areas within the feature maps. Uncertainity Gated Feature Fusion is used to adaptively balance the CNN and Transformer features depending on the real time confidence of the model's predictions.This classification is executed based on a tight combination of expert module that permits the individualization in decision limits on the recognition of various classes of diseases. In training, an auxillary intra class variance regularization loss is used to promote compact feature representations as well as generalization.

The AdamW optimizer is used to perform model training with learning-rate warm-up, cosine annealing and adaptive learning rate decay. The methods used in mitigating overfitting are regularization techniques which include label smoothing, dropout and early stopping. The model parameters are also maintained as an exponential moving average in order to stabilize the validation results.

Besides quantitative assessment, Explainable Artificial Intelligence (XAI) is also added to improve the transparency of the model. Grad-CAM++ is an extension of CNN backbone to produce class-specific localization maps, which use the regions that make the largest contribution to the predictions of a model. Visual explanations are superimposed on the original fundus images to give the intuitive understanding of the work of the model. The visualization of Attention Rollout out of the transformer is also presented to provide an alternative interpretability to various model components.

Accuracy, precision, recall, F1-score, and confusion matrices on validation and test datasets are used to test model performance. By incorporating XAI into the framework, it will be accurate as well as interpretable and applicable in clinical decision-support situations.

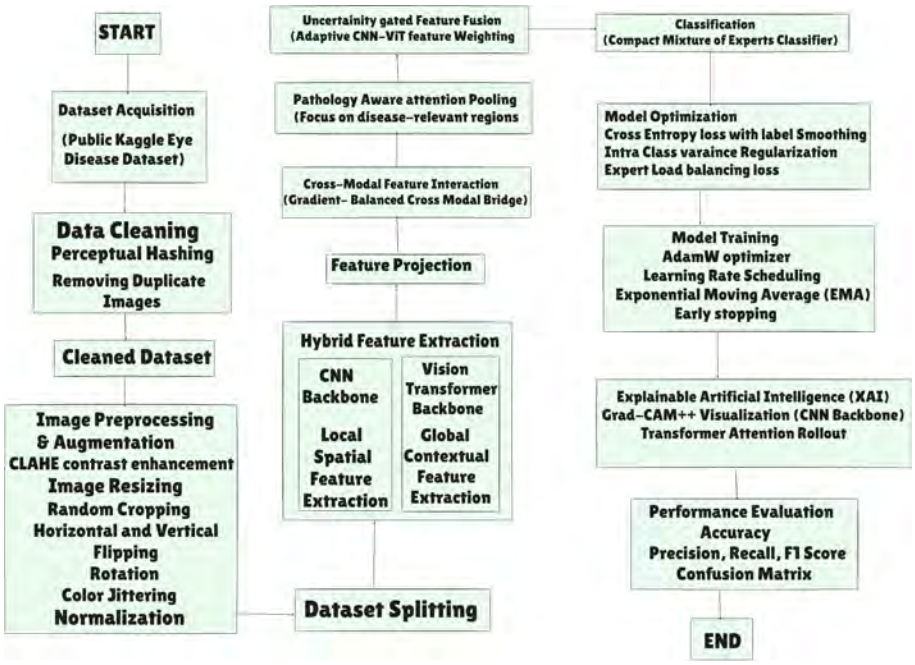


Figure 4.1: Methodology Overview

4.1.1 Architecture Overview

The MuninFormer architecture is a well coordinated pipeline wherein each system is constructed upon the outputs of the other piping stage and this information is then converted into a healthy disease prediction by a cascade of specialized processing modules. It is important to follow the full path of data flow in the architecture, analyzing how the information is narrowed, merged and standardized at every step.

The complete forward pass of the proposed model can be visualized under six interrelated stages. At first, the input fundus image enters the two parallel feature extraction path ways of CNN and ViT at once. The CNN branch captures the local patterns of the fundus image and the ViT branch models the global spatial relationships. Secondly, these diversified feature representations are mapped into a common embedding space facillating the effective alignment and the cross modal interaction. Thirdly, the Gradient-Balanced Cross-Modal Bridge allows the CNN features to attend to and integrate relevant global context of the ViT features, and balance learning cues to both arms. Fourth, the Pathology-Aware Attention Pooling projects the entire information of the spatial CNN to a small global representation with learned class-specific attention patterns. Fifth, the Uncertainty-Gated Feature Fusion is a dynamical combination of the CNN and ViT global representation, which depends on the prediction confidence of the model. Lastly the Compact Mixture-of-Experts Classifier yields the ultimate disease outputs and the Intra-Class Variance Regularizer which is applied to the training coding shapes the feature space to encourage generalization.

The difference between MuninFormer and more basic hybrid architectures is the sophisticated interaction between these elements. Instead of dealing with CNN and ViT branches extracted features independently where their outputs are simply

merged,MuninFormer presents several cross-modal interaction points and adaptive processing facilities that enable the architecture to make the most of the strengths of each branch providing a compensation of their own weaknesses.

4.1.2 Dual Backbone Feature Extraction

4.1.2.1 Design Philosopy

The dual-backbone design is a fusion of convolutional and attention-based feature extractors to obtain complementary retinal information. Convolutional neural networks focus on local pathological features including microaneurysms, exudates, and hemorrhages with the aid of translation-equivariant local filtering, whereas vision transformers predict long-range spatial associations between lesions and anatomical landmarks by using self-attention. This dual-branch architecture by combining parallel streams of Transformer and CNN enhances the detectiveness of the model to cracks in the middle of cluttered backgrounds. This design in particular solves the resolution constraints that the Transformer has per say by leveraging the capability of the CNN to retain the fine-grained structural details [48]. Extraction of cross-scale and multi-level features is made possible by the implementation of dual backbone architecture and helps in balancing the detailed aspect of the locality with the general global context and thus increases the overall robustness of the model [51]. The lightweight backbones, which are MobileNetV3-Small and ViT-Tiny, are chosen to focus on the parameter efficiency, and the inspiration in the architectures is brought in the following fusion and aggregation steps.

4.1.2.2 CNN Branch: MobileNetV3-Small with Feature Calibration

MobileNetV3-Small has been used as a lightweight backbone in the CNN branch which is used for extracting local retinal texture. MobileNetV3 is a particular model that is tailored towards the mobile environments; it has a perfect balance between computation efficiency and high-performance in image classification [18]. Thus, it is made sufficiently to be used and deployed in resource-limited clinical environments.

For an input fundus image $\mathbf{x} \in \mathbb{R}^{B\times3\times224\times224}$, the CNN backbone serves mainly as a feature extractor helping to generate a spatially organized feature tensor given in equation 4.1 rather than a final classification output.

$$\mathbf{F}_{\text{CNN}}^{\text{raw}} \in \mathbb{R}^{B\times576\times7\times7} \tag{4.1}$$

Individual feature channels are designed to capture the specific morphological signatures of retinal abnormalities of the fundus images such as exudates, hemorrhages, and microaneurysms.

To enhance the accuracy of these local descriptors before cross interaction with another mode, a lightweight channel-wise calibration module [15] is integrated. This calibration module of equation 4.2 performs global context aggregation with adaptive channel reweighting, which allows the network to highlight diagnostically informative channels whilst cutting off less relevant responses where $\mathbf{s}$ denotes the learned

channel importance weights.

$$\mathbf{F}_{\text{CNN}}^{\text{cal}} = \mathbf{F}_{\text{CNN}}^{\text{raw}} \odot \mathbf{s} \tag{4.2}$$

In the final stage, the calibrated feature tensors are flattened and projected into a sequential format given in equation 4.3, preparing them for processing by the attention layers.

$$\mathbf{F}_{\text{CNN}}^{\text{seq}} \in \mathbb{R}^{B\times 49\times 576} \tag{4.3}$$

This sequential feature map is corresponding to a 7×7 spatial grid. By maintaining the integrity of localized pathological markers, this representation provides a robust foundation for subsequent attention-driven global modeling.

4.1.2.3 ViT Branch: Vision Transformer Tiny

Simultaneously with the CNN branch, the global contextual features are obtained with the help of Vision Transformer Tiny (ViT-Tiny). In contrast to convolutional architectures, Vision Transformers (ViTs) are not based on the use of convolutional layers as feature extractors, but the ViTs model representations of global relationships, with the help of self-attention mechanisms [17]. The ViT models long distance spatial relationship by self attention and making it explicit in identification of the distribution of lesions with respect to the anatomical landmarks like optic disc and macula.

The input image is divided into non overlapping 16×16 patches producing 196 patch tokens. The generated token sequences after linear and positional encoding is run through a stack of transformer layers. As it is customary, the classification token is abandoned and the patch-level representations are kept.

$$\mathbf{F}_{\text{ViT}} \in \mathbb{R}^{B\times 196\times 192} \tag{4.4}$$

The patch-level information shown in equation 4.4 in each token is enhanced with global information self-attention. These characteristics supplement the localized representations of the CNN and are the source of context of global interaction that is subsequently cross-modal.

4.1.2.4 Feature Projection to a Common Embedding Space

The CNN and ViT branches generate feature sequences which vary in channels dimensions. Both representations are to allow an interaction across modalities coded into a common embedding space of size C = 192 through the use of linear transformations and layer normalization and dropout.

$$\mathbf{F}_{\text{CNN}}^{\text{proj}} \in \mathbb{R}^{B\times 49\times 192} \quad \mathbf{F}_{\text{ViT}}^{\text{proj}} \in \mathbb{R}^{B\times 196\times 192} \tag{4.5}$$

This correspondence shown in equation 4.5 s necessary so that the spatial tokens of each modality can be located within a similar feature space, allowing successful attention fusion but retaining their synergistic spatial scales.

4.1.3 Gradient-Balanced Cross-Modal Bridge (GBCMB)

4.1.3.1 Cross-Modal Gradient Balancing and Information Exchange

A frequent issue in multimodal neural architecture is that modalities can be underfitted because the modalities may conflict on the gradients they give during cross-modal fusion and the modalities do not learn equally on cross-modal fusions [53]. This is an especially acute problem in CNN hybrid models, in which CNNs generate spatially concentrated gradients based on the localized patterns, and ViTs spread gradients more across the input picture with self-attention. Otherwise, the imbalance can cause one of the modalities to prevail the complementary benefit of multi-modal fusion, which is reduced during the optimization process.

The Gradient-Balanced Cross-Modal Bridge (GBCMB) is aimed at dealing with this issue using a learnable gradient scaling mechanism embedded in a cross-attention framework. This module fulfills two major goals: (1) allowing CNN functions to take into account ViT-based global context, and (2) preserving the gradients between the two of the branches. GBCMB learns attention patterns with respect to the ViT tokens (global context) are relevant to each CNN token. The gradient balance and (local feature) will make sure that both modalities learn effectively. Moreover, the module is used to carry out non-linear transformations in order to combine the.transformed cross-attended information to refined and consolidated representations.

4.1.3.2 Cross-Modal Attention Formulation

GBCMB employs a cross-attention mechanism used for respective feature embeddings before computing attentions [49] given in equation 4.6 in which CNN features act as queries, while ViT features serve as keys and values. Given projected feature sequences $\mathbf{F}_{\text{CNN}}^{\text{proj}} \in \mathbb{R}^{B\times N_c\times C}$ and $\mathbf{F}_{\text{ViT}}^{\text{proj}} \in \mathbb{R}^{B\times N_v\times C}$, queries, keys, and values are computed as:

$$\mathbf{Q} = \mathbf{W}_Q\mathbf{F}_{\text{CNN}}^{\text{proj}}, \quad \mathbf{K} = \mathbf{W}_K\mathbf{F}_{\text{ViT}}^{\text{proj}}, \quad \mathbf{V} = \mathbf{W}_V\mathbf{F}_{\text{ViT}}^{\text{proj}} \tag{4.6}$$

where $\mathbf{W}_Q, \mathbf{W}_K, \mathbf{W}_V \in \mathbb{R}^{C\times C}$ are learnable projection matrices.

The cross-attention output shown in equation 4.7 is computed using scaled dot-product attention:

$$\text{Attention}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^\top}{\sqrt{d_k}}\right)\mathbf{V} \tag{4.7}$$

where d_k denotes the per-head embedding dimension. The softmax ensures attention weights sum to 1 across the context sequence, making them interpretable as a probability distribution over which ViT tokens to attend to.

4.1.3.3 Implicit Gradient Balancing via Reciprocal Scaling

The key inspiration of GBCMB lies in its gradient-balancing mechanism. A learnable scalar parameter shown in equation 4.8 β controls reciprocal scaling between the two modalities during training:

$$\alpha = \text{clamp}(\beta, 0.5, 2.0). \tag{4.8}$$

Our approach is also closest in CLIP with the difference that we apply the scaling at the feature level to symmetrically regulate gradient flow between CNN and ViT branches [19]. During the forward pass, modality features are scaled by the formula 4.9:

$$\tilde{\mathbf{F}}_{\text{CNN}} = \alpha \cdot \mathbf{F}_{\text{CNN}}^{\text{proj}}, \quad \tilde{\mathbf{F}}_{\text{ViT}} = \frac{1}{\alpha} \cdot \mathbf{F}_{\text{ViT}}^{\text{proj}}. \tag{4.9}$$

At inference time, no scaling is applied. Although the scaling in equation 4.9 is applied in the forward pass, its primary effect manifests during backpropagation. By the chain rule of equation 4.10, gradients with respect to the original features satisfy:

$$\frac{\partial \mathcal{L}}{\partial \mathbf{F}_{\text{CNN}}^{\text{proj}}} = \alpha \cdot \frac{\partial \mathcal{L}}{\partial \tilde{\mathbf{F}}_{\text{CNN}}}, \quad \frac{\partial \mathcal{L}}{\partial \mathbf{F}_{\text{ViT}}^{\text{proj}}} = \frac{1}{\alpha} \cdot \frac{\partial \mathcal{L}}{\partial \tilde{\mathbf{F}}_{\text{ViT}}}. \tag{4.10}$$

As a result, when $\alpha > 1$, gradients flowing into the CNN branch are amplified while those entering the ViT branch are attenuated; the opposite occurs when $\alpha < 1$. Since β is optimized end-to-end, the network adaptively reduces extreme imbalance in learning dynamics between modalities.

Unlike explicit gradient normalization or loss reweighting techniques, this balancing is achieved implicitly through reciprocal feature scaling, allowing standard backpropagation without custom gradient manipulation.

4.1.3.4 Feature Refinement

The cross-attended features are further refined using a feed-forward network (FFN) with residual connections shown in equation 4.11 and layer normalization in equation 4.12:

$$\mathbf{H} = \text{CrossAttn}(\tilde{\mathbf{F}}_{\text{CNN}}, \tilde{\mathbf{F}}_{\text{ViT}}) \tag{4.11}$$

$$\mathbf{F}_{\text{CNN}}^{\text{enh}} = \mathbf{H} + \text{FFN}(\text{LayerNorm}(\mathbf{H})) \tag{4.12}$$

The output $\mathbf{F}_{\text{CNN}}^{\text{enh}} \in \mathbb{R}^{B \times N_c \times C}$ is CNN features that have been enhanced by global context of ViT branch. Due to cross-attention, which functions on pre-scaled features, balanced gradient magnitude is maintained through this design from analogous to temperature based scaling in CLIP [19]. The output represents CNN features enriched with relevant global context while maintaining balanced gradient flow across modalities. These features are passed to the subsequent Pathology-Aware Attention Pooling module.

4.1.4 Pathology-Aware Attention Pooling (PAAP)

The Pathology-Aware Attention Pooling (PAAP) is a class-dependent pooling scheme that is developed in response to the problem of pooling spatial features of medical images. The classical pooling techniques like global average pooling or max pooling consider all the spatial locations, whereas in the case of medical images, the diagnostically useful information can be concentrated in certain areas, which differ across different types of diseases. This observation is in line with earlier work in medical image analysis, which demonstrates that diagnostically significant patterns are usually spatially localized, and different across diseases, which in turn is the motivation of attention-based or region-aware pooling strategies [10], [11]. PAAP surpasses this

weakness by mastering the ability to pay attention to the most informative areas to each pathology.

4.1.4.1 PAAP Architecture Overview

PAAP works by training individual attention questions per disease group, which allows the model to concentrate on other parts of the image based on the pathology under consideration. The attention mechanism gives a more important role to physical locations, which are applicable to the particular illness. The key components of PAAP are:

- **Class-Conditional Attention Queries:** Separate queries are trained on diseases classes, and the model is able to target separate regions that are best instructive to a given pathology.
- **Spatial Attention Computation:** Attention weights are calculated with the help of scaled dotproduct attention, features the queries in the CNN feature highlighting of critical areas.
- **Aggregation:** The features that are attended are concatenated and somehow aggregated to constitute a condensed global representation that is projected to the embedded dimensional for classification.

4.1.4.2 Class-Conditional Attention Modeling

In contrast to generic attention systems that are taught one global significance distribution, PAAP uses its own set of learnable attention queries under every class. These queries are not anatomically fixed points, but acquire to encode spatial patterns giving the most valuable indicators of a certain pathology. This design enables other classes of diseases to serve other parts of the same image.

Our pathology-aware attention pooling is inspired by query-based attention mechanisms such as DETR [13], but differs in that we employ class-specific sets of learnable queries to capture heterogeneous disease-related spatial patterns.The PAAP framework incorporates a trainable query shown in equation 4.13 allowing the model to iteratively refine the values through backpropagation.

$$\mathbf{Q} \in \mathbb{R}^{K \times Q \times C} \tag{4.13}$$

where K is the number of classes, Q denotes the number of queries per class, and C is the embedding dimension. The querying of multiple queries by class enables the model to learn complementary spatial patterns, e.g. a variety of lesion types or spatial patterns related to a single disease.

4.1.4.3 Spatial Attention Computation

The spatial attention computation used in our model follows the query-based attention mechanism used in DETR[13] which extends with class-specific queries to enable pathology-aware spatial pooling. Attention weights, on each class k and

query q, are computed on spatial tokens with scaled dot-product attention shown in equation 4.14.

$$\mathbf{A}_{k,q} = \text{softmax}\left(\frac{\mathbf{Q}_{k,q}\mathbf{K}^{\top}}{\sqrt{C}}\right) \tag{4.14}$$

where $\mathbf{A}_{k,q} \in \mathbb{R}^{N}$. This operation allows each query to sequentially attend to every spatial positions, selecting the features whose patterns perfectly match the characteristics of the previously learned classes.

We derive the attenuated representation for every class-query pair using the following equation 4.15.

$$\mathbf{Z}_{k,q} = \mathbf{A}_{k,q}\mathbf{V} \tag{4.15}$$

In the equation $\mathbf{Z}_{k,q}$ represents a query-specific summary vector and class-specific summary vector acquired through weighted spatial CNN features by their relevance to a given pathology.

4.1.4.4 Aggregation and Global Representation

In every class, the features attended of all queries are put together to create a class-level representation, which represents several spatial features of the pathology. When the predictions of the classes cannot be trusted during initial stages of training, representations of classes are summed up with a uniform average between classes shown in equation 4.16.

$$\mathbf{Z} = \frac{1}{K}\sum_{k=1}^{K}\text{Concat}(\mathbf{Z}_{k,1}, \ldots, \mathbf{Z}_{k,Q}). \tag{4.16}$$

Our single-class average-out method and combination of various query-level representations with linear projection is inspired by the query-based attention model, like DETR[13], and the need to find reliable class predictions in systems where the class prediction is less reliable. This combination approach is stable in learning and provides the opportunity of all classes to aid the pooled representation. The resulting vector shown in equation 4.17 is reconstructed back into the embedding dimension.

$$\mathbf{F}_{\text{CNN}}^{\text{global}} = \mathbf{W}_{\text{out}}\mathbf{Z}, \tag{4.17}$$

where $\mathbf{W}_{\text{out}} \in \mathbb{R}^{C\times(Q\cdot C)}$. The multi-query representation is concatenated and the projection matrix $\mathbf{W}_{\text{out}}$ reduces it into a global feature representation of fixed dimension, which allows complementary spatial evidence to be easily integrated.

4.1.5 Uncertainty-Gated Feature Fusion (UGFF)

Uncertainty-Gated Feature Vision (UGFF) serves as an adaptive mechanism that integrates global features from CNN and ViT backbones by weighting their contributions according to input specific predictive uncertainty. The confidence of predictions that the model provides depends on the inputs. The uncertainty-gated aspect of the feature fusion is based on uncertainty-weighted learning models [9], and expanded to adaptively combine CNN and Transformer representations depending on input-specialized predictive uncertainty. CNN features prove to be more reliable on some images with obvious local lesions and ViT features are more reliable on some

images with diffuse patterns. The variability is handled by UGFF shown in figure 4.2 which modulates the contribution of each modality dynamically depending on the estimated uncertainty of the model.

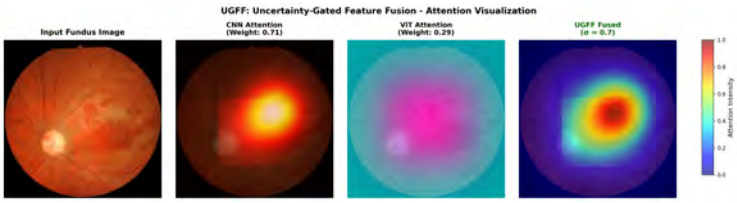


Figure 4.2: The attention maps for CNN, ViT, and their fused representation based on uncertainty.

4.1.5.1 UGFF Architecture

UGFF consists of the following key components:

- **Uncertainty Estimation:** To estimate the uncertainty of the model, a lightweight head that takes the concatenated CNN and ViT features is used to determine the confidence of the model on a per-input basis. The signal of uncertainty directs the fusion process and the model is able to highlight the most dependable modality of each input.
- **Modality-Specific Feature Processing:** All modalities (CNN and ViT) are processed separately by an independent processing path before fusion to transform and normalize the modality.
- **Uncertainty-Guided Fusion:** The processed CNN and ViT features are mixed with weights which get dynamically updated according to the uncertainty value. CNN features are more weighted when the uncertainty is great and equally both modalities are weighted when the uncertainty is small.
- **Fusion Alignment:** The resulting fused representation is in turn projected to a shared embedding space via a linear projection matrix which ensures the desired combination of CNN and ViT features and provides the fused representation with downstream classification.

4.1.5.2 Uncertainty Estimation

UGFF estimates a scalar uncertainty shown in equation 4.18 where value for each input using the concatenated global CNN and ViT features.

$$\mathbf{f}_{\text{comb}} = \text{Concat}\left(\mathbf{F}_{\text{CNN}}^{\text{global}}, \mathbf{F}_{\text{ViT}}^{\text{global}}\right) \tag{4.18}$$

where $\mathbf{f}_{\text{comb}} \in \mathbb{R}^{2C}$. The concatenated vector $\mathbf{f}_{\text{comb}} \in \mathbb{R}^{2C}$ jointly encodes local CNN evidence and global ViT context, serving as the input to the uncertainty estimation head.

The proposed uncertainty estimation head is based on the principle of learnable gating mechanisms of mixture-of-experts models, which generates a scalar uncertainty score which flexibly changes multimodal feature contributions[8]. This combined

representation is processed by a lightweight uncertainty estimation head given in equation 4.19 implemented as a two-layer multilayer perceptron in equation 4.20.

$$\mathbf{h} = \mathrm{ReLU}(\mathbf{W}_1 \mathbf{f}_{\mathrm{comb}}) \tag{4.19}$$

$$u = \sigma(\mathbf{W}_2 \mathbf{h}) \tag{4.20}$$

where $\mathbf{W}_1 \in \mathbb{R}^{\frac{C}{2} \times 2C}$, $\mathbf{W}_2 \in \mathbb{R}^{1 \times \frac{C}{2}}$, $\sigma(\cdot)$ denotes the sigmoid function, and $u \in [0, 1]$ represents the estimated uncertainty. Values close to zero indicate high confidence, while values close to one indicate high uncertainty.The figure 4.3 highlights an overview of the uncertainty estimation.

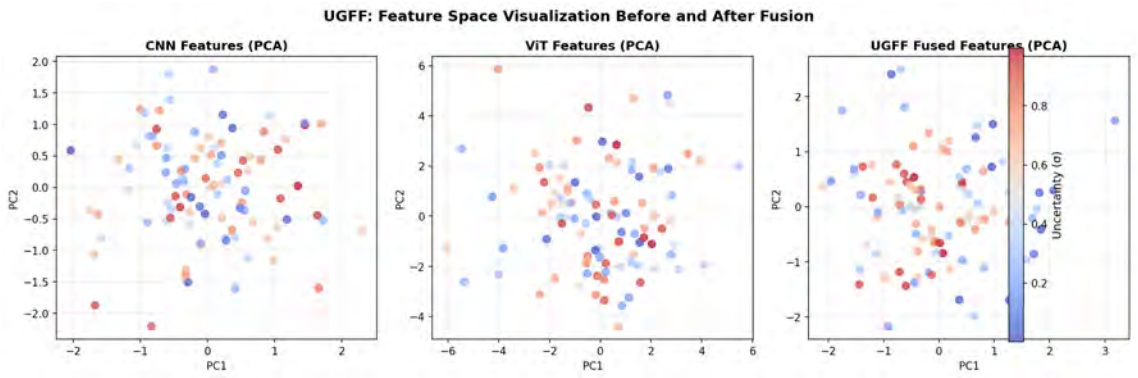


Figure 4.3: UGFF Feature Space Visualization Before and After Fusion.

4.1.5.3 Modality-Specific Feature Processing

As is customary in multimodal learning, every modality is represented by a separate projection and normalization path before fusion, or can be represented by an embedding head like those of CLIP [19], but modality-specific. Before fusion, each modality is passed through an independent processing path to enable modality-specific transformation and normalization shown in equation 4.21 and equation 4.22.

$$\tilde{\mathbf{F}}_{\mathrm{CNN}} = \mathrm{GELU}\big(\mathrm{LN}(\mathbf{W}_{\mathrm{CNN}} \mathbf{F}_{\mathrm{CNN}}^{\mathrm{global}})\big) \tag{4.21}$$

$$\tilde{\mathbf{F}}_{\mathrm{ViT}} = \mathrm{GELU}\big(\mathrm{LN}(\mathbf{W}_{\mathrm{ViT}} \mathbf{F}_{\mathrm{ViT}}^{\mathrm{global}})\big) \tag{4.22}$$

where $\mathbf{W}_{\mathrm{CNN}}, \mathbf{W}_{\mathrm{ViT}} \in \mathbb{R}^{C \times C}$ and LN denotes layer normalization.

4.1.5.4 Uncertainty-Guided Fusion

The processed features are combined using uncertainty-dependent weights by the equation 4.23.

$$w_{\mathrm{CNN}} = 0.5 + 0.3u, \quad w_{\mathrm{ViT}} = 0.5 - 0.3u \tag{4.23}$$

This formulation ensures that the weights are positive and sum to one for all values of u. When uncertainty is low, both modalities contribute equally. The more uncertain they are, the more they are emphasized on CNN features, which indicates their strength to localized pathological evidence.

The fused representation is computed as shown in the equation 4.24.

$$\mathbf{F}_{\mathrm{fused}} = w_{\mathrm{CNN}} \cdot \tilde{\mathbf{F}}_{\mathrm{CNN}} + w_{\mathrm{ViT}} \cdot \tilde{\mathbf{F}}_{\mathrm{ViT}}. \tag{4.24}$$

The uncertainty-dependent weighting is based on the principle of uncertainty-aware gating applied to previous results and generates a convex mixture of CNN and ViT representations, which is positive and stable [8], [9].The fused representations shown in equation 4.25 are ultimately projected into a shared embedding spacing through a linear layer to ensure proper dimensional consistency. space 4.25:

$$\mathbf{F}_{\text{fused}} = \mathbf{W}_{\text{fusion}}\mathbf{F}_{\text{fused}}, \tag{4.25}$$

where $\mathbf{W}_{\text{fusion}} \in \mathbb{R}^{C\times C}$. This linear projection uses a learnable transformation which reencodes the uncertainty weighted features into a single embedding space in dimension C. The projection matrix$\mathbf{W}_{\text{fusion}} \in \mathbb{R}^{C\times C}$ can be trained to optimally combine and re-scale the fused representation, reducing biases related to modality that are added in the process of weighted fusion. By reconciling the features obtained from different modailities, this procedure facilitates deeper synergy between the CNN and ViT elements. Ultimately, it generates a standardized and a fixed global representation which is optimized for the final classification task. A visual representation has been shown in figure 4.4.

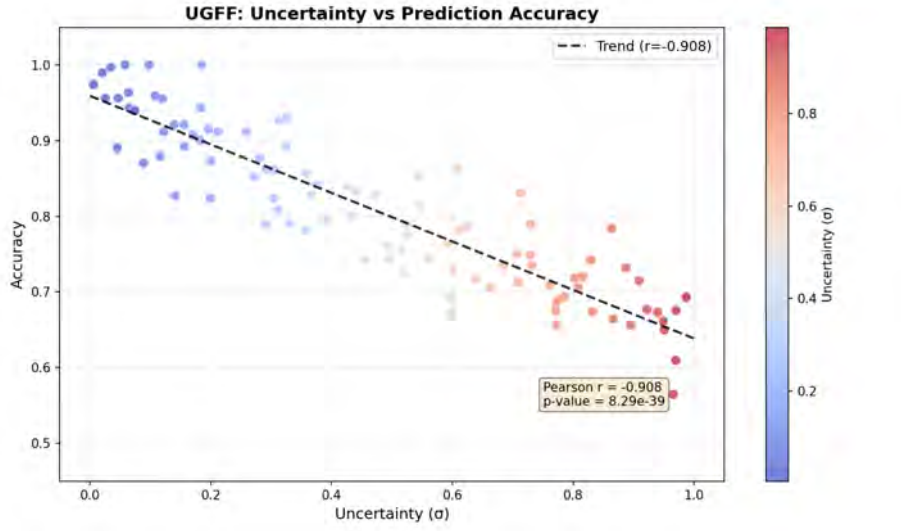


Figure 4.4: Uncertainty vs Prediction Accuracy of UGFF

4.1.5.5 Learning Dynamics and Output

The uncertainty estimation head is trained in the end to end without supervision. Using backpropagation, the network is able to learn values of uncertainty that enhance the performance of classification by giving preference to the more accurate modality with each input.

The last output $\mathbf{F}_{\text{fused}} \in \mathbb{R}^{C}$ is the input to the next Compact Mixture- of-Experts classifier, and the uncertainty values are a signal to be understood, the signal of the certainty of the model on its predictions.

4.1.6 Compact Mixture-of-Experts Classifier (CMOE)

Compact Mixture-of-Experts (CMOE) is a small scale implementation of MoE that operates with a small number of experts, and shallow networks of experts. CMOE is

a small classification component that is intended to substitute a single large classifier with a network of expert networks that are specialized. Every specialist is concerned with training a small decision boundary of various dimensions of the classification problem. The suggested CMOE classifier is based on the paradigm of mixture-of-experts, which uses a set of lightweight expert networks and a gating mechanism based on input, and uses load-balancing regularization against expert collapse[1], [8]. The model makes use of a gating network to assign dynamic weights to every expert in accordance with the input features. This enables specialization to be adaptive without making the models more complex as a way of guaranteeing efficient and effective classification.

4.1.6.1 CMOE Architecture

The CMOE architecture consists of multiple expert networks and a gating network. The key components of the architecture are as follows:

- **Expert Networks:** CMOE is based on E expert networks, each of which is a two-layer small MLP. These professionals have the specialization of learning tight decision boundaries that concentrate on various facts of the classification task.
- **Gating Network:** The gating network identifies the relative significance of every expert on a particular input. It gives a probability distribution of the experts with respect to the input features.
- **Expert Aggregation:** Individual experts come up with class logits and the end result is determined as a weighted average of all the individual expert outputs. The gating mechanism modulates the weighting coefficients adaptively and conditioned on the specific features of the input data.
- **Load Balancing Regularization:** CMOE has a load balancing regularizer to maintain balanced expert usage and eliminate collapse (when the majority of inputs are directed to one expert), where the load balancing regularizer will be punitive against unbalanced expert utilization.

4.1.6.2 Expert Networks

CMOE consists of E expert networks, each implemented as a small two-layer MLP. Each specialist is trained as a two-layer lightweight multilayer perceptron, with the typical mixture-of-experts formulation, in which each specialist is trained on specialized decision functions on the input feature space [1]. Given the fused global feature vector $\mathbf{F}_{\text{fused}} \in \mathbb{R}^C$, the output of the e-th expert is computed as shown in eqaution 4.26.

$$\mathbf{o}_e = \mathbf{W}_{e,2}\ \text{ReLU}(\mathbf{W}_{e,1}\mathbf{F}_{\text{fused}}) \tag{4.26}$$

where $\mathbf{W}_{e,1} \in \mathbb{R}^{D\times C}$, $\mathbf{W}_{e,2} \in \mathbb{R}^{K\times D}$, D denotes the expert hidden dimension, and K is the number of classes. The small hidden dimension encourages each expert to learn compact, specialized decision boundaries.

4.1.6.3 Gating Network

A gating network determines the relative contribution of each expert for a given input. The gating network is a standard mixture-of-experts model, it is a softmax-normalized distribution over experts that depend on the input features [1]. The gating logits are computed as shown in equation 4.27.

$$\mathbf{g} = \text{softmax}(\mathbf{W}_{\text{gate}}\mathbf{F}_{\text{fused}}) \tag{4.27}$$

where $\mathbf{W}_{\text{gate}} \in \mathbb{R}^{E\times C}$ and $\mathbf{g} \in \mathbb{R}^{E}$ is a probability distribution on experts. Soft gating makes sure that the experts are differentiable and play a role in the final prediction.

4.1.6.4 Expert Aggregation

The end prediction is the weighted average of expert predictions, according to the conventional mixture-of-experts combination process [1]. The prognosis of each expert is generated, and the overall prognosis is acquired as a weighted sum according to equation 4.28.

$$\mathbf{y} = \sum_{e=1}^{E} g_e \cdot \mathbf{o}_e \tag{4.28}$$

In this formulation, the classifier is able to take a combination of several specialized views with adjustments to the importance of experts on a sample-by-sample basis.

4.1.6.5 Load Balancing Regularization and Output

CMOE uses a load balancing regularization term to avoid the phenomenon of expert collapse, in which the gating network directs most of the inputs to one expert. Expert usage is a moving average that is exponential shown in equation 4.29.

$$\mathbf{u}_e \leftarrow \alpha\mathbf{u}_e + (1-\alpha)\,\mathbb{E}[g_e] \tag{4.29}$$

where α is a smoothing coefficient. In order to avoid the risk of expert collapse, we use a load-balancing regularization method, similar in use to previous mixture-of-experts models, that promotes equal use of experts in training [8]. The load balancing loss rewards the failure to follow the uniform utilization of the experts as given in equation 4.30.

$$\mathcal{L}_{\text{balance}} = \lambda \sum_{e=1}^{E} \left(u_e - \frac{1}{E} \right)^2 \tag{4.30}$$

This regularization encourages balanced expert participation while preserving classification performance. This standardization will encourage equal professional participation and sustain the performance of the classification.

It uses the resulting logits $\mathbf{y} \in \mathbb{R}^K$ to calculate the primary classification loss. The gating probabilities can bring further insights on the interpretability as they indicate the experts lend themselves most to every prediction.

4.1.7 Intra-Class Variance Regularization (ICVR)

The proposed intra-class variance regularization shares the same concept with the previous literature that has regularized the idea of compact class-wise feature distributions to enhance generalization, like center loss and Fisher-based discriminant objectives [6]. Intra-Class Variance Regularization (ICVR) is a training-time regularization method that promotes compact distributions of features within each of the classes and is aimed at improving the generalization of a model. The topology of the learned feature space is very critical in the generalization of a model. Features that belong to the same or similar class are likely to be dispersed in models that are prone to overfitting and this leads to brittle decision boundaries. In contrast, robust models with high generalization capacity structure the representation space such that features belonging to the same category form dense and distinct clusters with clear inter-class margins.

The explicit method to do this in ICVR punishes variance in the distribution of features of each group, thus encouraging more reliable and general decision boundaries. ICVR facilitates the reduction of intra-class scattering to make the model learn to arrange the class features in a compact way, without changing the forward architecture of the model. This norming procedure is an additional loss to the main training goal, which brings about the stabilization of the feature space and decreases overfitting.

4.1.7.1 Centroid Estimation

ICVR maintains a running centroid for each class in the embedding space. The suggested intra-class regularization of the variance is largely similar to center-based compactness of features goals, featuring momentum-steadied centroid of classes and uncoupled proxy objectives to stabilize training [6]. Let $\mathbf{F}_{\text{fused}} \in \mathbb{R}^{B\times C}$ denote the fused feature representations and $\mathbf{y} \in \{1,\dots,K\}^B$ the corresponding class labels. For each class k, a centroid $\boldsymbol{\mu}_k \in \mathbb{R}^C$ is updated using a momentum-based scheme shown in equation 4.31.

$$\boldsymbol{\mu}_k \leftarrow \begin{cases} \mathbf{m}_k, & \text{if } n_k = 0 \\ \gamma\boldsymbol{\mu}_k + (1-\gamma)\mathbf{m}_k & \text{otherwise} \end{cases} \tag{4.31}$$

where $\mathbf{m}_k$ is the mean feature vector of class k in the current batch, n_k tracks whether the centroid has been initialized, and $\gamma \in [0, 1)$ is the momentum coefficient. This update stabilizes centroid estimates by reducing sensitivity to batch-level noise.

4.1.7.2 Variance Regularization Loss

The intra-class variance loss given in equation 4.32 penalizes the average squared distance between features and their corresponding class centroids.

$$\mathcal{L}_{\text{ICVR}} = \frac{1}{K}\sum_{k=1}^{K} \mathbb{E}_{i:y_i=k}\left[\left\|\mathbf{F}^{(i)}_{\text{fused}} - \text{stopgrad}(\boldsymbol{\mu}_k)\right\|_2^2\right] \tag{4.32}$$

The centroid vectors 4.32 are detached from the computational graph to prevent trivial solutions in which centroids move toward individual samples rather than

guiding feature learning.

Averaging across classes ensures balanced regularization even under class imbalance.

4.1.7.3 Integration with Training Objective

ICVR is applied only during training and acts as an auxiliary regularization term. The total training objective shown in equation 4.33 is given by:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{CE}} + \mathcal{L}_{\text{balance}} + \lambda_{\text{ICVR}}\mathcal{L}_{\text{ICVR}} \tag{4.33}$$

where $\mathcal{L}_{\text{CE}}$ denotes the primary classification loss, $\mathcal{L}_{\text{balance}}$ is the mixture-of-experts load balancing loss, and λ_{ICVR} controls the strength of the regularization. A small weighting factor ensures that ICVR promotes compact feature clusters without dominating the optimization process.

4.1.8 Training Methodology

4.1.8.1 Loss Function

The training objective combines a primary classification loss with auxiliary regularization terms that promote balanced expert utilization and robust feature geometry. The total loss is defined as in equation 4.34.

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{CE}} + \mathcal{L}_{\text{balance}} + \lambda_{\text{ICVR}}\mathcal{L}_{\text{ICVR}} \tag{4.34}$$

where $\mathcal{L}_{\text{CE}}$ denotes the cross-entropy classification loss, $\mathcal{L}_{\text{balance}}$ is the mixture-of-experts load balancing loss, and $\mathcal{L}_{\text{ICVR}}$ is the intra-class variance regularization term. The weighting factor λ_{ICVR} is set to a small value to ensure regularization does not dominate the optimization process.

To reduce the issue of overconfident predictions, label smoothing is applied on the classification targets as shown in equation 4.35.

$$\tilde{y}_k = (1 - \epsilon)y_k + \frac{\epsilon}{K} \tag{4.35}$$

where y_k is the one-hot label, K is the number of classes, and $\epsilon = 0.08$. TThis leads to calibrated probability estimates better and enhances generalization.

4.1.8.2 Optimizer Configuration

The model is trained using the AdamW optimizer with the following hyperparameters: learning rate 3×10^{-5}, weight decay 0.025, $\beta_1 = 0.9$, $\beta_2 = 0.999$, and numerical stability constant $\epsilon = 10^{-8}$. AdamW is selected for its decoupled weight decay, which provides consistent regularization across both pretrained and newly introduced parameters. The relatively low learning rate is chosen to preserve the pretrained features of MobileNetV3 and ViT-Tiny while allowing fine-tuning on the fundus imaging domain.

4.1.8.3 Learning Rate Scheduling

The optimization process follows a learning rate protocol that incorporates an initial warmup phase which is followed by a cosine annealing decay schedule. During the warmup phase (first 15 epochs), the learning rate increases linearly from near zero to the initial learning rate. This is followed by cosine annealing in equation 4.36 until convergence, allowing smooth decay of the learning rate:

$$\eta(t) = \eta_{\min} + \frac{1}{2}(\eta_{\max} - \eta_{\min}) \left[1 + \cos\left(\pi \frac{t - T_{\text{warm}}}{T_{\max} - T_{\text{warm}}}\right)\right] \tag{4.36}$$

In addition, a ReduceLROnPlateau strategy is applied as a supplementary mechanism to reduce the learning rate when validation performance stagnates.

4.1.8.4 Exponential Moving Average

An exponential moving average (EMA) shown in equation 4.37 of model parameters is maintained during training.

$$\boldsymbol{\theta}_{\text{EMA}} = \alpha \boldsymbol{\theta}_{\text{EMA}} + (1 - \alpha)\boldsymbol{\theta} \tag{4.37}$$

with decay factor $\alpha = 0.9995$. EMA weights are used exclusively for validation and final evaluation, providing more stable and reliable performance estimates.

4.1.8.5 Gradient Stabilization

To prevent gradient explosion, particularly in the presence of transformer components and multiple auxiliary losses, gradient norms are clipped to a maximum value of 1.0 as given in equation 4.38.

$$\|\nabla \mathcal{L}\| \leftarrow \min(\|\nabla \mathcal{L}\|, 1.0) \tag{4.38}$$

4.1.8.6 Early Stopping

Training is terminated early if validation accuracy fails to improve for 80 consecutive epochs. Additionally, training is stopped if the gap between training and validation accuracy exceeds 5% after sufficient convergence, preventing excessive overfitting.

4.1.8.7 Batch Configuration

A batch size of 20 is used during training to balance gradient stability and memory constraints. For validation and testing, a larger batch size of 40 is employed to improve computational efficiency, as gradients are not required.

4.2 Model Specification

The classification of medical images has a number of challenges that essentially differ with ones that are faced during natural image classification. The pathologic characteristics tend to be delicate, spatially confined, and extremely not only different in the same diagnosis group. Moreover, medical data are generally small and can have the unavoidable annotation errors. These properties require a model structure

that puts emphasis on robustness, adaptive reasoning and interpretation instead of bare representational ability. Therefore, the proposed model has clear architectural mechanisms that will deal with these obstacles on the levels of representation, fusion, and decision making.

The suggested architecture MuninFormer adheres to the design of hybrid convolutional and transformer based architectures. The main reason behind this selection is that Convolutional Neural Networks (CNNs) and Vision Transformers (ViTs) complement each other. CNNs offer high inductive bias of local spatial microtextures like lesions, vascular anatomy and texture abnormalities, which play a vital role in medical images. Their capability of modeling long-range dependencies is however limited. Vision Transformers, by contrast, are quite effective at learning the contextual relationships globally by using self-attention, but usually have to be trained on large datasets to generalise effectively. The proposed architecture will integrate the two paradigms to help capture both the fine-grained local pathology and global anatomical context jointly with the preservation of data efficiency.

To ensure successful communication between CNN and ViT branches, extracted feature projections of both backbones are mapped into an embedding space of fixed dimensionality. This is required to ensure the two branches are together since they are different significantly in their feature statistics, spatial resolution and semantic structure. Let $F_{\text{cnn}} \in \mathbb{R}^{B \times N_{\text{cnn}} \times C}$ and $F_{\text{vit}} \in \mathbb{R}^{B \times N_{\text{vit}} \times C}$ denote features of the projected CNN and ViT feature sequences respectively. Linear projection and then normalization make sure that the two modalities play a similar role in the fusion process, thus enhancing training, steadiness and the averting of premature pre-emption of one branch by the other. The comparison of the extraction of local and global features with CNN and ViT backbones, respectively, is presented in Figure 4.5. These two streams of features are then projected to the same dimension embedding space where they are then fused to create cross-modal interaction and balance on propagated gradients.

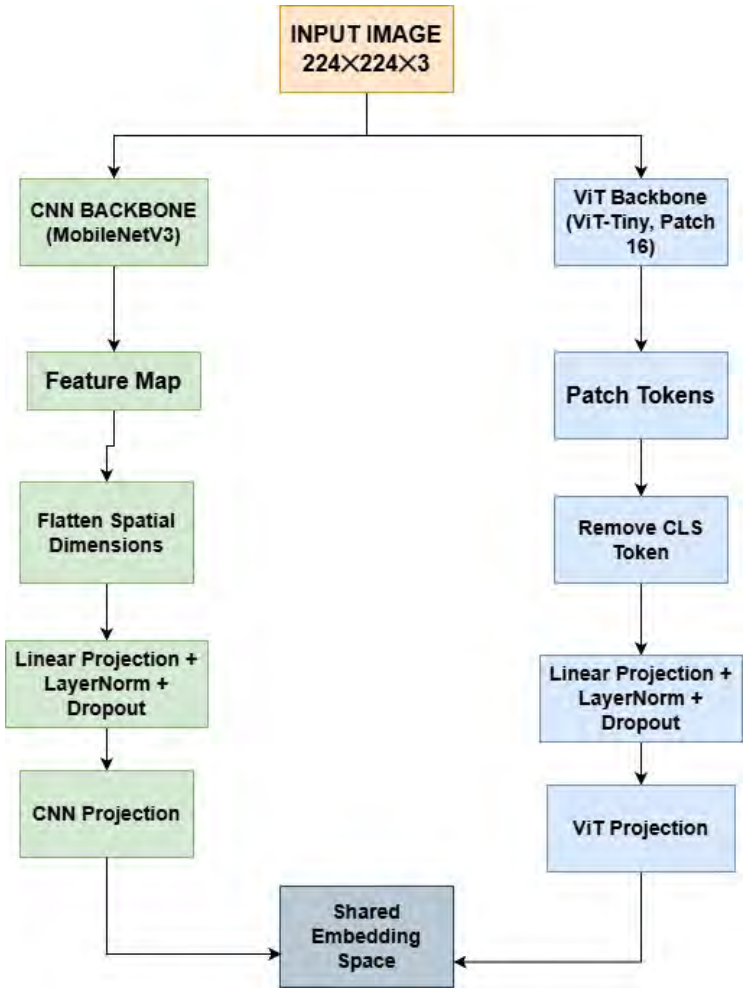


Figure 4.5: Dual Feature Extraction of CNN and ViT branches

One of the most perilous issues in the hybrid architectures is the unbalanced gradient flow during training whereby one of the modalities continues to receive gradient updates that are disproportionately huge. This may result in inadequate exploitation of the complementary arm and poor feature learning. To overcome this problem, Gradient-Balanced Cross-Modal Bridge is provided. In this module, the cross-attention of CNN and ViT features are performed with adaptive scaling in training to equalize the magnitude of the gradient. The design explicitly controls the learning dynamics instead of implicitly balancing the system, hence both the local and global representation are optimized together in the course of training. Figure 4.6 gives us an overview of the architecture of the Gradient-Balanced Cross-Modal Bridge (GBCMB). The CNN feature tokens serve as query in a cross-attention operation on ViT tokens, and allow injecting the global contextual information into spatially localized representations. Adaptive feature scaling is used in training which also implicitly controls the magnitude of cross-modal gradients leading to optimization stability without explicit gradient normalization.

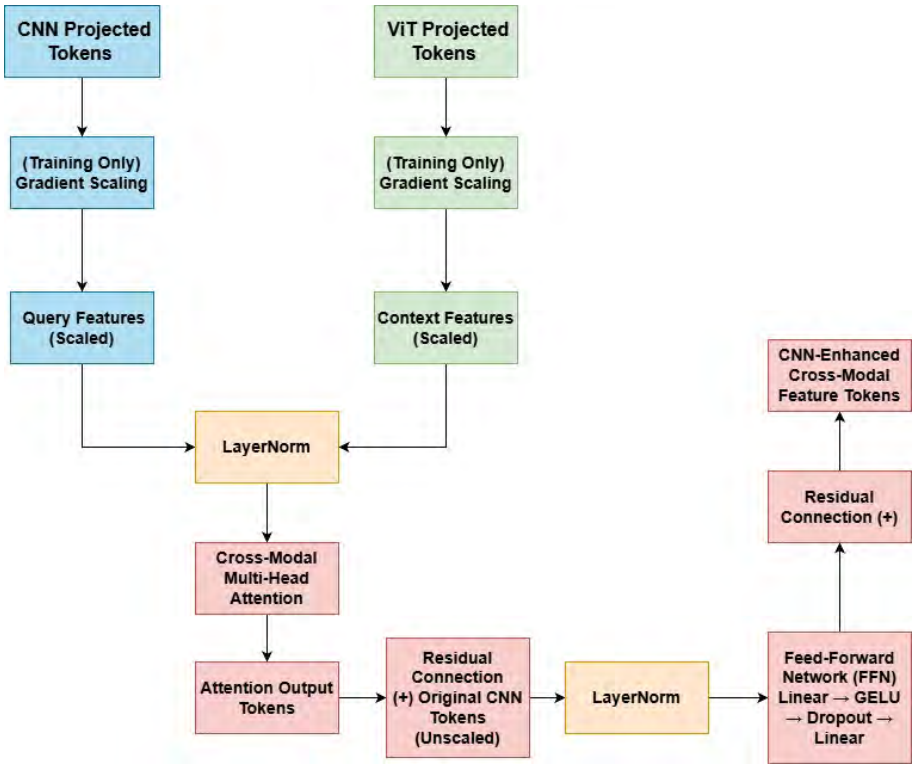


Figure 4.6: Proposed Gradient-Balanced Cross-Modal Bridge

After cross-modal interaction, the model has to combine spatial features into a small global representation that can be classified. Conventional global average pooling is a situation in which the spatial locations are equally important, which is not the case for clinical reality in which the pathological regions tend to be localized and dependent on classes. In order to eliminate this shortcoming, a Pathology-Aware Attention Pooling mechanism is utilized. In this module, learnable and class-conditional query vectors are introduced that consider spatial characteristics concerning particular types of diseases. Given a set of class-specific queries Q_k and feature keys K, attention weights are computed as in the given equation 4.39

$$A_k = \text{softmax}\left(\frac{Q_k K^\top}{\sqrt{d}}\right) \tag{4.39}$$

where d denotes the embedding dimension. The attention-weighted resulting features highlight the diagnostically significant areas and generate interpretable attention maps which conform to clinical expectations.

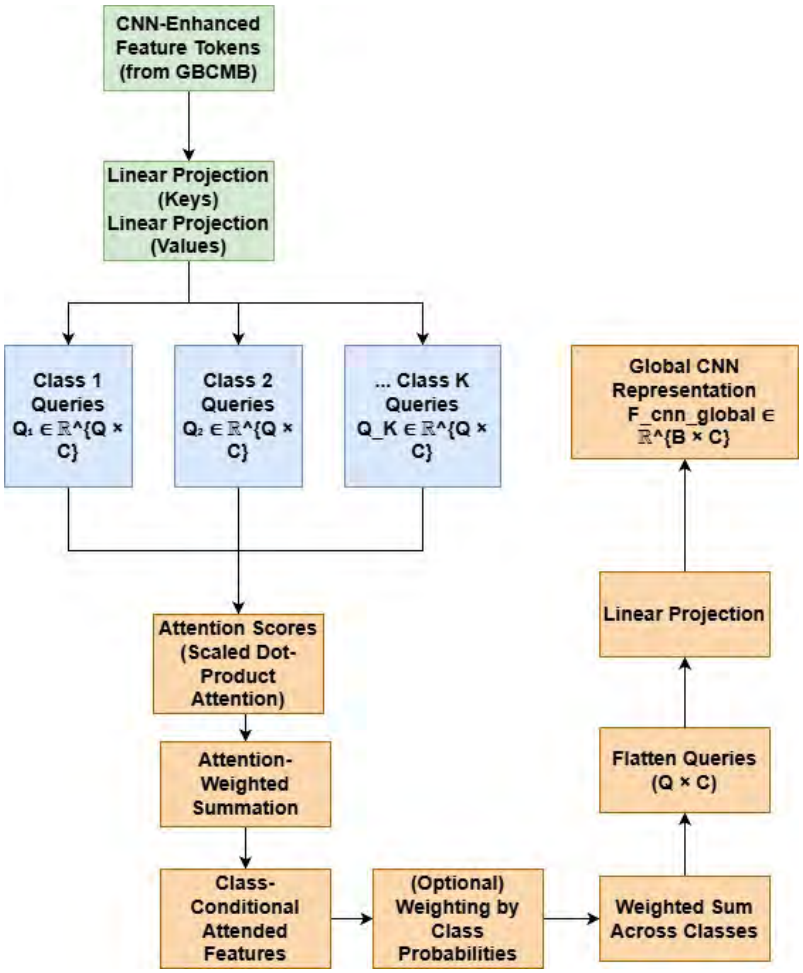


Figure 4.7: Proposed Pathology-Aware Attention Pooling

Figure 4.7 depicts that PAAP substitutes the traditional global pooling with class-conditional multi-query attention that selectively chooses the spatial features of the pathology-relevant regions.

Although both CNN and ViT branches play valuable roles, their relative contribution to information is valuable. Samples have different levels of importance. On ambiguous or blurry images usually strident local features can be more predictive, and in more distinct cases, global contextual reasoning may dominate. In order to allow this variability, an Uncertainty-Gated Feature. Mechanism of fusion is presented. The present module approximates a scalar uncertainty value $u \in [0, 1]$ of the composite feature representation and bases on it to adjust the adaptive weighting of inputs of every branch. The fused feature representation F_{fused} given in the equation 4.40 is computed as

$$F_{\text{fused}} = \alpha(u)F_{\text{cnn}} + \beta(u)F_{\text{vit}} \tag{4.40}$$

where $\alpha(u)$ and $\beta(u)$ are monotonic functions of the estimated uncertainty. Such an adaptive fusion approach allows input based reasoning that is especially crucial in the medical decision making scenarios.

To achieve the last classification step, a Compact Mixture-of-Experts (CMOE) classifier will be used instead of a single monolithic classifier. Task Medical classification

tasks are usually associated with inter-class complexities in which some class differences are more difficult compared to others. The decision boundaries are learned by one classifier and this makes the model more complex as well as prone to overfitting. The proposed design breaks this task into several lightweight network of experts each of which focuses on various aspects of the classification problem. A gating network attaches weights to input dependent on each expert and the eventual forecast is given as a weighted sum of expert outputs. The design encourages specialization with efficiency of parameters.

Even in an architecturally regularized model, deep models can still overfit on instance specific features instead of being able to learn features that are invariant across all classes. In order to counterbalance this threat, Intra-Class Variance Regularization (ICVR) mechanism is introduced in the training. This auxiliary goal makes features of the same class to clump closely around a centroid of the class in the embedding space. Officially, the regularization loss minimizes the mean squared error between feature vectors and centrals of the classes and thus encourages tight and discriminative class representations. This limitation is a supplement to conventional regularization methods like dropout and weight decay.

A key feature of the suggested design is that the interpretability would be created as a result of the architecture and not something that would be brought up as a post-hoc analysis. The pooling mechanism generates attention maps, the uncertainty estimates generated during fusion and expert selection probabilities generated by the classifier give several perspectives of understanding model behavior. This design option promotes openness and promotes confidence in clinical implementation conditions.

Finally, interpretability, adaptability, training stability, and representational power compromise purposefully in the proposed model design. The design has a problem focus and each design element is presented to solve a certain shortcoming noticed in the traditional methods. This pre-specification forms a platform on which the implementation and experimental evaluation in later chapters will be based.

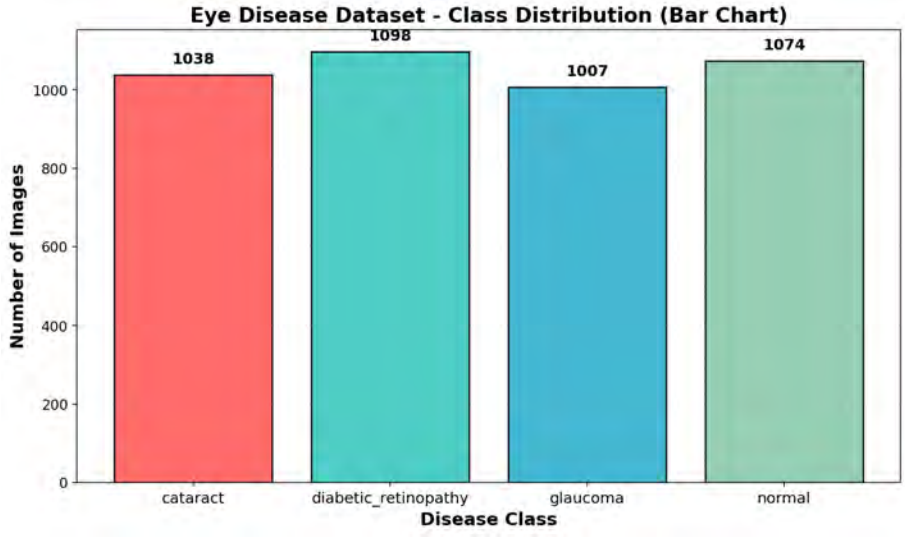
4.2.1 Datasets

This study utilizes two publicly available retinal fundus image datasets: the Eye Disease Classification (EDC) dataset and the AMDNet23 dataset. Each dataset serves a distinct purpose within the experimental pipeline, enabling both multi-disease classification analysis and dataset-level distribution assessment.

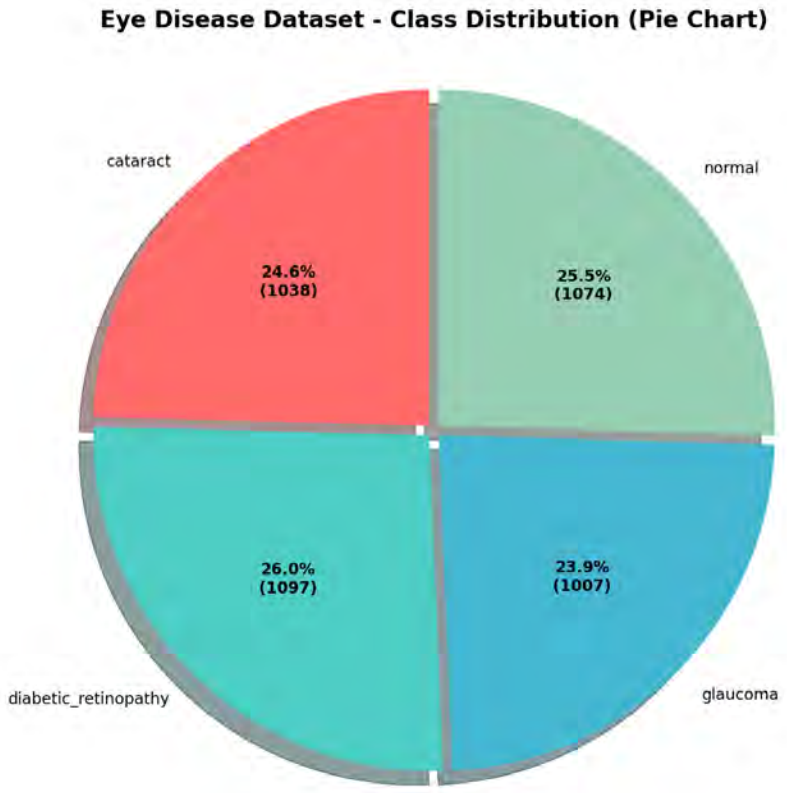
4.2.2 Eye Disease Classification (EDC) Dataset

In the study, a dataset which was publicly available was used which included images of retinal fundus. The Kaggle dataset of Eye Disease Classification [20] is made up of 4,217 color retinal fundus pictures split into four separate categories as shown in figure 4.9. The catagories are Normal (1,074 images), Diabetic Retinopathy (1,098 images), Cataract (1,038 images) and Glaucoma(1,007 images). All the images are coded with a single label of a class referring to one ocular condition. The balance

between images distribution is approximately even as depicted in Figure 4.8 and each with about 1,000 images of the different classes obtained by different sources which includes databases like IDRiD, HRF and Ocular Recognition. This diversity introduces differences in resolution, color and image device features which are to be standard pretraining model training.



(a) Class Distribution (Bar Chart)



(b) Class Distribution (Pie Chart)

Figure 4.8: Distribution of eye disease classes within the studied dataset

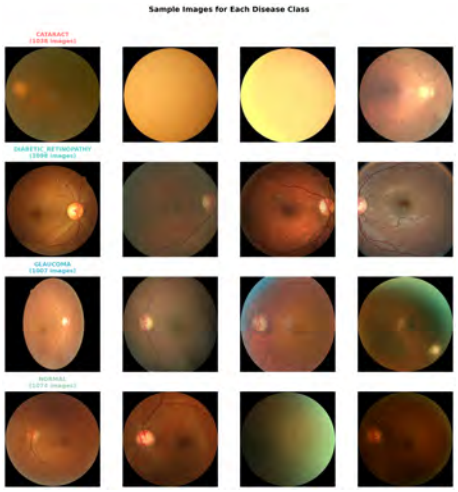
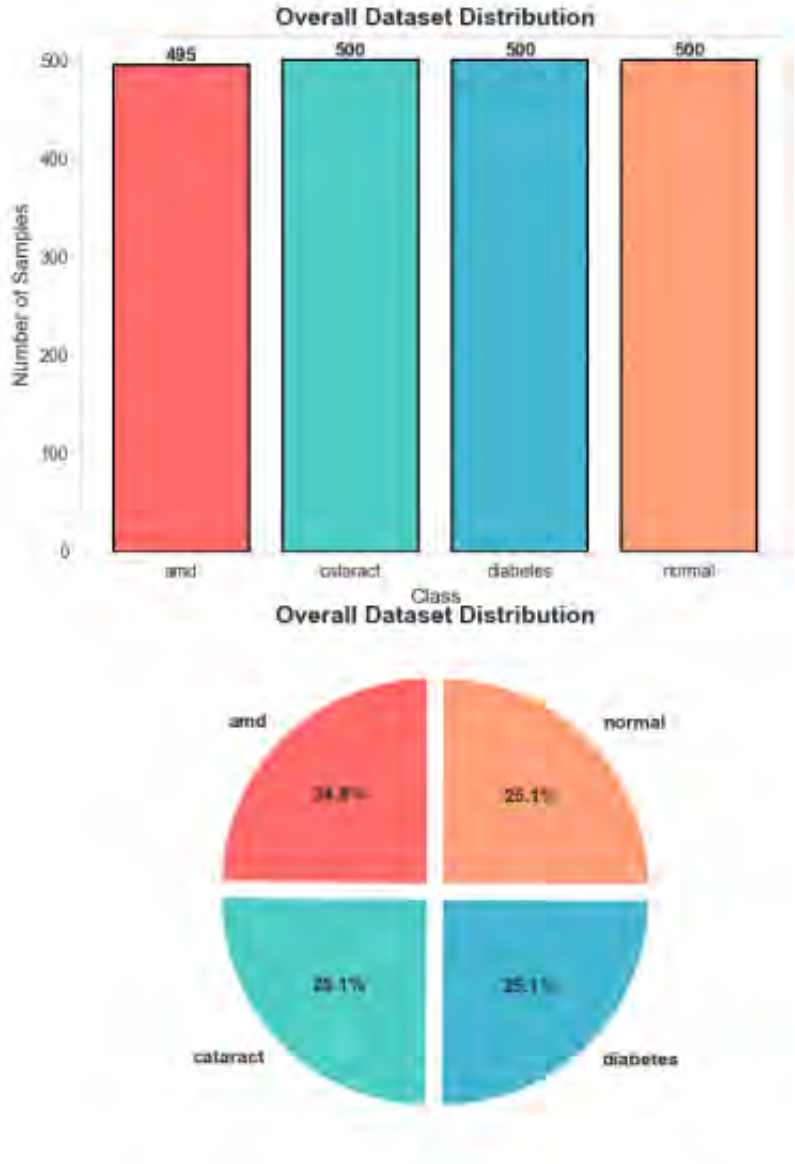
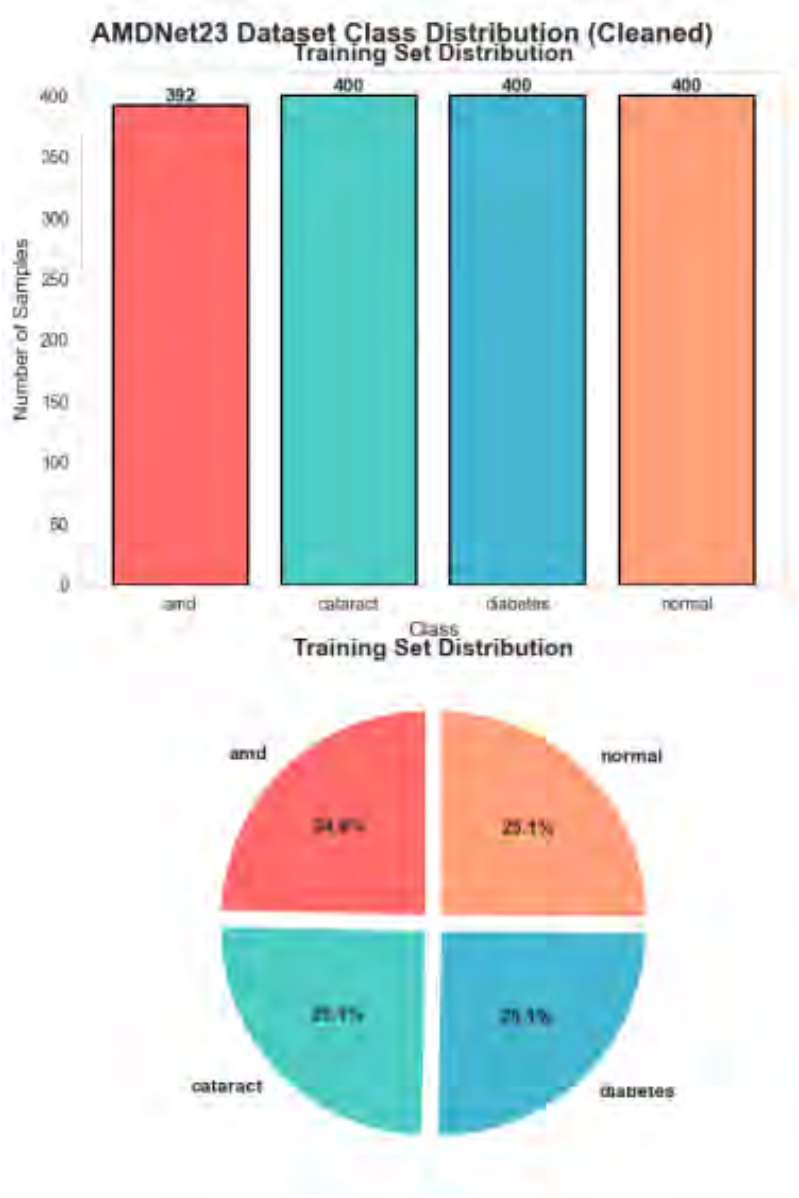


Figure 4.9: Sample retinal fundus images for each disease class: Cataract, Diabetic Retinopathy, Glaucoma, and Normal.

4.2.3 AMDNet23

The AMDNet23 dataset [47] is a curated collection of 2,000 high-quality retinal fundus images compiled from six publicly available sources, namely Ocular Disease Recognition, DR_200, Fundus Dataset, RFMiD, HRF, and ARIA. The dataset contains four balanced disease classes—Normal, Diabetic Retinopathy, Cataract, and Age-related Macular Degeneration (AMD)—with 400 images per class. All images were enhanced using advanced preprocessing techniques, including gamma correction and Contrast Limited Adaptive Histogram Equalization (CLAHE), to improve visual clarity and diagnostic relevance. AMDNet23 is designed to support research in deep learning–based ocular disease detection and multi-class fundus image classification.


Overall Dataset Distribution
Number of Samples
Class
amd
cataract
diabetes
normal
495
500
500
500
0
100
200
300
400
500
Overall Dataset Distribution
amd
normal
cataract
diabetes
25.1%
25.1%



sxdsxd

Figure 4.9: Class-wise distribution of the AMDNet23 dataset

4.2.4 Dataset Cleaning

To prevent data redundancy on the datasets and potential information leakage, the dataset was screened for duplicate and near-duplicate images using perceptual hashing. Each image was converted to grayscale and processed using a difference-hash–based method, enabling detection of visually identical or highly similar samples despite minor pixel-level variations. This procedure identified three duplicate groups, including cases where identical images appeared across different class directories on the EDC dataset and on the AMDNet23 Dataset 157 duplicate images were identified and removed.

Table 4.1: Dataset cleaning and quality control summary.

Check	Count
Duplicate images removed on EDC	3
Duplicate images removed on AMDNet23	157

4.2.5 Dataset Preprocessing

We train the fundus image image dataset in our pipeline with extensive data augmentation and preprocessing procedures that are medical imaging specific. The preprocessing includes CLAHE (Contrast Limited Adaptive Histogram Equalization) that is especially useful with fundus images due to the ability to improve local contrast to display blood vessels, lesions, and other pathological phenomena.

4.2.5.1 Training Augmentation Pipeline

For training, we apply the following augmentation sequence using the Albumentations library:

1. **CLAHE Enhancement**: Applied with clip limit of 2.0 and tile grid size of 8×8 with probability $p = 0.5$. This adaptive histogram equalization enhances local contrast while preventing over-amplification of noise.

2. **Random Resized Crop**: Images are randomly cropped and resized to 224×224 pixels with scale range $(0.8, 1.0)$ and aspect ratio range $(0.9, 1.1)$. This introduces translation and scale invariance.

3. **Geometric Transforms**: Horizontal flip ($p = 0.5$), vertical flip ($p = 0.3$), and rotation within ± 10 ($p = 0.5$) are applied to increase data variability.

4. **Color Augmentation**: Either ColorJitter (brightness, contrast, saturation ± 0.15, hue ± 0.05) or RandomBrightnessContrast (± 0.15) is applied with combined probability $p = 0.5$.

5. **Blur/Noise Injection**: Either Gaussian blur (kernel size 3–5) or Gaussian noise (variance 5–20) is applied with probability $p = 0.2$ to improve robustness.

6. **Coarse Dropout**: Random rectangular regions (1–6 holes, size 8–20 pixels) are filled with zeros with probability $p = 0.15$, acting as regularization.

7. **Normalization**: Images are normalized using ImageNet statistics (mean $= [0.485, 0.456, 0.406]$, std $= [0.229, 0.224, 0.225]$).

4.2.5.2 Validation/Test Preprocessing

For validation and testing, we apply a deterministic preprocessing pipeline:

1. **CLAHE Enhancement**: Applied with probability $p = 1.0$ to ensure consistent contrast enhancement across all evaluation images.

2. **Resize**: Images are resized to 231×231 pixels (approximately $1.03\times$ the target size).

3. **Center Crop**: A center crop of 224×224 pixels is extracted.

4. **Normalization**: Same ImageNet statistics as training.

4.2.6 Dataset Splitting

The dataset is split into training, validation, and test sets using stratified sampling to maintain class distribution across all splits. The split ratios are 70% for training, 15% for validation, and 15% for testing. Algorithm 1 presents the dataset splitting and loading procedure.

Algorithm 1 Dataset Splitting and Loading

Require: CSV file with image paths and labels, Image directory
Ensure: Training, Validation, and Test DataLoaders
Load dataset CSV into DataFrame D
Extract labels $y \leftarrow \text{factorize}(D[\text{'label'}])$
// **First split: 70% train, 30% temporary**
$idx_{train}, idx_{temp} \leftarrow \text{StratifiedSplit}(D, \text{test\_size} = 0.30, \text{stratify} = y)$
// **Second split: 50% validation, 50% test (of the 30%)**
$idx_{val}, idx_{test} \leftarrow \text{StratifiedSplit}(idx_{temp}, \text{test\_size} = 0.50, \text{stratify} = y[idx_{temp}])$
Create $\mathcal{D}_{train} \leftarrow \text{FundusDataset}(D[idx_{train}], \text{train\_transform})$
Create $\mathcal{D}_{val} \leftarrow \text{FundusDataset}(D[idx_{val}], \text{val\_transform})$
Create $\mathcal{D}_{test} \leftarrow \text{FundusDataset}(D[idx_{test}], \text{val\_transform})$
return DataLoader($\mathcal{D}_{train}$, batch_size=20, shuffle=True)
return DataLoader($\mathcal{D}_{val}$, batch_size=40, shuffle=False)
return DataLoader($\mathcal{D}_{test}$, batch_size=40, shuffle=False)

4.3 Implementation of Selected Design

This section is about the process description of the implementation of the proposed MuninFormer architecture to classify fundus images. Its implementation includes the full pipeline of data preprocessing, model training and visualization of explainability. All the elements were introduced in Python with PyTorch as the main deep learning platform.

4.3.1 Training Configuration

We configure a set of training hyperparameters to ensure stable optimization and effective generalization of the proposed MuninFormer model. The model is trained for a maximum of 300 epochs using a batch size of $B = 20$, which balances memory consumption and gradient stability for the dual-backbone architecture. Fundus images are resized to 224×224 pixels, and the task involves four disease categories: cataract, diabetic retinopathy, glaucoma, and normal.

The search algorithm uses a starting learning rate of 3×10^{-5} that is gradually increased in the initial 15 epochs to stabilize the initial training dynamics, especially with attention-based components. Learning rate is then scaled down to the lowest

value which is 1×10^{-6}. The weight decay along with the label smoothing are implemented to promote regularization to reduce the overfitting, whereas the gradient clipping is employed to avoid irregular updates. Early termination of training at 80 epochs of patience is taken when the performance ceases to increase. Also, convergence and generalization are further stabilized with exponential moving average (EMA) of model weights.

Parameter	**Value**	**Description**
Maximum Epochs	300	Upper limit for training iterations
Batch Size	20	Samples processed per training step
Initial Learning Rate	3×10^{-5}	Learning rate after warmup
Minimum Learning Rate	1×10^{-6}	Lower bound for learning rate decay
Warmup Epochs	15	Linear warmup period for stable training
Weight Decay	0.025	L2 regularization coefficient
Label Smoothing	0.08	Regularization to prevent overconfidence
Gradient Clip Norm	1.0	Maximum allowed gradient magnitude
Early Stopping Patience	80	Epochs without validation improvement
EMA Decay	0.9995	Exponential moving average coefficient

Table 4.2: Training configuration and hyperparameters

4.3.2 Model Architecture Hyperparameters

In our implementation, we have used the following architectural hyperparameters as summarized in Table 4.3. These values were obtained by systematic experimentation to obtain the best performance within the parameter budget constraint.

Parameter	Value	Description
Image Size	224×224	Input resolution for both backbones
CNN Backbone	MobileNetV3-Small	Lightweight CNN for local feature extraction
ViT Backbone	ViT-Tiny-Patch16	Vision Transformer for global context
Shared Embedding Dimension	192	Common projection space dimension
Number of Classes	4	Disease categories in fundus dataset
Dropout Rate	0.15	Applied in projection layers
Number of Experts (CMOE)	4	Specialized classifier networks
Expert Hidden Dimension	48	Internal dimension of each expert
PAAP Queries per Class	3	Learnable attention queries
Cross-Attention Heads	4	Multi-head attention in GBCMB

Table 4.3: Model Architecture Hyperparameters

4.3.3 Inspired Component Loss Weights

To balance the various loss terms in our multi-objective training, we configure specific weights for each auxiliary loss component. The Intra-Class Variance Regularization (ICVR) loss is weighted by $\lambda_{ICVR} = 0.05$, which encourages tight clustering of same-class features without overpowering the primary classification objective. The Mixture-of-Experts (MoE) load balancing loss is weighted by $\lambda_{MoE} = 0.1$, which promotes uniform utilization of all expert networks.
The total loss function is formulated as shown in the equation 4.41:

$$\mathcal{L}_{total} = \mathcal{L}_{CE} + \lambda_{ICVR} \cdot \mathcal{L}_{ICVR} + \lambda_{MoE} \cdot \mathcal{L}_{balance} \tag{4.41}$$

where $\mathcal{L}_{CE}$ is the cross-entropy classification loss with label smoothing of 0.08.

4.3.4 Model Architecture Implementation

The MuninFormer model is implemented as a PyTorch `nn.Module` class integrating all five inspired components. Figure 4.10 illustrates the complete architecture flow.

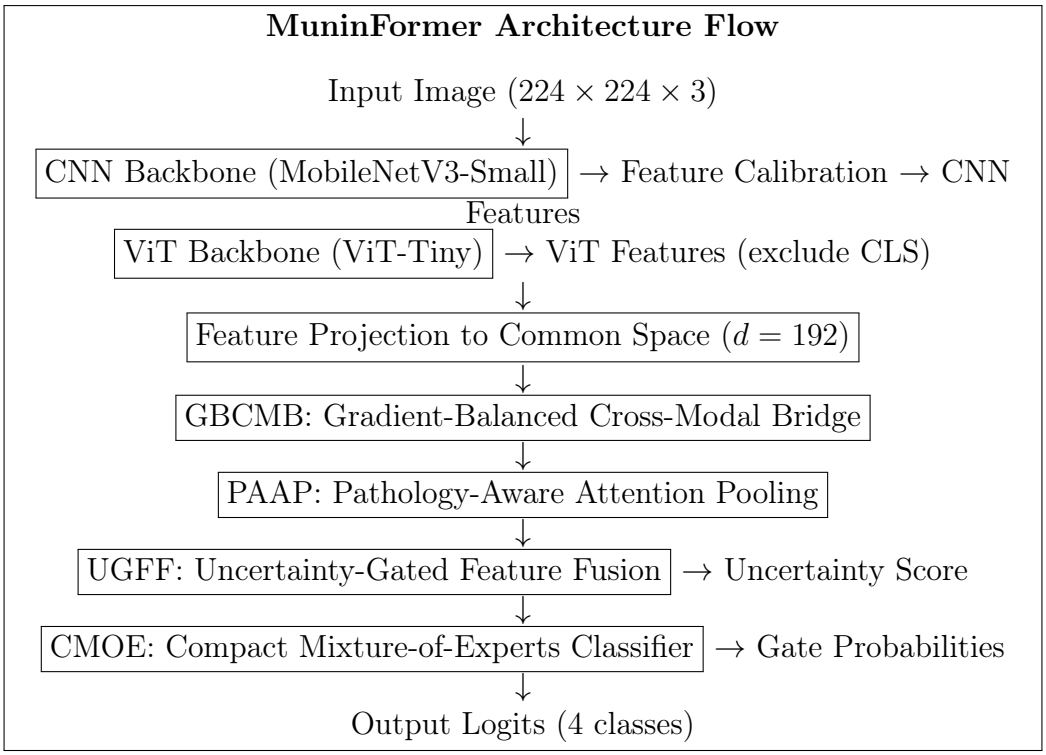


Figure 4.10: MuninFormer Architecture Flow Diagram

4.3.4.1 Backbone Integration

The pre-trained backbones are integrated using the `timm` library:

- **CNN Backbone**: MobileNetV3-Small is loaded with `features_only=True` and `out_indices=[4]` to extract the final convolutional feature maps without the classification head. A lightweight channel attention module (Squeeze-and-Excitation style) is applied for feature calibration.
- **ViT Backbone**: ViT-Tiny-Patch16-224 is loaded with `num_classes=0` to obtain transformer encoder outputs. The CLS token is excluded, and patch tokens are used as spatial features.

Both feature sets are projected to the common embedding dimension ($d = 192$) through linear layers followed by LayerNorm and Dropout.

4.3.5 Inspired Components Implementation

This subsection details the implementation of each inspired component in the MuninFormer architecture, providing both algorithmic descriptions and explanations of the underlying mechanisms.

4.3.5.1 Pathology-Aware Attention Pooling (PAAP)

PAAP replaces standard global pooling with class-conditional attention, enabling the model to focus on disease-specific regions in fundus images shown in 2. Unlike conventional pooling that treats all spatial locations equally, PAAP learns separate

attention patterns for each disease class, allowing the model to identify pathology-specific features such as hemorrhages for diabetic retinopathy or optic disc changes for glaucoma.

Algorithm 2 PAAP Forward Pass

Require: Input features $X \in \mathbb{R}^{B \times N \times C}$, Class queries $Q \in \mathbb{R}^{K \times Q \times C}$
Ensure: Pooled features $Z \in \mathbb{R}^{B \times C}$, Attention maps $A \in \mathbb{R}^{B \times K \times Q \times N}$

$K \leftarrow W_K \cdot X$ ▷ Key projection: $[B, N, C]$
$V \leftarrow W_V \cdot X$ ▷ Value projection: $[B, N, C]$
$Q_{exp} \leftarrow \text{expand}(Q, B)$ ▷ Expand queries: $[B, K, Q, C]$
// **Compute attention scores**
$A \leftarrow \text{softmax}\left(\frac{Q_{exp} \cdot K^T}{\sqrt{C}}\right)$ ▷ $[B, K, Q, N]$
// **Attend to values**
$Z_{attended} \leftarrow A \cdot V$ ▷ $[B, K, Q, C]$
// **Average across classes and queries**
$Z_{flat} \leftarrow \text{mean}(Z_{attended}, \text{dim} = 1)$ ▷ $[B, Q, C]$
$Z_{flat} \leftarrow \text{reshape}(Z_{flat}, [B, Q \times C])$
$Z \leftarrow W_{out} \cdot Z_{flat}$ ▷ Output projection: $[B, C]$
return Z, A

4.3.5.2 Uncertainty-Gated Feature Fusion (UGFF)

UGFF addresses the challenge of combining CNN and ViT features by learning to adaptively weight each modality based on estimated prediction uncertainty. When the model is uncertain about an input, it shifts the weighting toward the more robust CNN features; when confident, it can leverage the ViT's global context more heavily 3.

Algorithm 3 UGFF Forward Pass

Require: CNN features $F_{CNN} \in \mathbb{R}^{B \times C}$, ViT features $F_{ViT} \in \mathbb{R}^{B \times C}$
Ensure: Fused features $F_{fused} \in \mathbb{R}^{B \times C}$, Uncertainty $u \in \mathbb{R}^{B \times 1}$

$F_{concat} \leftarrow \text{concat}(F_{CNN}, F_{ViT})$ ▷ $[B, 2C]$
// **Estimate uncertainty**
$u \leftarrow \sigma(\text{MLP}_{uncertainty}(F_{concat}))$ ▷ $[B, 1]$, σ = sigmoid
// **Process each modality**
$F'_{CNN} \leftarrow \text{GELU}(\text{LayerNorm}(\text{Linear}(F_{CNN})))$
$F'_{ViT} \leftarrow \text{GELU}(\text{LayerNorm}(\text{Linear}(F_{ViT})))$
// **Compute adaptive weights**
$w_{CNN} \leftarrow 0.5 + 0.3 \times u$ ▷ Range: $[0.5, 0.8]$
$w_{ViT} \leftarrow 0.5 - 0.3 \times u$ ▷ Range: $[0.2, 0.5]$
// **Weighted fusion**
$F_{fused} \leftarrow W_{fusion} \cdot (w_{CNN} \odot F'_{CNN} + w_{ViT} \odot F'_{ViT})$
return F_{fused}, u

4.3.5.3 Gradient-Balanced Cross-Modal Bridge (GBCMB)

GBCMB addresses a critical training challenge in hybrid architectures: the gradient imbalance between CNN and ViT branches. During backpropagation, gradients flowing through cross-attention can have vastly different magnitudes for each modality, causing one branch to dominate learning. GBCMB introduces learnable scaling factors that automatically balance gradient flow during training 4.

Algorithm 4 GBCMB Forward Pass

Require: Query features $F_Q \in \mathbb{R}^{B \times N_Q \times C}$, Context features $F_C \in \mathbb{R}^{B \times N_C \times C}$
Ensure: Enhanced features $F_{out} \in \mathbb{R}^{B \times N_Q \times C}$

if training mode **then**
  $\gamma \leftarrow \text{clamp}(\gamma_{learnable}, 0.5, 2.0)$
  $F'_Q \leftarrow F_Q \times \gamma$ ▷ Scale query
  $F'_C \leftarrow F_C / \gamma$ ▷ Inverse scale context
else
  $F'_Q \leftarrow F_Q$; $F'_C \leftarrow F_C$
end if
// **Cross-attention**
$F_{norm} \leftarrow \text{LayerNorm}(F'_Q)$
$F_{attn}, \_ \leftarrow \text{MultiHeadAttention}(F_{norm}, F'_C, F'_C)$
$F_{res} \leftarrow F_Q + F_{attn}$ ▷ Residual connection
// **Feed-forward network**
$F_{out} \leftarrow F_{res} + \text{FFN}(\text{LayerNorm}(F_{res}))$
return F_{out}

4.3.5.4 Compact Mixture-of-Experts Classifier (CMOE)

CMOE replaces the standard single-head classifier with multiple specialized expert networks, each potentially learning to recognize different aspects of the input. A gating network is trained to send the inputs to the best suited experts so that specialization is achieved, and computational efficiency exists in algorithm 5.

Algorithm 5 CMOE Forward Pass

Require: Input features $X \in \mathbb{R}^{B \times C}$, Number of experts $E = 4$
Ensure: Logits $Z \in \mathbb{R}^{B \times K}$, Gate probabilities $G \in \mathbb{R}^{B \times E}$
1: // **Compute gating weights**
2: $G \leftarrow \text{softmax}(W_{gate} \cdot X)$ ▷ $[B, E]$
3: // **Get predictions from all experts**
4: **for** $e = 1$ to E **do**
5: $Z_e \leftarrow \text{Expert}_e(X)$ ▷ $[B, K]$, Expert: Linear-ReLU-Linear
6: **end for**
7: $Z_{stack} \leftarrow \text{stack}([Z_1, ..., Z_E])$ ▷ $[B, E, K]$
8: // **Weighted combination**
9: $Z \leftarrow \sum_{e=1}^{E} G[:, e] \times Z_{stack}[:, e, :]$ ▷ Einstein summation
10: // **Update expert usage statistics (training only)**
11: **if** training mode **then**
12: $\text{expert\_usage} \leftarrow 0.9 \times \text{expert\_usage} + 0.1 \times \text{mean}(G, \text{dim} = 0)$
13: **end if**
14: **return** Z, G

A critical challenge in mixture-of-experts models is "expert collapse," where the gating network learns to route all inputs to a single expert, leaving others unused. To prevent this, we track exponential moving averages of expert usage during training (Lines 11-13). The load balancing loss (defined in Section 4.3.3) penalizes deviation from uniform expert usage by the equation 4.42

$$\mathcal{L}_{balance} = 0.1 \times \text{mean}\left(\left|\text{expert\_usage} - \frac{1}{E}\right|\right) \tag{4.42}$$

The gate probabilities G returned by this module provide interpretability, showing which experts are most active for different inputs. This can reveal whether experts specialize for different disease types or image characteristics.

4.3.5.5 Intra-Class Variance Regularization (ICVR)

ICVR improves the discriminability of learned features by encouraging samples of the same class to cluster tightly in the feature space. This regularization supplements the cross-entropy loss which merely guarantees that the boundaries of classes are accurate but does not directly promote tight class clusters shown in algorithm 6.

Algorithm 6 ICVR Forward Pass

Require: Features $F \in \mathbb{R}^{B \times C}$, Labels $y \in \mathbb{R}^{B}$, Centroids $\mu \in \mathbb{R}^{K \times C}$
Ensure: ICVR loss $\mathcal{L}_{ICVR}$

if not training mode **then**
  return 0
end if
// **Update centroids with momentum**
for $k = 1$ to K **do**
  $mask_k \leftarrow (y == k)$
  if $\text{sum}(mask_k) > 0$ **then**
    $F_k \leftarrow F[mask_k]$ ▷ Features of class k
    $\bar{F}_k \leftarrow \text{mean}(F_k, \text{dim} = 0)$
    $\mu_k \leftarrow 0.9 \times \mu_k + 0.1 \times \bar{F}_k$ ▷ Momentum update
  end if
end for
// **Compute intra-class variance loss**
$\mathcal{L}_{ICVR} \leftarrow 0$
for $k = 1$ to K **do**
  $mask_k \leftarrow (y == k)$
  if $\text{sum}(mask_k) > 1$ **then**
    $F_k \leftarrow F[mask_k]$
    $d_k \leftarrow \|F_k - \mu_k\|_2^2$ ▷ Squared distances to centroid
    $\mathcal{L}_{ICVR} \leftarrow \mathcal{L}_{ICVR} + \text{mean}(d_k)$
  end if
end for
return $\mathcal{L}_{ICVR}/K$

4.3.6 Training Protocol

The entire training process incorporates all the elements that are well planned by optimization plans. The entire training loop is shown in algorithm 7.

Algorithm 7 MuninFormer Training Protocol

Require: Training data $\mathcal{D}_{train}$, Validation data $\mathcal{D}_{val}$, Configuration cfg
Ensure: Trained model M^*

Initialize model $M \leftarrow$ MuninFormer(num_classes $= 4$)
Initialize EMA model M_{EMA} with decay $= 0.9995$
Initialize optimizer $\mathcal{O} \leftarrow$ AdamW(M.params, $lr = 3 \times 10^{-5}$)
Initialize learning rate schedulers: CosineAnnealing, ReduceLROnPlateau
Initialize loss function $\mathcal{L}_{CE} \leftarrow$ CrossEntropyLoss(label_smoothing $= 0.08$)
best_acc $\leftarrow 0$, no_improve $\leftarrow 0$
for epoch $= 1$ to 300 **do**
    M.train()
    if epoch ≤ 15 **then**
        $lr \leftarrow 3 \times 10^{-5} \times \frac{\text{epoch}}{15}$
        Update $\mathcal{O}$ learning rate
    end if
    for each batch (X, y) in $\mathcal{D}_{train}$ **do**
        $\mathcal{O}$.zero_grad()
        $\hat{y}$, aux_data $\leftarrow M(X, \text{labels} = y)$
        $\mathcal{L}_{cls} \leftarrow \mathcal{L}_{CE}(\hat{y}, y)$
        $\mathcal{L}_{aux} \leftarrow$ aux_data['icvr_loss'] + aux_data['moe_balance_loss']
        $\mathcal{L}_{total} \leftarrow \mathcal{L}_{cls} + \mathcal{L}_{aux}$
        $\mathcal{L}_{total}$.backward()
        $\mathcal{O}$.step()
        Update EMA model: M_{EMA}.update(M)
    end for
    val_acc $\leftarrow$ evaluate($M_{EMA}, \mathcal{D}_{val}$)
    if val_acc $>$ best_acc **then**
        best_acc $\leftarrow$ *val_acc*
        $M^* \leftarrow M_{EMA}$.state_dict()
        no_improve $\leftarrow 0$
    else
        no_improve $\leftarrow$ no_improve $+ 1$
    end if
    if no_improve ≥ 80 **then**
        break
    end if
end for
return M^*

Explanation of Algorithm 7: TThe training process incorporates the best training deep neural networks by integrating training optimization techniques that guarantee stable convergence and good generalization.

4.3.7 Exponential Moving Average (EMA)

EMA given in equation 4.43 maintains a smoothed version of model weights updated at each training step:

$$\theta_{EMA}^{(t)} = \alpha \cdot \theta_{EMA}^{(t-1)} + (1 - \alpha) \cdot \theta^{(t)} \tag{4.43}$$

and $\alpha = 0.9995$ is the decay factor and $\theta^{(t)}$ are the current model weights. EMA has a number of advantages such as less noise in weight updates resulting in a more continuous convergence, which generally facilitates better generalization compared to raw training weights and is an implicit ensemble across training trajectory.

4.3.8 Evaluation Function

To evaluate the model performance based on various measures such as accuracy, macro F1-score,preciison,recall and confusion matrix that are applicable in medical image classification, we develop a holistic evaluation function.

Algorithm 8 Model Evaluation

Require: Model M, Test DataLoader $\mathcal{D}_{test}$, Class names
Ensure: Metrics dictionary containing accuracy, F1-score, precision, recall, confusion matrix

M.eval()
Initialize lists: all_preds $\leftarrow []$, all_targets $\leftarrow []$
loss_sum $\leftarrow 0$
for each batch (X, y) in $\mathcal{D}_{test}$ **do**
    with no_grad():
        $\hat{y}, \_ \leftarrow M(X)$
        loss_sum $\leftarrow$ loss_sum + CrossEntropy$(\hat{y}, y) \times |X|$
        all_preds.append(argmax$(\hat{y}, \text{dim} = 1)$)
        all_targets.append(y)
end for
$y_{pred} \leftarrow$ concatenate(all_preds)
$y_{true} \leftarrow$ concatenate(all_targets)
// **Compute metrics**
accuracy $\leftarrow \frac{\sum(y_{pred} == y_{true})}{|y_{true}|}$
f1 $\leftarrow$ f1_score$(y_{true}, y_{pred}, \text{average} = \text{'macro'})$
precision $\leftarrow$ precision_score$(y_{true}, y_{pred}, \text{average} = \text{'macro'})$
recall $\leftarrow$ recall_score$(y_{true}, y_{pred}, \text{average} = \text{'macro'})$
cm $\leftarrow$ confusion_matrix(y_{true}, y_{pred})
return {'acc' : accuracy, 'f1' : f1, 'precision' : precision, 'recall' : recall, 'cm' : cm}

4.3.9 Explainability Implementation

The explainability module provides visualization methods for model interpretation, enabling clinicians to understand the model's decision-making process.

4.3.9.1 GradCAM++ for CNN Features

GradCAM++ is implemented using forward and backward hooks to capture activations and gradients from the CNN backbone's final convolutional block.

Algorithm 9 GradCAM++ Visualization

Require: Model M, Input image X, Target class c (optional)
Ensure: Heatmap $H \in \mathbb{R}^{224\times224}$

Register forward hook on target layer to capture activations A
Register backward hook on target layer to capture gradients G
X.requires_grad $\leftarrow$ True
$\hat{y}, \_ \leftarrow M(X)$
if c is None **then**
    $c \leftarrow \text{argmax}(\hat{y})$
end if
// **Backward pass for class** c
one_hot $\leftarrow$ zeros_like($\hat{y}$); one_hot$[0, c] \leftarrow 1$
$\hat{y}$.backward(gradient = one_hot)
// **GradCAM++ weighting**
$G^2 \leftarrow G^2$; $G^3 \leftarrow G^3$
$\alpha_{num} \leftarrow G^2$
$\alpha_{denom} \leftarrow 2 \times G^2 + \sum_{h,w}(A) \times G^3 + \epsilon$
$\alpha \leftarrow \alpha_{num}/\alpha_{denom}$
$w \leftarrow \sum_{h,w}(\alpha \times \text{ReLU}(G))$ ▷ Weights per channel
$H \leftarrow \text{ReLU}(\sum_c(w \times A))$ ▷ Weighted combination
$H \leftarrow \text{interpolate}(H, \text{size} = (224, 224))$
$H \leftarrow \text{normalize}(H)$
$H \leftarrow H^{0.6}$ ▷ Gamma correction
return H

4.3.10 Model Integration

The complete MuninFormer model was implemented by integrating all components into a single PyTorch `nn.Module`. Table 4.4 summarizes the final model configuration.

Table 4.4: Final Model Configuration

Parameter	**Value**
Total Parameters	7.36 Million
Input Size	$224 \times 224 \times 3$
Embedding Dimension	192
Number of Classes	4
PAAP Queries per Class	3
CMOE Experts	4 (hidden dim: 48)
Cross-Attention Heads	4
Dropout Rate	0.15

The model architecture follows the data flow: Input → Dual Backbones → Feature Projection → GBCMB (bidirectional) → PAAP → UGFF → CMOE → Output Logits.

4.3.10.1 Parameter Count Verification

The model parameter count was verified to meet the design target of 8–10 million parameters:

Table 4.5: Parameter Count by Component

Component	**Parameters**
MobileNetV3-Small Backbone	1,529,968
ViT-Tiny Backbone	5,717,416
Feature Projections	73,856
PAAP Module	111,060
UGFF Module	112,132
GBCMB Module	296,832
CMOE Classifier	38,404
Lightweight Calibration Module	12,288
Other Components (norms, dropouts, etc.)	15,044
Total	**7.36M**

The total of approximately 7.36 million parameters falls within the design target range of 8–10 million, confirming successful implementation of a compact yet capable architecture.

4.3.11 Final Model Specification

Table 4.6 summarizes the final model specifications achieved through our implementation.

Table 4.6: Final Model Specifications

Specification	**Value**
Total Parameters	7.36 Million
Trainable Parameters	7.357 Million
Input Size	$224 \times 224 \times 3$
Output Classes	4
Inspired Components	PAAP, UGFF, GBCMB, CMOE, ICVR
Backbone Networks	MobileNetV3-Small + ViT-Tiny
Framework	PyTorch 2.0+
Training Hardware	NVIDIA CUDA GPU,Kaggle Cloud (NVIDIA Tesla T4)

The implementation achieves the design target of approximately 8–10 million parameters while incorporating all five inspired components for enhanced medical image classification performance. This modular architecture enables the verification of every component separately and the end-to-end training of the entire system with the standard deep learning practices.

Chapter 5

Result Analysis

5.1 Performance Evaluation

This section discusses the evaluation criteria used to assess the performance of the model, focusing on key metrics such as accuracy, precision, recall, and the confusion matrix. The testing methods used to evaluate the model's performance include the train-test split.

5.1.1 Evaluation Metrics

In order to fully assess the performance of the proposed model and the baseline models, we used a set of evaluation metrics that are widely applied to multi-label classification, especially to the medical analysis of images. As one fundus image can have more than one concurrent ocular pathology, the measures of evaluation are not adequate through the conventional use of one label per image. Thus, threshold-dependent and threshold-independent measures were employed to identify various model performance characteristics in the entire range of diseases.

All performance measures were calculated with the help of macro-averaging in order to cope with the asymmetric disease labeling. In this method, the metrics are initially computed separately in terms of each class and averaged between all the classes in equal measure. This allows the contribution of every disease to the total evaluation of the dataset, irrespective of the frequency of occurrence in the dataset.

Besides predictive performance, we also provide various measures of model complexity to determine how well it is computed. These are the total number of trainable parameters, the number of floating-point operations (FLOPs), and the size of storage of the model. These measures can be used together to give an understanding of the resource needs of each model.

Through the reporting of both complexity and performance measures, we can compare the performance and complexity measures in a balanced and more balanced way that takes into account not only the predictive accuracy but also the computational feasibility. The overall review is crucial to the assessment of the appropriateness of the models to clinical implementation in the real world, where resource efficiency is a major factor of concern.

5.1.1.1 Accuracy

Accuracy is the percentage of the right classification of samples out of all evaluated samples. It is calculated by the following equation 5.1

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

with TP (true positives) indicating how many of the positive samples were correctly predicted, TN (true negatives) indicating how many of the negative ones were correctly predicted, FP (false positives) indicating how many negative ones were falsely identified as positive, and FN (false negatives) indicating how many positive ones were falsely identified as negative.

Accuracy is an overall measure of the classification performance which is intuitive and measures the fraction of correct prediction. But in class-imbalanced datasets, accuracy can be misleading because high scores can be obtained by promoting the majority class. As such, accuracy is claimed as well as class-balanced measures such as macro-F1 to maintain a more holistic analysis.

5.1.1.2 Precision and Recall

Precision and recall evaluate a classifier's performance with respect to positive predictions and actual positive samples, respectively. Precision is defined in the equation 5.2:

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{5.2}$$

where TP (true positives) denotes the number of correctly predicted positive samples and FP (false positives) represents negative samples incorrectly classified as positive.

Recall measures the model's ability to correctly identify all positive samples and is defined in the following equation 5.3:

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{5.3}$$

where FN (false negatives) denotes the number of positive samples incorrectly classified as negative.

Precision reflects how reliable positive predictions are, while recall indicates how effectively the model captures all relevant positive cases. Recall is specifically significant in the medical image classification to reduce missed diagnoses and precision respectively are used to minimize false alarms.

5.1.1.3 Macro-F1 Score

The Macro-F1 score determines the performance of classification by computing the F1-score of each individual class and averaging them (without weighting). In each of the classes k, the F1-score is given in the following equation 5.4:

$$\text{F1}_k = \frac{2 \cdot \text{Precision}_k \cdot \text{Recall}_k}{\text{Precision}_k + \text{Recall}_k} \tag{5.4}$$

where Precision_k and Recall_k denotes the precision and recall for class k.
The Macro-F1 score is then computed by the following equation 5.5:

$$\text{F1}_{\text{macro}} = \frac{1}{K} \sum_{k=1}^{K} \text{F1}_k \tag{5.5}$$

in which K is the number of classes altogether.
Macro-F1 is very useful in determining the performance on skewed data sets, as it gives equal weight to all classes without weighting them, and is therefore more appropriate in that scenario. where is the total number of classes.

5.1.1.4 Per-Class Performance Metrics

Besides aggregate scores, per-class precision, recall and F1 scores are also used. has been reported to give fine-grained details about performance in classes. Such an analysis is especially significant in the medical use, where other applications should be distinguished classification errors may have clinical implications of various types, including incorrect diagnoses.

5.1.1.5 Confusion Matrix

Prediction errors are analyzed in more detail using a confusion matrix $\mathbf{C} \in \mathbb{R}^{K \times K}$. $C_{i,j}$ is the number of the samples of class i that are expected to be of class j. A correct prediction is diagonal and the incorrect prediction is the off-diagonal that emphasize examples of confusion between certain pairs of classes.

5.1.1.6 Model Complexity Metrics

The effectiveness of the suggested model besides predictive performance is in addition. It is measured in terms of the number of parameters, the computational complexity in GFLOPs, and model size in mega bytes (MB). The metrics give an idea on the viability of the implementation of the model in resource-constrained environments.

Number of Parameters: The count of parameters indicates the overall number of weights which can be learned in the model such as convolutional kernels, linear projection matrices and bias values. It is computed in the equation 5.6:

$$\text{Params} = \sum_{l=1}^{L} |\mathbf{W}_l| \tag{5.6}$$

In which, $\mathbf{W}_l$ represents the space of learnable parameters in layer l, and L represents the number of layers. A smaller number of parameters suggests the smaller size of the model and fewer memory requirements.

GFLOPs: The GFLOPs is a unit of measure of computational complexity, which is a count of giga floating point operations. The quantity of arithmetic operations

needed to process a single input while forward pass is being initiated shown in equation 5.7.

$$\text{GFLOPs} = \frac{\text{Total FLOPs per forward pass}}{10^9} \tag{5.7}$$

The lower GFLOPs translate to the lower cost of computation, faster inference, and in better adavantages for real time or edge device deployment scenarios.

Model Size: Model size, which is reported in megabytes (MB) is the size of memory which is needed to store the model parameters. It is computed by the following equation 5.8:

$$\text{Model Size (MB)} = \frac{\text{Params} \times b}{8 \times 10^6} \tag{5.8}$$

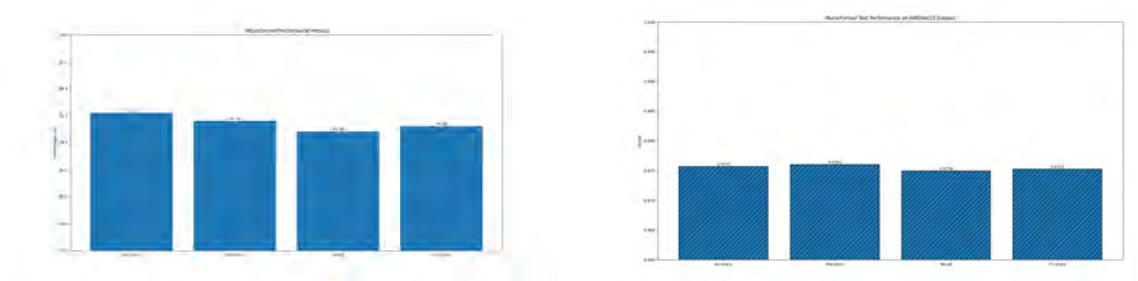
where b denotes the number of bits used to represent each parameter (e.g., $b = 32$ for single-precision floating-point). Smaller model sizes can enable easy storage, load faster and run on devices with small memory storage.

5.1.2 Quantitative Results and Analysis

The suggested MuninFormer model shows good and high performance within both evaluation datasets. As in Table 5.1, MuninFormer is an achiever of precision of 94.31% on the EDC dataset [20], with a close value (94.28%) of precision, recall (94.24%), and F1-score (94.26%), which represents equal classification behavior. The model also performs better on the AMDNet23 dataset [47] achieving an accuracy of 97.57%, the precision of 97.61%, the recall of 97.50% and F1-score are 97.53%. Such a steady consistency between accuracy and recall are especially significant to medical diagnosis, where false positive as well as false negative can occur critical clinical implications. The findings indicate the strength of the findings and the generationalness.generalization ability of MuninFormer on the various fundus imaging standards. A figure 5.1 showing their performance metric in barchart has also been provided.

Dataset	**Accuracy (%)**	**Precision (%)**	**Recall (%)**	**F1-score (%)**
EDC [20]	94.31	94.28	94.24	94.26
AMDNet23 [47]	97.57	97.61	97.50	97.53

Table 5.1: Quantitative metric performance of **MuninFormer** model on EDC and AMDNet23 datasets



(a) EDC dataset (b) AMDNet23 dataset

Figure 5.1: Final test performance of **MuninFormer** on (a) EDC and (b) AMDNet23 datasets

Table 5.2 and table 5.3 reveals that there is a high-level of class-wise discrimination with the proposed MuninFormer model in both datasets. Diabetic Retinopathy in EDC dataset is of high quality with a high accuracy, recall, and F1-score of 99.39, which implies it is highly reliable to identify it and is essential in clinical intervention to be done at an early age. The classification of cataracts is also strong with a precision of 97.38, and a recall of 95.51. On the contrary, Glaucoma and the Normal classes perform relatively poorly on EDC with the recall of 91.39 and the precision of 89.57, respectively. These findings indicate that there are difficulties that are related to delicate early-stage visual patterns and feature overlap between the healthy and mildly affected cases. MuninFormer shows remarkably good results on all categories of the AMDNet23 dataset, with almost perfect recall in AMD (100%), and perfect precision in Cataract (100%), whereas Diabetes and Normal classes have F1-scores of over 95. Altogether, the findings suggest that MuninFormer is effective in generalization between datasets and provides opportunities to improve it by introducing better class balancing, disease-specific features learning, and inclusion of multi-scale clinical cues.

Class	**Precision (%)**	**Recall (%)**	**F1-score (%)**
Cataract	97.38	95.51	96.44
Diabetic Retinopathy	99.39	99.39	99.39
Glaucoma	90.78	91.39	91.08
Normal	89.57	90.68	90.12

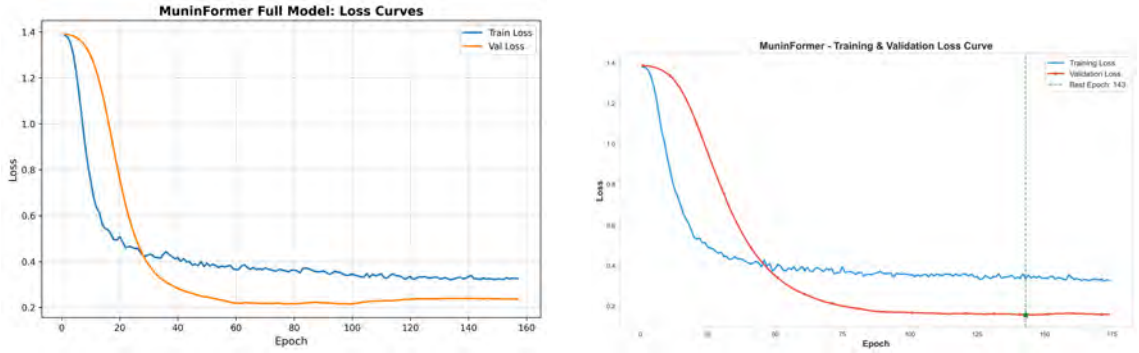
Table 5.2: Per-class performance of MuninFormer on EDC dataset [20].

Class	**Precision (%)**	**Recall (%)**	**F1-score (%)**
Cataract	100.00	99.00	99.50
AMD	97.39	100.00	98.68
Diabetes	97.89	93.00	95.38
Normal	95.15	98.00	96.55

Table 5.3: Per-class performance of MuninFormer on the AMDNet23 dataset [47].

Figure 5.2 shows the trainer loss convergence curve and validation loss convergence curve of the proposed MuninFormer model between the EDC and AMDNet23

datasets. The two scenarios have a comparable training and validation loss of approximately 1.4 that implies that both have low initial discriminative ability prior to feature learning.The curves follow a sharp upward trend during the initial training, and the training loss reduces faster than the validation loss, which points to the extraction of features and dynamic optimization.



(a) EDC dataset (b) AMDNet23 dataset

Figure 5.2: Training and validation loss curves of the proposed **MuninFormer** model on EDC and AMDNet23 datasets

In the case of EDC dataset, training loss stably approaches 0.33 during later epochs, whereas the validation loss approaches with smooth and little variation which implies that overfitting is controlled and there is good generalization. The near intersection of the two curves during the training process is an indication of balanced learning behaviour and good regularisation.

The same convergence trend is found on the AMDNet23 dataset where a more pronounced and sustained reduction in validation loss is found. The minimum loss is achieved at epoch 143 that is taken as the best checkpoint to be evaluated finally. The convergence and gap between the training and validation curves is also smooth, which is an additional factor that demonstrates that the learning is stable and has a high level of generalization on unknown data.

In general, these loss curves indicate that in both datasets, MuninFormer could converge reliably and remain stable when trained longer and does not overfit too much, which shows how well this model can be used in medical image classification.

The computational efficiency of the proposed MuninFormer model is analyzed to assess its suitability for resource-constrained deployment. As shown in Table 5.4, the model requires only 1.18 GFLOPs per forward pass at $224 \times 224 \times 3$ input resolution, significantly lower than conventional deep convolutional and transformer architectures that often exceeds from 5-15 GFLOPs. This low computational cost enables smooth inference on devices with limited processing capabilities without requiring specialized hardware acceleration.

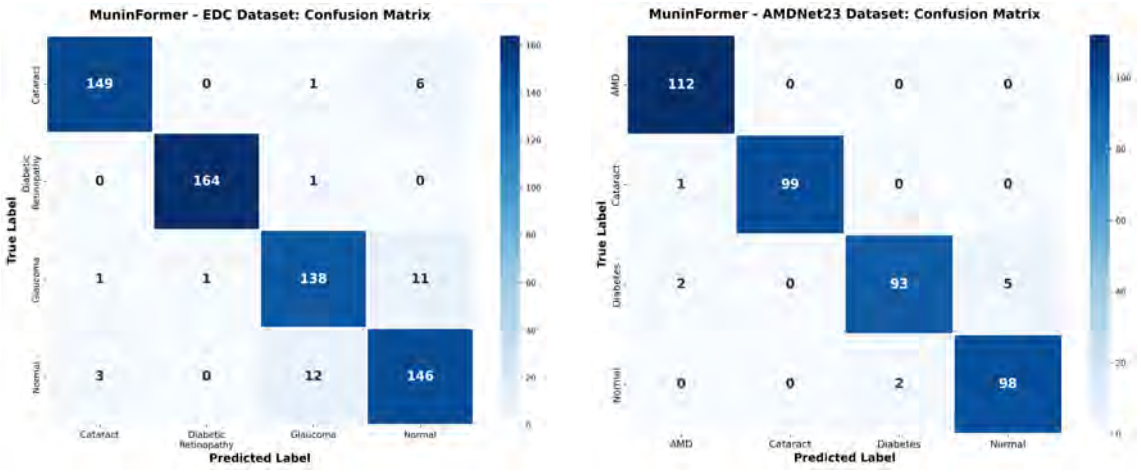
Metric	Value
Input Resolution	$224 \times 224 \times 3$
Trainable Parameters	7.36 M
Model Size	28.11 MB
GFLOPs (per forward pass)	1.18 G
Inference Time (ms)	11.13 ms

Table 5.4: Computational complexity and memory footprint of the proposed **MuninFormer** model

The model's memory footprint is equally compact at 28.11 MB with 7.36 million parameters, facilitating efficient storage and faster loading on edge devices. The inference time of 11.13 ms per image enables near real-time processing suitable for time-sensitive clinical applications.

The combination of low GFLOPs, compact model size, and fast inference demonstrates a favorable balance between computational efficiency and diagnostic accuracy, making MuninFormer well-suited for deployment on mobile devices, portable diagnostic tools, and real-time clinical decision-support systems.

Figure 5.3 indicates the confusion matrices of the suggested MuninFormer model on EDC and AMDNet23 test sets. Both cases have high levels of diagonal dominance, which is a sign of credible separation of classes into disease categories. On EDC dataset, Diabetic Retinopathy has almost perfect recognition (164/165 correct) which shows that the model can take into account certain pathological characteristics. Cataract is also very strong (149/156 correct) and most misclassifications are detected as Normal possibly because of global blur patterns and not local lesions.



(a) EDC dataset (b) AMDNet23 dataset

Figure 5.3: Confusion matrices of the proposed **MuninFormer** model on the EDC and AMDNet23 test sets

The majority of the errors on EDC can be observed between the Glaucoma and the normal classes, where minor structural indicators like early optic disc alterations are harder to differentiate. Conversely, the AMDNet23 data is more separated in

terms of classes, with AMD having a perfect recall and low ambiguity between classes. Generally, the confusion matrices suggest that the MuninFormer is trained to learn clinically meaningful representations, and wrongfully identified examples mainly consist of the similar image classes, which opens future research directions by using disease-specific supervision and learning features based on optic discs in particular.

Table 5.5 summarizes the class-wise AUC performance of the proposed MuninFormer model on the EDC and AMDNet23 datasets. On the EDC dataset, MuninFormer achieves a high macro-average AUC of 0.9886, indicating strong overall class discrimination. Diabetic Retinopathy exhibits perfect separability (AUC = 1.0000), while Cataract also demonstrates strong discriminative capability (AUC = 0.9924). Comparatively lower AUC values are observed for Glaucoma (0.9853) and Normal (0.9727), which is consistent with previously observed confusion between these visually similar categories.

Class	**AUC (AMD)**	**AUC (EDC)**
AMD (Age-related Macular Degeneration)	1.0000	–
Cataract	1.0000	0.9924
Diabetic Retinopathy	0.9996	1.0000
Glaucoma	–	0.9853
Normal	0.9997	0.9727
Macro-average	0.9999	0.9886

Table 5.5: Comparison of ROC–AUC scores for AMD and EDC datasets (one-vs-rest).

On the AMDNet23 dataset, MuninFormer demonstrates near-perfect discrimination across all classes, achieving a macro-average AUC of 0.9999. All disease categories record AUC values close to or equal to 1.0000, reflecting highly separable feature representations and strong generalization performance. Overall, the consistently high AUC scores across both datasets confirm the robustness of MuninFormer in distinguishing both pathological and normal fundus conditions, while highlighting the increased difficulty of Glaucoma-related discrimination in more heterogeneous datasets such as EDC.

5.1.3 Explainability and XAI Result Analysis

5.1.3.1 Qualitative Analysis of Model Attention

Figure 5.4 compares qualitatively explainability outputs of the three interpretability mechanisms of cataract classification based on three CNN branches, Grad-CAM++, attention rollout based on the Vision Transformer branch.

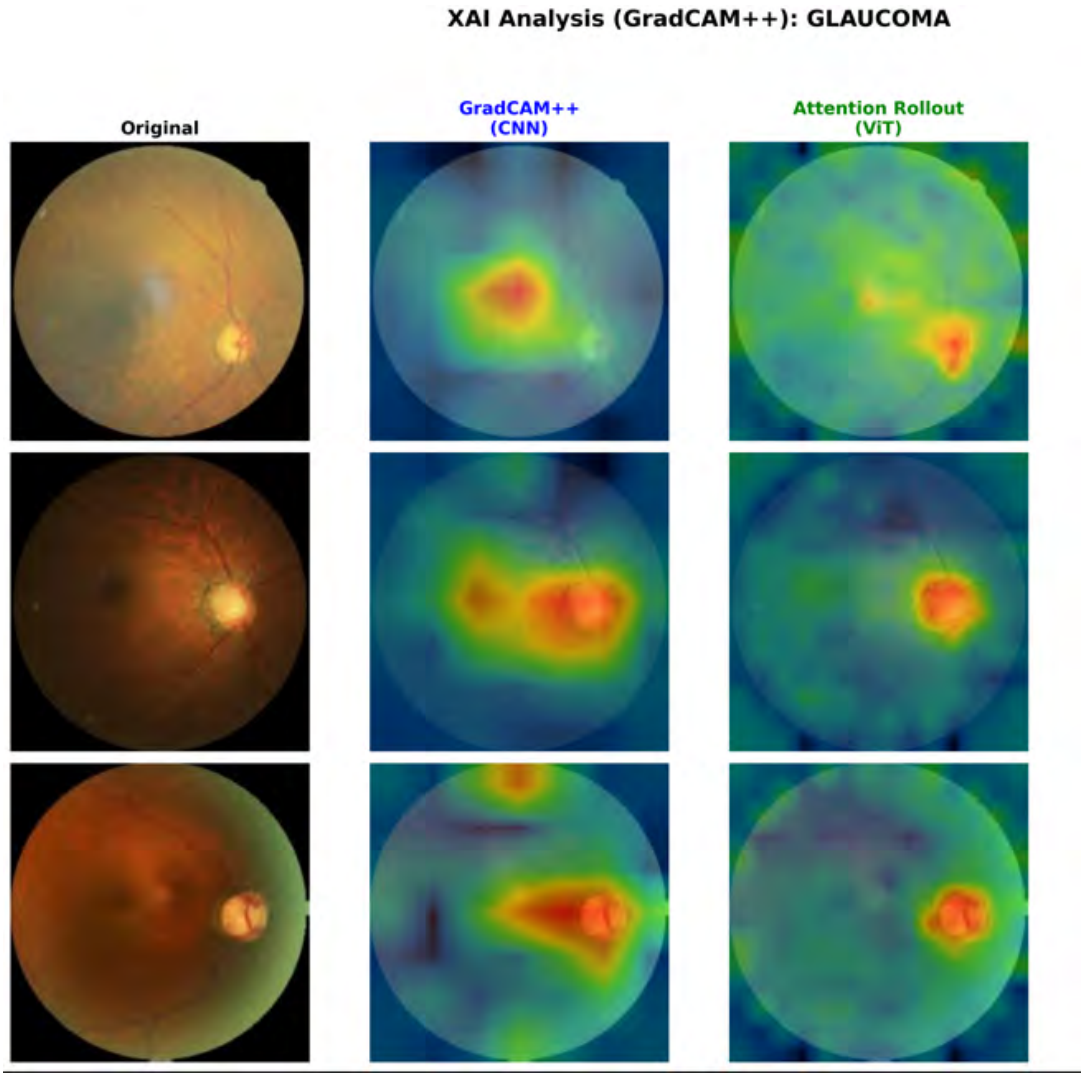


Figure 5.4: Qualitative results of explainability outputs of glaucoma classification with Grad-CAM++ (CNN), attention rollout (ViT).

The visualizations of the Grad-CAM++ highlight local high-activation points mainly on dense patterns of opacification and shadows at the lens. Although such maps are effective in extracting local cues relevant to cataract formation, the patterns of attention seem to be fragmented and sometimes they are vulnerable to background light, indicative of the locality bias of convolutional neural networks.

Compared to, Vision Transformer branch attention rollout results contain smoother and more globally distributed attention maps. These visualizations point to the fact that the transformer serves wider retinal structures and not individual areas. Nevertheless, the resulting attention is usually diffuse and less likely to be focused on clinically significant cataract areas which can lower its suitability in localized pathological interpretation.

5.2 Analysis of Design Solutions

This chapter gives a detailed review of the suggested MuninFormer architecture, in terms of accuracy, efficiency, feasibility, and clinical applicability. The discussion shows the strengths and limitations of the model based on ample experimental findings and the ablation study done.

5.2.1 Solution Overview

MuninFormer is a lightweight hybrid CNN-Transformer architecture, which tackles the major clinical needs, such as high diagnostic performance, parameter performance, explainability, and strong multi-class separation between cataract, diabetic retinopathy, glaucoma and normal images. The model combines complementary feature extractor pathways through the MobileNetV3-Small as backbone, which corresponds to local pathological information capture, and the Vision Transformer Tiny, which corresponds to the global capture of the spatial relationships of fundus images. To improve performance and clinical interpretability, Muninformer encompasses a number of new elements, such as gradient-balanced cross-modal fusion to integrate CNN-ViT stably, uncertainty-regulated fusion to make decisions adaptively, a simple mixture-of-experts classifier to specialize to pathology, and intraclass variance regularization to require discriminative and compact representations.

5.2.2 Deployment Feasibility

The model can be highly implemented on the clinical workstations and cloud systems and can be easily implemented on edge and mobile devices with optimization. It is possible to train with a single consumer-level GPU and inference can be performed by a CPU. MuninFormer works with standard RGB fundus images and gives the probability of classes, scores of uncertainty, and maps of spatial attention. The pipeline for deployment assists with the transformation to optimized inference formats to facilitate customization into clinical processes.

5.2.3 Strengths of the Proposed approach

MuninFormer is a hybrid CNN-Transformer architecture that has very high parameter efficiency, allowing the model for both local and global features accurately and can be still deployed on small-scale hardware. Through the attention maps it provides strong clinical explainability. This model has an impressive state-of-the-art performance capability, and it can provide steady and replicable training, detect diabetic retinopathy, classify multiple diseases, infer in real-time, and has a low false negative rate, which is highly important in patient safety.

5.2.4 Limitation and Weakness

MuninFormer also possesses a number of technical and clinical drawbacks, such as slower inference as compared to pure CNN models when using attention mechanisms and a higher complexity in deploying the model on ultra-low-power devices. The deterministic input resolution can limit the detection of fine details and the use of two CNN-ViT pathways can require a larger model than a single-path one, but

knowledge distillation may overcome this. The domain of some cases has a variation in performance in comparison to disease classes, and generalization can be influenced by a change in imaging devices or population, which implies that domain adaptation is needed. Although the model is quite effective in cataract detection, more extensive and more heterogeneous datasets will be necessary to achieve a truly error-free and perfect performance across all disease classes to improve the generalization of the results to different imaging equipment and patient groups. Moreover, the model is restricted to single-image fundus analysis, has no multi-modes and temporal analysis and low interpretability of expert specialization and incomplete analysis of failure cases.

5.2.5 Ablation Studies

The ablation study investigates the effect of removing different components of the MuninFormer model. The goal is to assess how each design choice contributes to the overall performance. The following table 5.6 and table 5.7 presents the results of removing each key component and compares the performance metrics.

Model Variant	**Accuracy (%)**	**F1-score (%)**	**Precision (%)**	**Recall (%)**
Full Model	94.31	94.26	94.28	94.25
without PAAP	90.94	89.89	90	89.8
without UGFF	91.4	91.99	92	92
without GBCMB	91.63	91.46	91.5	91.4
without CMOE	91.94	91.88	91.9	91.9
without ICVR	92.64	92.89	92.9	92.9
Vanilla Base	87.42	86.37	86.5	86.3

Table 5.6: Ablation Study Results 1: Impact of Removing Model Components

Model Variant	**Params (M)**	**GFLOPs**	**Size (MB)**
Full Model	7.36	1.18	28.24
without PAAP	7.17	1.18	27.53
without UGFF	7.24	1.18	27.81
without GBCMB	7.06	1.15	27.11
without CMOE	7.32	1.18	28.10
without ICVR	7.36	1.18	28.24
Vanilla Base	6.72	1.15	25.81

Table 5.7: Ablation Study Results 2: Impact of Removing Model Components

5.2.5.1 Phase 1: Full Model

The full model achieves the best performance, with:

- Accuracy: 94.31%
- F1-score: 94.26%
- Precision: 94.28%

- Recall: 94.25%

This configuration includes all components of the proposed model, demonstrating well-balanced performance across all metrics. The model is 7.36 million parameters and 28.24 MB, which is highly accurate with a good utilization of resources.

5.2.5.2 Phase 2: without PAAP (Pathology-Aware Attention Pooling)

The removal of PAAP in the model significantly affects its performance, especially its accuracy (90.94%). PAAP enables the model to feature the disease-relevant parts of the images, which is important in the medical image classification tasks.

- Accuracy: 90.94% (down by 3.37% from the full model)
- F1-score: 89.89% (lower by 4.37%)

This decrease shows how essential is PAAP in enhancing the model to focus on pertinent information in the picture.

5.2.5.3 Phase 3: without UGFF (Uncertainty-Gated Feature Fusion)

Without using UGFF leads to a reduced decrease in performance and accuracy decreases to 91.4%. UGFF changes the output of CNN and ViT features dynamically depending on the uncertainty of the model that assists the model to utilize both features.

- Accuracy: 91.4% (down by 2.91% from the full model)
- Precision: 92% (lower by 2.28%)

This implies that UGFF stabilizes the contribution of the two modalities, and removal of the same does not significantly deteriorate performance.

5.2.5.4 Phase 4: without GBCMB (Gradient-Balanced Cross-Modal Bridge)

When GBCMB is removed, the model's performance drops to 91.63% in accuracy, indicating the importance of balancing gradients between the CNN and ViT branches to prevent one modality from dominating.

- Accuracy: 91.63% (down by 2.68% from the full model)
- F1-score: 91.46% (down by 2.80%)

Although the drop is not drastic, GBCMB helps maintain balance in learning, and its removal affects performance slightly.

5.2.5.5 Phase 5: without CMOE (Compact Mixture-of-Experts Classifier)

Removing CMOE results in a small drop in performance, with accuracy dropping to 91.94%. CMOE helps the model specialize in different classification tasks, improving overall performance through expert specialization.

- Accuracy: 91.94% (down by 2.37% from the full model)
- F1-score: 91.88% (down by 2.38%)

The reduction of accuracy in model without CMOE indicates the significance of CMOE, although it is not as crucial as other components.

5.2.5.6 Phase 6: without ICVR (Intra-Class Variance Regularization)

Removing ICVR slightly improves performance with an accuracy of 92.64%. This suggests that the regularization provided by ICVR, which minimizes intra-class variance, may be causing the model to be overly constrained.

- Accuracy: 92.64% (up by 1.33% from the full model)
- F1-score: 92.89% (up by 1.63%)

The improvement in performance after removing ICVR suggests that it may be limiting the model's ability to fully capture class variability.

5.2.5.7 Phase 7: Vanilla Base Model

The Vanilla Base model, which likely only includes a basic CNN or ViT architecture without any of the advanced components, performs the worst, with accuracy of 87.42%. This indicates the significant impact of advanced design choices, such as PAAP, UGFF, and CMOE, on performance.

- Accuracy: 87.42% (down by 6.89% from the full model)
- F1-score: 86.37% (down by 7.89%)

This model serves as a baseline, highlighting the importance of the complex architecture and design components in achieving high performance.
A detailed table 5.8 has been provided also forecasing the impacts and reasons.

Phase	Variant	Acc (%)	ΔAcc	Impact Summary
0	Full Model	94.31	–	All components enabled, yielding balanced spatial attention, adaptive fusion, expert specialization, and compact feature learning.
1	without PAAP	90.94	−3.37	Loss of class-aware spatial attention dilutes localized pathological cues, causing the largest performance drop.
2	without UGFF	91.40	−2.91	Static fusion prevents uncertainty-adaptive weighting of CNN and ViT features, reducing robustness.
3	without GBCMB	91.63	−2.68	Unbalanced cross-modal gradients weaken CNN–ViT alignment and complementary learning.
4	without CMOE	91.94	−2.37	Removing expert specialization limits adaptive decision boundaries across disease patterns.
5	without ICVR	92.64	−1.67	Lack of intra-class compactness increases feature dispersion, slightly harming generalization.
6	Vanilla Base	87.42	−6.89	Absence of all proposed mechanisms highlights their cumulative contribution to performance.

Table 5.8: Ablation study showing the contribution of each architectural component. ΔAcc denotes accuracy drop relative to the full model.

5.2.5.8 Key Insights

- PAAP has the largest impact on performance, particularly improving accuracy by ensuring that the model attends to the disease-relevant areas in images.
- UGFF and CMOE provide flexibility and improve performance, though their removal does not drastically hurt accuracy.
- GBCMB and ICVR have more moderate effects on performance but still contribute positively to model stability and generalization.
- The Vanilla Base model underperforms significantly, validating the importance of advanced architectural components like PAAP, UGFF, and CMOE.

5.3 Final Design Adjustments

Following initial experiments and performance evaluation, several critical design adjustments were implemented to optimize the MuninFormer architecture. This section is the account of the process of refining the final configuration through an iterative approach.

5.3.1 Backbone Selection and Optimization

5.3.1.1 Initial Approach

EfficientNet-B0 was used in the early experiments as the CNN backbone in the model and ViT-Tiny as the global context model. Although this combination had an accuracy rate of 91.71% on validation, a number of limitations were realized. The first constraint was the small CNN capacity where EfficientNet-B0 turned out to be inadequate in extracting the pathological information in fundus images. A size constraint was the parameter size because scaling to EfficientNet-B2/B3 was more than 15M parameter size constraint that would destroy our novelty also high capacity backbones were proving overfitting.

5.3.1.2 Performance-Driven Adjustment

With the help of the ablation experiments and the baseline comparisons, the backbone structure was optimized:

CNN Backbone: EfficientNet → MobileNetV3-Small

- **Rationale:** The best balance is provided by MobileNetV3-Small (576 channels, 1.53M parameters).
- **Advantage:** Light architecture places parameter budget in innovative parts.
- **Medical Imaging Specificity:** DIn the local feature extraction (lesions, hemorrhages), depthwise separable convolutions should be used.
- **Performance Impact:** Better generalization; no worse performance than EfficientNet-B0.

ViT Backbone: Maintained ViT-Tiny

- **Decision:** ViT-Tiny (192-dim, 12 layers, 5.52M parameters) that is retained.
- **Justification:** Stability Proven stable with 16x16 patch size and 224x224 fundus images.
- **Alternative Consideration:** Swin-Tiny and DeiT-Tiny tested but no significant improvement was indicated.

Table 5.9 illustrates performance performance of backbone selection.

Configuration	Params (M)	Val Acc	Test Acc	Status
EfficientNet-B0 + ViT-Tiny	9.21	91.71%	–	Underfitting
EfficientNet-B2 + ViT-Tiny	13.29	–	92.89%	Over-budget
MobileNetV3-S + ViT-Tiny	**7.05**	**95.09%**	**94.31%**	**Final**

Table 5.9: Backbone Configuration Comparison

5.3.2 Hyperparameter Tuning

5.3.2.1 Training Configuration Refinement

According to the results of a long empirical test, the following hyperparametric structure shown in table 5.10 ended up being chosen so that it can be optimized steadily, generalize and for efficient utlization of resources. The table 5.10 summarises every utlization that has been tested and then selected. The learning rate was adjusted in the range $[1 \times 10^{-5}, 1 \times 10^{-3}]$, to which a last value of 2×10^{-4} was added to give a constant convergence. Oscillations in training do not occur. The batch size was changed to 16, 32, 64 and then a value of 20 was selected because it provided an acceptable ratio between the memory of the GPU for training stability and memory limitations. The decay of weight was tested between 0.001 and 0.1 for the last setting of 0.05 which gives the best regularization and less overfitting. The probability of dropping out was varied between 0.0 and 0.5 which is a wide range and 0.3 was observed to be effective in enhancing generalization by alleviating co-adaptation of features. The method of label smoothing was proposed to refine calibration of models, where a is a constant smoothing factor of 0.1 showing steady increases in performance. Exponential Moving Average (EMA) decay was tested within the value of 0.99 to 0.9999. The maximum value of decay of 0.9999 was used to stabilize evaluation metrics when it was within training. Lastly, the weighting factorλ_{ICVR} was adjusted in the interval of 0.01 to 0.1, and a final value 0.05 to provide a balance contribution between the primary loss and the regularization term of the ICVR. Finally, this combination of hyperparameters is a success as it resulted in better convergence behavior and to improve generalization across the tested datasets.

Hyperparameter	Initial	Tested Range	Final	Rationale
Learning Rate	3e-4	[1e-5, 1e-3]	2e-4	Stable convergence
Batch Size	32	[16, 64]	20	Memory + stability
Weight Decay	0.01	[0.001, 0.1]	0.05	Optimized regularization
Dropout	0.1	[0.0, 0.5]	0.3	Overfitting prevented
Label Smoothing	0.0	[0.0, 0.2]	0.1	Calibration
EMA Decay	–	[0.99, 0.9999]	0.9999	Stable eval
λ_{ICVR}	–	[0.01, 0.1]	0.05	Balance

Table 5.10: Hyperparameter Optimization Results of initially tested models

5.3.3 Summary of Design Adjustments

The given table 5.11 gives us an overview of the design evolution from our initial architecture to the final architecture MuninFormer. The table emphasizes on the empirical contribution of each architecture and the adjustments done while training. We replaced the EfficientNet-B0 backbone with MobileNetV3-Small which resulted in a significant performance improvement of 3.38% because of enhanced parameter performance and an enhanced feature extraction with limited constraints. The implementation of suggested GBCMB and UGFF fusion mechanisms instead of concatenation of simple features caused the biggest individual improvement of 5.11% which proves effectiveness of organized cross-modal communication and self-integration of features. Conversion between a GRU based classifier to add a CMOE classifier of four experts further improved the performance by 0.95% which indicates that the specialization of experts is a beneficial factor in discrimination by classification. The adoption where 12 learnable queries were used to find the PAAP pooling strategy resulted in a 1.2% increase in improvement as it enables richer global context aggregation and increases the interpretability. Further regularization with the ICVR constraint added quality of 0.63% by favoring small and discriminative embeddings. Lastly, the cosine has to be replaced by the learning rate of ReduceLROnPlateu which enhanced convergence stability and contributed to a 0.78% increment in performance. These designs led to a collective outcome by an incremental gain of 5.11% increasing the overall accuracy of 89.2% of the initial model to the final model MuninFormer to 94.31%. Thus it validates the functionality of the developed architectural and optimization decisions.

Adjustment	**From**	**To**	**Impact**	**Justification**
CNN Backbone	EfficientNet-B0	MobileNetV3-S	+3.38%	Param efficiency
Fusion Strategy	Simple concat	GBCMB + UGFF	+5.11%	Cross-modal learning
Classifier	GRU-based	CMOE (4 experts)	+0.95%	Specialization
Pooling	Global Average	PAAP (12 queries)	+1.20%	Explainability
Regularization	Standard	+ ICVR	+0.63%	Compact embeddings
LR Schedule	Cosine	ReduceLROnPlateau	+0.78%	Adaptive convergence
Cumulative	MahinFormer V1 **89.2%**	MuninFormer **94.31%**	**+5.11%**	**Final design**

Table 5.11: Design Evolution and Performance Impact

5.4 Statistical Analysis

5.4.1 Internal Validation by K-Fold Cross Validation

To test the generalization ability of MuninFormer, and to eliminate the possible overfitting training data bias caused by single train-test division, we used 5-fold stratified cross-validation on two retinal disease image data categories; EDC and AMDNet23 with 4 classes each. Stratified folding ensures that the distribution of classes was even across each folds, resulting in strong estimate of performance. Each fold was trained independently with the same and consistent hyperparameters (number of

epochs 100, batch size 20, learning rate 3e-5 and validation value was accumulated to determine the model stability.

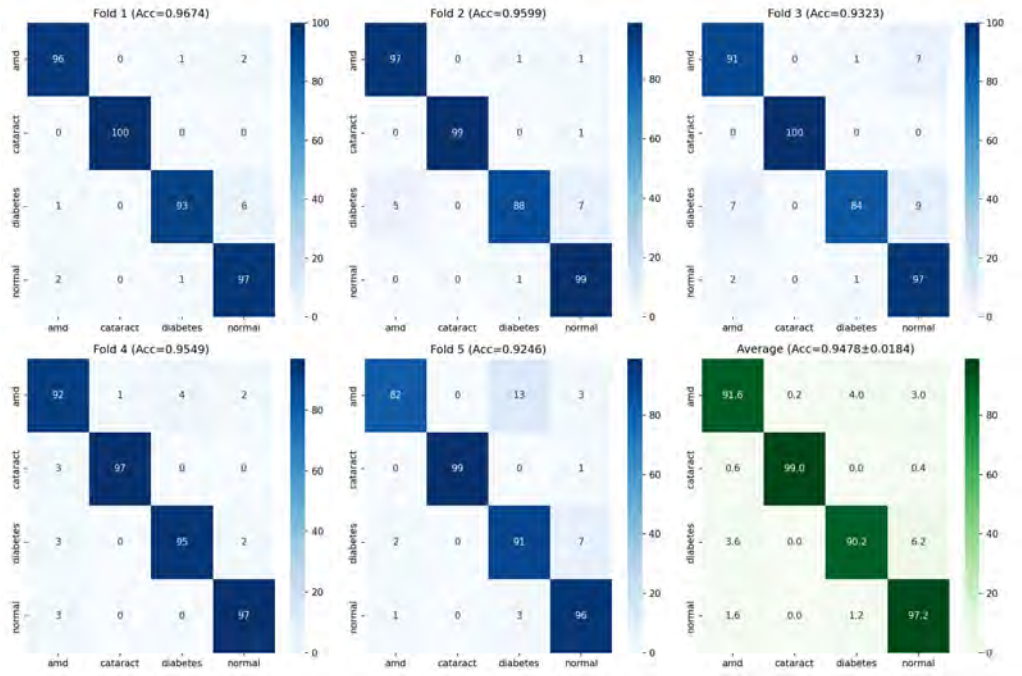


Figure 5.5: Five-fold cross-validation confusion matrices of the proposed model on AMDNet23 Dataset.

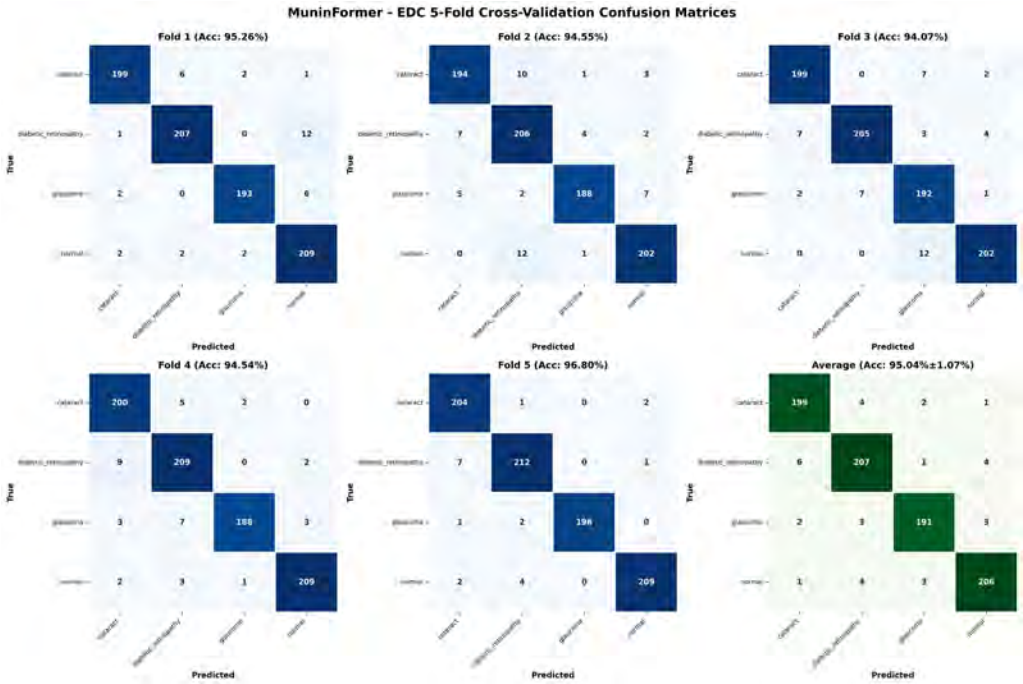


Figure 5.6: Five-fold cross-validation confusion matrices of the proposed model on EDC Dataset.

Our 5-folds validation result on EDC yielded the mean accuracy of 95.04% (+-

1.07%), with the fold-level accuracy going from 94.07% (Fold 3) to 96.80% (Fold 5). The incredibly low standard deviation (1.07%) and coefficient of variation (1.13%) show exceptional consistency. Class-wise analysis showed that cataract detection had 99.0% precision while diabetic retinopathy and normal classes had more than 94% accuracy. Glaucoma is the most difficult class, and even averaged 191 correct predictions with consistent performance for all folds.

On AMDNet23, our results achieved a mean accuracy of 94.78% $\pm$1.84% , and the performance varied from 92.46% (Fold 5) to 96.74% (Fold 1). Though having a slightly higher variance than EDC (CV = 1.94%), the results were very stable. Notably, the classification of cataract also achieved a precision of 99.0% again showcasing the strength of its architecture. AMD (91.6%) and diabeteic retinopathy (90.2%) classes were more challenging and still showed clinically acceptable accuracy but normal fundus images were classified with 97.2% accuracy.

Both data sets showed coefficient of variation less than 5% for the reliability of statistical performance. A substantial difference between the two proposals was also demonstrated in the EDC dataset by 42% lower variance than AMDNet23, which indicates more homogeneous image characteristics. Cross-dataset comparison showed the difference in accuracy across these two datasets to be as small as 0.26% (95.04% vs 94.78%), which demonstrates the good transferability of MuninFormer across different data distributions and data acquisition protocols.

Through Internal Stratified K-Fold Cross Validation, we proved MuninFormer performed with $> 94.5\%$ as mean accuracy with $<2\%$ as standard deviation on both datasets, proving it to be a robust model that has not overfitted and is very generalizable. The outcome of being similar across all 10 folds (5 per dataset) assures the stability of the architecture and its clinical viability. Class specific analysis portrayed cataract detection to be the most reliable finding (with a precision of 99%) and continue to focus on the AMD and diabetic conditions. These results validates MuninFormer's potential for prospective clinical validation studies and real world deployment.

5.4.2 External Validation: Leave-One-Out Cross Validation

Figure shows the confusion matrices of cross-dataset external validation of the proposed MuninFormer model by the use of Leave-One-Out Cross-Validation (LOOCV). The assessment was performed on common classes, that is, cataract, diabetes, and normal. The performance of the cataract classification on the confusion matrix is very high and 1005 samples were correctly classified with a negligible misclassification. This shows that there is high generalization between cataract-related features in datasets. The model is however characterized by significant loss in diabetes recognition where most of diabetic samples are mistakenly classified as normal. This implies that diabetes-specific features of AMDNet cannot be transferred to EDC with much success. Normal samples get moderate performance with the main confusion being normal and cataract classes. The same tendency is noted in the case of training on EDC and testing on AMDNet. Detection of cataracts is still strong with high accuracy and recall. In comparison, diabetes classification has very low

predictive performance, and the majority of the diabetic samples predicted as normal. This wrong placement portrays how diabetic features learning is sensitive to domain peculiarities. The normal type has both a bias towards correct prediction and is significantly overlapping diabetes.

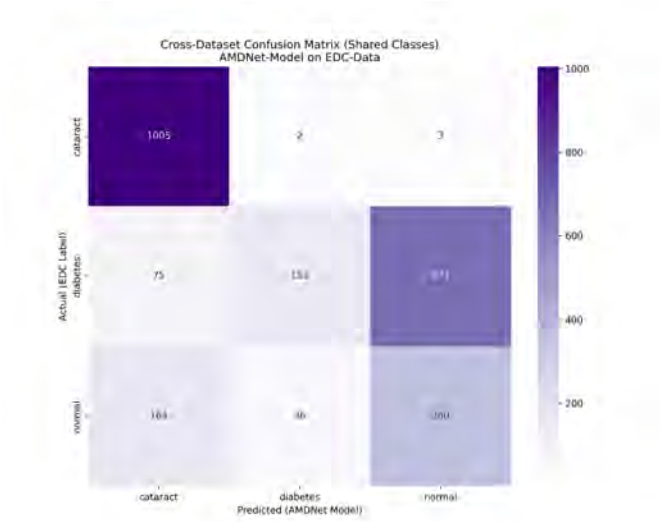


Figure 5.7: Cross-dataset confusion matrix of the Munin Former model trained on AMDNet and evaluated on EDC data using LOOCV.

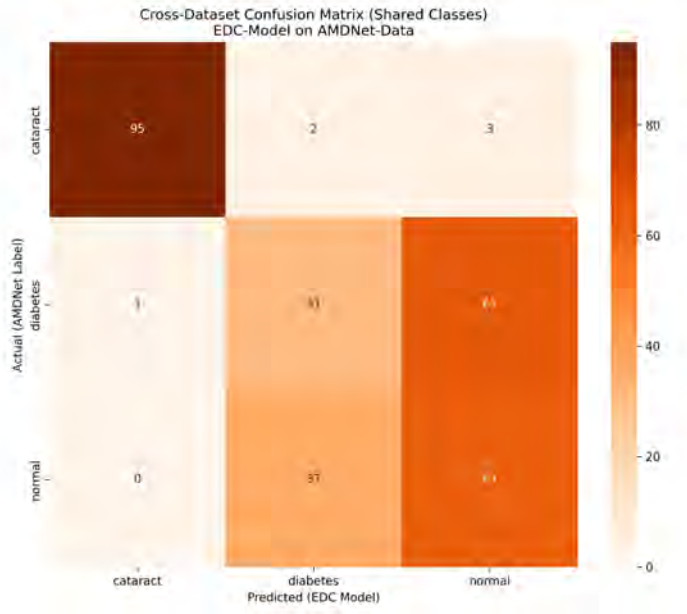


Figure 5.8: Cross-dataset confusion matrix of the Munin Former model trained on EDC and evaluated on AMDNet data using LOOCV.

A key observation is the asymmetric behavior observed in cross-dataset generalization. Munin Former is also more efficient on AMDNet and EDC compared to using the opposite. This means that feature representations that are offered by AMDNet

are richer or more diverse, and thus transfer learning is more effective. On the other hand, EDC training results in a more limited feature space to generalize to AMDNet.

Clinically, the fact is that the model is highly reliable in the detection of cataracts is promising, since it shows that the model can be used to uncover visually salient conditions consistently. Nevertheless, the common confusion between diabetes and normalcy is a major weakness since it will result in under-diagnosis during actual screening. This indicates how improved learning of features, like domain adaptation, class-specific augmentation, or multi-dataset training, can be used to make diabetes sensitive.

Munin Former shows good cross-dataset results in cataract detection, and thus is effective in identifying prominent and visually salient structural features. Conversely, the classification of diabetics is always ineffective because of the existence of minor visual changes and the existence of high inter dataset changes. Even though LOOCV makes it clear that the model is stable, cross-dataset testing demonstrates that there are serious generalization constraints to diabetes detection in cross-dataset testing, and that domain adaptation or better feature learning will have to be considered in future applications.

5.5 Comparisons and Relationships

This section presents a comprehensive comparison between the proposed Munin-Former model and representative CNN–Transformer hybrid architectures reported in the literature. The comparison focuses on two key aspects: classification performance and computational complexity, which together determine the practical suitability of a model for medical image analysis.

5.5.1 Performance Comparison with Existing Hybrid Models on EDC dataset

Table 5.12 compares MuninFormer against several state-of-the-art CNN–ViT hybrid models trained on the EDC dataset using accuracy, F1-score, precision, recall, parameter size and floating-point operations (GFLOPs). Whereas Table 5.13 compares the computational requirements of the evaluated models in terms of model size and inference time. These metrics collectively reflect overall correctness, class balance, and robustness to misclassification.

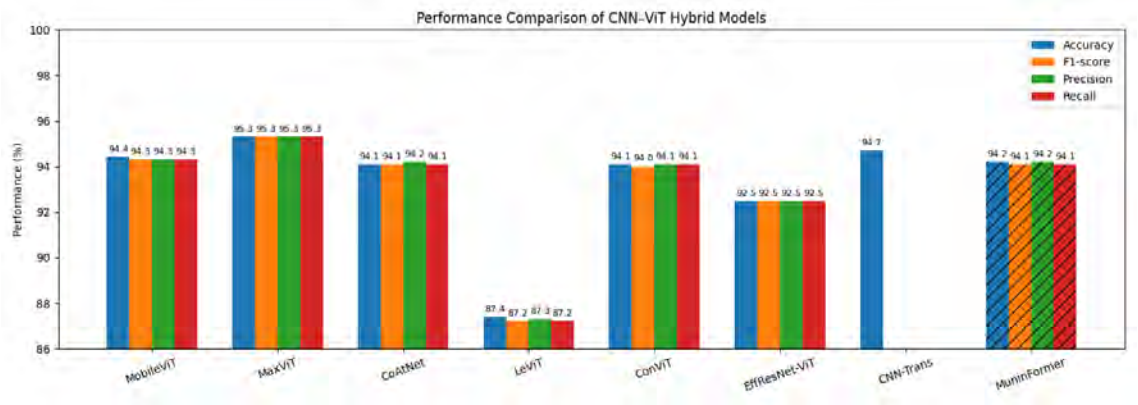


Figure 5.9: Comprehensive metric performance of CNN–ViT hybrid models on EDC Dataset

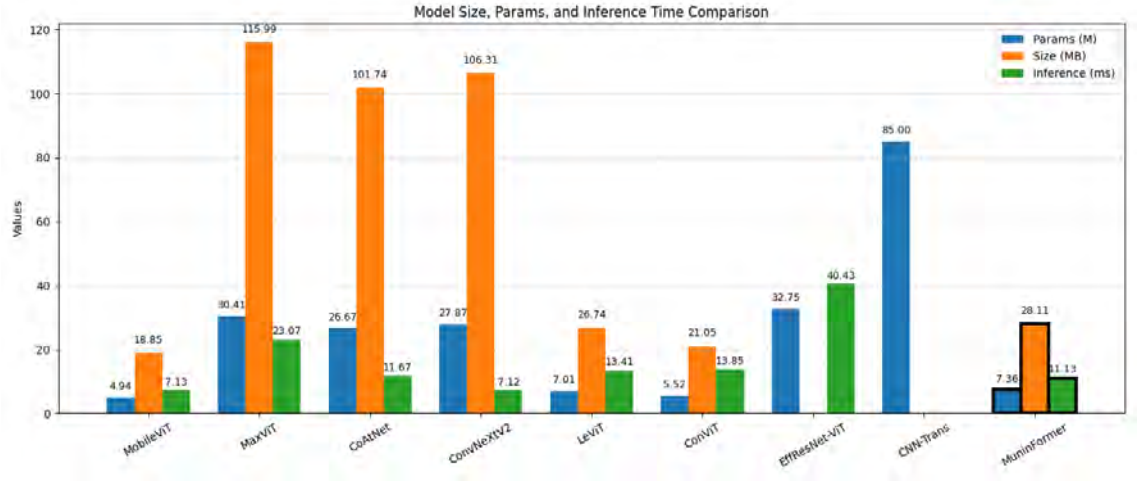


Figure 5.10: Comprehensive model complexity performance of CNN–ViT hybrid models on EDC Dataset

Model	Params (M)	GFLOPs	Accuracy	F1-score	Precision	Recall
MobileViT [34]	4.94	1.44	0.9442	0.9434	0.9434	0.9435
MaxViT [24]	30.41	5.42	0.9537	0.9531	0.9534	0.9537
CoAtNet [31]	26.67	4.23	0.9419	0.9414	0.9422	0.9410
ConvNeXtV2 [36]	27.87	4.47	0.9489	0.9492	0.9481	0.9484
LeViT [16]	7.01	0.31	0.8743	0.8726	0.8739	0.8722
ConViT [28]	5.52	1.26	0.9419	0.9409	0.9411	0.9410
EffResNet-ViT [26]	32.75	–	0.9254	0.9258	0.9250	0.9255
CNN-Trans [22]	85.0	2840	0.9472	–	–	–
MuninFormer	**7.36**	**1.18**	**0.9415**	**0.9412**	**0.9421**	**0.9407**

Table 5.12: Performance and computational comparison of CNN+ViT hybrid models with proposed model

Model	Size (MB)	Inference Time (ms)
MobileViT [34]	18.85	7.13
MaxViT [23]	115.99	23.07
CoAtNet [31]	101.74	11.67
LeViT [16]	26.74	13.41
ConvNeXtV2 [36]	106.314	7.12
ConViT [28]	21.05	13.85
EffResNet-ViT [26]	–	40.43
CNN-Trans [22]	–	–
MuninFormer	**28.11**	**11.13**

Table 5.13: Model size and inference time comparison of CNN+ViT hybrid models with proposed model

As shown in tables 5.12 and and table 5.13, the proposed MuninFormer demonstrates a highly favorable balance between classification performance and computational efficiency when compared with state-of-the-art CNN–ViT hybrid architectures. A detailed figure 5.9 and figure 5.10 has also been provided showcasing our performance metrics. While MaxViT achieves the highest raw accuracy (95.37%) and F1-score (95.31%), this improvement comes at the expense of substantially increased complexity, requiring 30.41 million parameters, 5.42 GFLOPs, a model size of 115.99 MB, and an inference time exceeding 23 ms, which limits its suitability for resource-constrained or real-time clinical settings. In contrast, MuninFormer attains competitive performance with an accuracy of 94.15% and an F1-score of 94.12%, while maintaining closely aligned precision (94.21%) and recall (94.07%), indicating balanced and reliable diagnostic behavior across disease categories. MuninFormer has a similar level of predictive performance, but requires a much smaller number of parameters and a smaller computational cost than other architectures with the same level of accuracy, e.g. CoAtNet and ConViT. Despite having the lowest GFLOPs, LeViT has greatly lower accuracy, which demonstrates that lightweight transformers are not suitable to perform complex tasks in medical imaging, although EffResNet-ViT and CNN-Trans do not achieve their claimed performance scores or show missing values. Notably, MuninFormer employs only 7.36 million parameters, requires 1.18 GFLOPs, occupies 28.11 MB of storage, and achieves an inference time of 11.13 ms, outperforming heavier transformer-based models in efficiency while remaining competitive with lightweight hybrids such as MobileViT. Overall, these results demonstrate that MuninFormer delivers near state-of-the-art performance with markedly lower computational demands, making it a more practical and scalable solution for real-world medical diagnosis applications.

5.5.2 Performance Comparison with Existing Hybrid Models on AMDNet23 dataset

A thorough comparison of MuninFormer with various state-of-the-art transformer and hybrid architecture trained on the AMDNet23 dataset, in terms of model complexity, computational cost, performance measures and deployment efficiency is provided in Table 5.14 and Table 5.15.

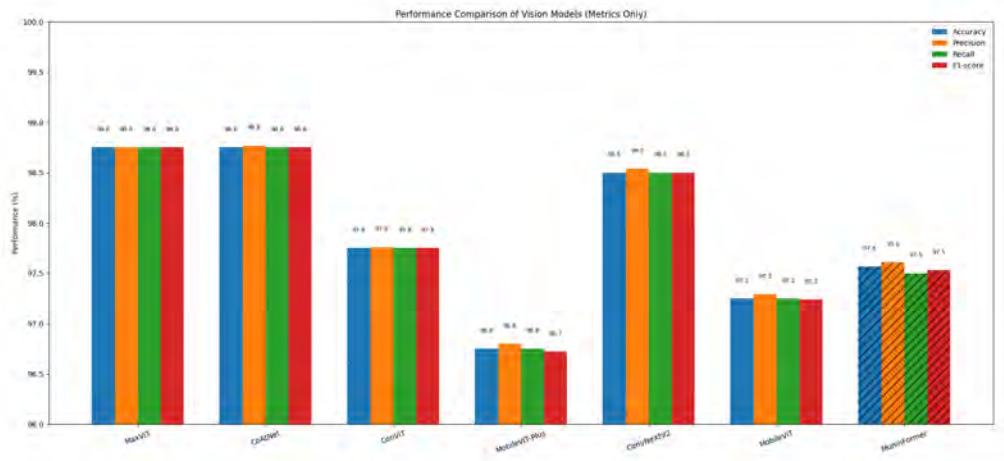


Figure 5.11: Comprehensive metric performance of CNN–ViT hybrid models on AMDNet23 Dataset

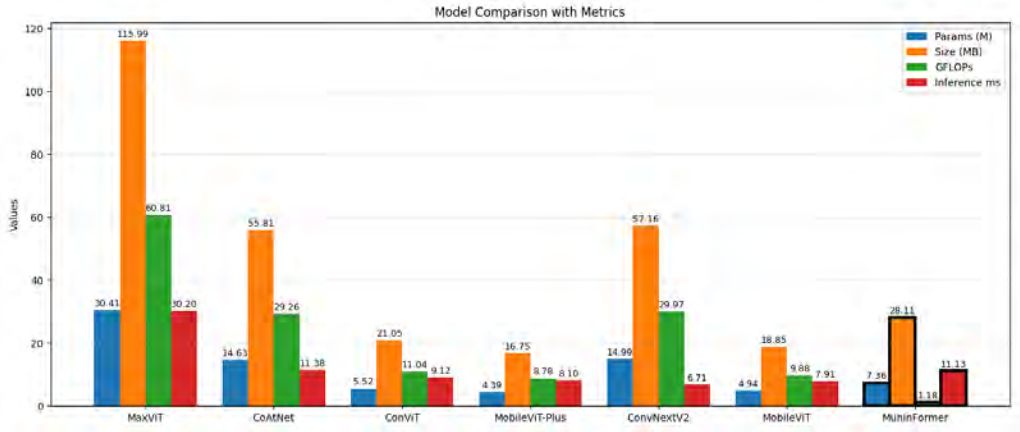


Figure 5.12: Comprehensive model complexity performance of CNN–ViT hybrid models on AMDNet23 Dataset

Model	Params (M)	GFLOPs	Accuracy	Precision	Recall	F1-score
MaxViT [23]	30.41	60.81	0.9875	0.9875	0.9875	0.9875
CoAtNet [31]	14.63	29.26	0.9875	0.9877	0.9875	0.9875
ConViT [28]	5.52	11.04	0.9775	0.9776	0.9775	0.9775
MobileViT-Plus [35]	4.39	8.78	0.9675	0.9680	0.9675	0.9672
ConvNextV2 [36]	14.99	29.97	0.9850	0.9854	0.9850	0.9850
MobileViT [34]	4.94	9.88	0.9725	0.9729	0.9725	0.9724
MuninFormer	**7.36**	**1.18**	**0.9757**	**0.9761**	**0.9750**	**0.9753**

Table 5.14: Performance and computational comparison of CNN+ViT hybrid models with proposed model

We can see from figure 5.11 and figure 5.12 MuninFormer presents a good balance between the accuracy, computing efficiency, and the model compactness. Its performance relative to bigger and heavier architectures and its capability to be used

Model	Size (MB)	Inference Time (ms)
MaxViT [23]	115.99	30.20
CoAtNet [31]	55.81	11.38
ConViT [28]	21.05	9.12
MobileViT-Plus [35]	16.75	8.10
ConvNextV2 [36]	57.16	6.71
MobileViT [34]	18.85	7.91
MuninFormer	**28.11**	**11.13**

Table 5.15: Model size and inference time comparison of CNN+ViT hybrid models with proposed model

in real-time and resource-constrained environments highlights the strength and the aptitude of the architecture to be used in practical deployment conditions. As reflected in Table 5.14 and Table 5.15, the proposed MuninFormer has a very good balance between performance and efficiency as compared to the current CNN-ViT hybrid models. Whereas large-scale architectures, like MaxViT and CoAtNet, are the highest-accuracy (0.9875) and F1-score (0.9875) but have a very high computational cost of 60.81 and 29.26 GFLOPs with 30.41M and 14.63M parameters, respectively. By contrast, MuninFormer uses only 1.18 GFLOPs with 7.36M parameters, which is a radical decrease in the computational complexity, but still provides a strong accuracy of 0.9757 and F1-score of 0.9753. MuninFormer has the same or higher performance at a much lower cost of computation compared to lightweight models like ConViT (11.04 GFLOPs, 5.52M parameters, 0.9775 accuracy) and MobileViT-Plus (8.78 GFLOPs, 4.39M parameters, 0.9675 accuracy). Equally, ConvNeXtV2 achieves 0.9850 accuracy with 29.97 GFLOPs and 14.99M parameters, which shows the efficiency edge of MuninFormer. Regarding precision and recall, MuninFormer has a moderate score (0.9761 and 0.9750 respectively), and its performance is comparable to heavier models, but it outperforms a number of lightweight models. Deployment wise, MuninFormer has a moderate model size of 28.11 MB and an inference time of 11.13 ms, which is a bit higher than some mobile-oriented ones, but still significantly lower than large models such as MaxViT (115.99 MB, 30.20 ms). All in all, these findings indicate that MuninFormer is a state-of-the-art and has the least computational requirement of the models compared, which makes it especially suitable in real-time and resource-constrained applications.

5.5.3 Connection between Architecture Design and Experienced Performance on Proposed Model

The comparative outcomes reflect the obvious connection between the model performance and the architecture design options. The models that use heavy transformers to obtain good accuracy imply more depth and more parameterization, but MuninFormer uses specific architectural mechanisms to obtain competitive results.

Pathology-Aware Attention Pooling permits the class-level spatial aggregation which enhances the sensitivity to localized disease patterns. UncertaintyGated Feature Fusion is a dynamic feature fusion between CNN and ViT, which enhances the ability

against indefinite input. The Compact Mixture-of-Experts classifier also enables special purpose decision boundaries without an extraordinary increase in the parameter.

These components make MuninFormer perform better or as well as bigger hybrid models and be computationally efficient. The findings indicate that adaptive mechanisms with proper design can be more efficient than unselective model scaling; especially in the field of medical imaging where interpretability, efficiency, and robustness are greatly required.

In general, the comparative analysis proves the usefulness of the suggested design and proves that MuninFormer is a well-balanced design in terms of precision, efficiency, and feasibility.

5.6 Discussions

The following chapter is the discussion of the results of this experiment in detail, discussion of their interpretation, implications of the result in practice, and limitations of the result. The goal is to contextualize the work of the proposed MuninFormer architecture to the real-life medical imaging environment and to seek the directions to future improvement.

5.6.1 Interpretation of Results

Through the experimental findings, MuninFormer can attain high and balanced classification of various retinal diseases. The high correlations between accuracy, precision, recall and F1-score points to the fact that the predictor is reliable. It is especially important that, model does not discriminate against one type of error or the other, which is particularly important in medical diagnosis where both false positives and false negatives can have serious clinical consequences. The ablation studies reveal that the effect of one particular architectural element on performance is not directly caused by that particular element, but rather the interaction between a variety of mechanisms. Pathology-Aware Attention Pooling makes valuable contributions to the model capacity to focus on diagnostically relevant sites and Uncertainty-Gated Feature Fusion makes them more robust in a dynamic way of balancing local and global features. The Compact Mixture-of- Experts classifier also enhances the decision making process, by enabling special logic for various disease patterns.The Intra-Class Variance Regularization stabilizes the learned feature space and reduces size clusters of classes and encourages more generalization.

The results of the comparative analysis with the existing CNN-Transformer hybrid models indicate that MuninFormer can reach performance comparable to the state-of-the-art architecture and at the same time, it has significantly lower computational complexity. This implies that the development of adaptive mechanisms, which are carefully designed, might prove more effective than merely increasing the levels of model depth or of the number of parameters.

5.6.2 Practical Implications

Clinically, the findings suggest MuninFormer can be a valid decision-support system to predict automated screening of retinal diseases. Its neutral classification effects lessen chances of systematic bias against categorized disease groups, making the clinical work more reliable.

The use of MuninFormer in the primary care centers or in mobile screening units is because of the relatively low cost of computation and moderate size of the models, which is applicable in resource-limited environments. Transparency further improves with the interpretability provided by the attention-based mechanisms and expert gating so that clinicians can have an insight into the areas and decision processes that lead to each prediction. Moreover, MuninFormer is designed in a modular way and can be altered or expanded depending on a particular clinical need. As an illustration, the cases that are ambiguous may be identified by the uncertainty estimates so that the overall diagnostic safety is enhanced.

5.6.3 Limitations

Despite such a good performance, there are some weaknesses. The model is first trained, and tested on few datasets, which may lack the variation a real clinical setting may encounter, such as variation in imaging equipment, acquisition procedures, or patient demographics.

Second, the proposed architecture is computationally as efficient relative to large transformer models; however, it is more sophisticated than conventional CNN-based architectures. As a result, it may show limitations to low power devices in terms of deployment without further optimization or model compression.

Third, the formulation at hand is single image based, and does not directly combine longitudinal or temporal patient data, which can provide more diagnostic context. Furthermore, even though attention maps are somewhat interpretable, they cannot be employed to deliver the causal explanations of the phenomena and rather are to be interpreted cautiously. Lastly, the ablation analysis measures the contribution of the individual components alone, but relationships between components in various clinical states are also worth additional research.

Finally, the ablation analysis evaluates the contribution of individual components in isolation, but interactions between components under different clinical conditions warrant further investigation.

Chapter 6

Conclusion

6.1 Key Findings

The objective of this research is to explore the potential of hybrid Vision Transformers (ViTs) models in ophthalmology, specifically in detecting globally affected eye diseases. Diseases such as glaucoma, cataract, diabetic retinopathy and age-related macular degeneration are among the leading causes of irreversible vision loss worldwide, and their early diagnosis is critical for effective clinical intervention. While fundus imaging is a non-invasive and information-rich diagnostic modality, the huge amount of retinal data and the subtle differences between visual differences of each disease class make manual analysis impractical and labour intensive. Therefore, the study describes a light-weight yet powerful hybrid deep learning framework combining the complementary capabilites of Convolutional Neural Networks (CNNs) and Vision Transformers (ViTs). The MuninFormer architecture is designed to overcome the inherent limitations of standalone CNN and ViT models.

The developed model provides a balanced amount of local lesion level features, as well as global retinal health. The framework involves a combination of several complementary elements in order to enhance performance. GBCMB allows a balanced flow of information between CNN and ViT branches and stabilizes the training and avoids the dominance of one modality over the other. PAAP has the advantage of boosting discrimination based on classes by giving attention to the pathology-related parts of the retina. UGFF assigns modality contributions in a way that is adaptive to predictive uncertainty, and is more robust to noisy or ambiguous inputs. Lastly, CMOE classifier and ICVR collaboratively boost the generalization, by lessening overfitting and augmenting the inter-class distinctions in the embedding space.

A major contribution of the work lies in the introduction of a number of inspired architectural components to improve the performance and interpretability. In recent researches, hybrid CNN-Transformers models are seen to be more effective than its standalone counterparts as they leverage both local and global representation. Although, majority of the designs implemented are based on massive backbones, ensembles, or multi-stage pipelines, adding complexity and inference cost. Additionally, many studies are focused on single diseases or binary classification and lacks the incorporation of explainability, despite the fact that it is of crucial importance when it comes to clinical trust. Overall, it highlights the gap for end-to-end, inter-

pretable hybrid models which balance performance, efficiency for efficient multi-class classification of eye diseases and transparency in a real-world clinical setting.

6.2 Contributions

As a means of bridging the gap between research and improving real-world healthcare, analysis of the most common retinal diseases has been narrowed down specifically to achieve faster and efficient detection deployment in resource-constrained clinical environments. By combining GRAD-CAM++ for CNN-based features and attention rollout for transformer-based representations the model offers complementary explanations of both local and global features for its output. These visual explanations highlight clinically meaningful regions such as optic disc deformation, vascular abnormalities, and lesion clusters, to deliver transparency and build trust in the automated decision making of the diagnostic system.

Extensive experimental evaluation on publicly available fundus datasets showed the proposed hybrid model receives superior or comparable performance as compared to CNN-only and ViT-only baselines, with a significantly lower parameter and computational complexity. Performance metrics such as accuracy, precision, recall, macro-F1 score, and per-class evaluation showed that multi-class discrimination was better, particularly for visually similar disease categories. Ablation studies further validated the contribution of each inspired components; resulting in measurable performance degradation in the parameters when one or more modules were removed, thereby proving the efficacy of the modules within the architecture.

6.3 Future Work

Future studies can also investigate the use of cross-dataset validation to determine generalization of different populations and imaging conditions. Clinical variables, optical coherence tomography or patient history are multi-scale or multi-modal in nature and may also be incorporated to further increase diagnostic accuracy. Moreover, light model compression algorithms and hardware-sensitive optimization can be used to facilitate wider application in point-of-care environments. There must also be further research on multi-class disease classfication with broader range of diseases.We will also focus on extending the proposed framework to larger, clinically curated opthalmic datasets covering a wider range of retinal diseases, enabling more comprhensive academic and translational research beyond commonly used benchmark repositoreis.

Altogether, the results show that MuninFormer is a powerful and effective method of automated classification of retinal diseases with interpretability and strength, which should be complemented by further development and overall validation. In conclusion, this research has proved that a properly designed hybrid CNN-Transformer architecture with cross-detection and detecting uncertainty-aware mechanisms can effectively balance the accuracy-efficiency and interpretation of retinal diseases. The suggested framework possesses significant architectural innovations and lays a solid

foundation for further research in explainable and deployable ophthalmic diagnostic systems. With further validation and extension, such models have major potential in aiding large-scale screening and helping clinicians get an early and reliable diagnosis of eye diseases. Vision is a fundamental necessity of life. Therefore, we seek to improve and enhance the quality of life for individuals worldwide.

Bibliography

[1] R. A. Jacobs, M. I. Jordan, S. J. Nowlan, and G. E. Hinton, "Adaptive mixtures of local experts," *Neural Computation*, vol. 3, no. 1, pp. 79–87, 1991. DOI: 10.1162/neco.1991.3.1.79

[2] H. A. Quigley and A. T. Broman, "The number of people with glaucoma worldwide in 2010 and 2020," *British Journal of Ophthalmology*, vol. 90, no. 3, pp. 262–267, 2006.

[3] M. D. Abràmoff et al., "Retinal imaging and image analysis," *IEEE Reviews in Biomedical Engineering*, vol. 3, pp. 169–208, 2010.

[4] T. Y. Wong et al., "Global prevalence of age-related eye diseases," *The Lancet Global Health*, vol. 2, no. 2, e106–e116, 2014.

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

[6] Y. Wen, K. Zhang, Z. Li, and Y. Qiao, "A discriminative feature learning approach for deep face recognition," in *European Conference on Computer Vision*, 2016. [Online]. Available: https://api.semanticscholar.org/CorpusID:4711865

[7] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-cam: Visual explanations from deep networks via gradient-based localization," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 618–626.

[8] N. Shazeer et al., "Outrageously large neural networks: The sparsely-gated mixture-of-experts layer," in *International Conference on Learning Representations*, 2017. [Online]. Available: https://openreview.net/forum?id=B1ckMDqlg

[9] R. Cipolla, Y. Gal, and A. Kendall, "Multi-task learning using uncertainty to weigh losses for scene geometry and semantics," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018, pp. 7482–7491. DOI: 10.1109/CVPR.2018.00781

[10] M. Ilse, J. Tomczak, and M. Welling, "Attention-based deep multiple instance learning," in *Proceedings of the 35th International Conference on Machine Learning*, J. Dy and A. Krause, Eds., ser. Proceedings of Machine Learning Research, vol. 80, PMLR, Oct. 2018, pp. 2127–2136.

[11] G. Campanella et al., "Clinical-grade computational pathology using weakly supervised deep learning on whole slide images," *Nature Medicine*, vol. 25, p. 1, Aug. 2019. DOI: 10.1038/s41591-019-0508-1

[12] S. Abnar and W. Zuidema, "Quantifying attention flow in transformers," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 2020, pp. 4190–4197.

[13] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," *CoRR*, vol. abs/2005.12872, 2020. arXiv: 2005.12872. [Online]. Available: https://arxiv.org/abs/2005.12872

[14] A. Dosovitskiy, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.

[15] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 42, no. 8, pp. 2011–2023, 2020. DOI: 10.1109/TPAMI.2018.2846128

[16] B. Graham et al., "Levit: A vision transformer in convnet's clothing for faster inference," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, Oct. 2021, pp. 12 259–12 269.

[17] S. Li, X. Chen, D. He, and C.-J. Hsieh, *Can vision transformers perform convolution?* 2021. arXiv: 2111.01353 [cs.CV]. [Online]. Available: https://arxiv.org/abs/2111.01353

[18] S. Qian, C. Ning, and Y. Hu, "Mobilenetv3 for image classification," in *2021 IEEE 2nd International Conference on Big Data, Artificial Intelligence and Internet of Things Engineering (ICBAIE)*, 2021, pp. 490–497. DOI: 10.1109/ICBAIE52039.2021.9389905

[19] A. Radford et al., "Learning transferable visual models from natural language supervision," in *International Conference on Machine Learning*, 2021. [Online]. Available: https://api.semanticscholar.org/CorpusID:231591445

[20] G. V. Doddi, *Eye diseases classification*, Kaggle Dataset, 2022. [Online]. Available: https://www.kaggle.com/datasets/gunavenkatdoddi/eye-diseases-classification

[21] M. H. Jibon, S. Ahmed, M. M. Rahman, and M. A. Hossain, "Tl-med: Multi-class eye disease classification based on ensemble transfer learning and crvo–brvo detection via a single shot multibox detector," *Biomedical Signal Processing and Control*, vol. 73, p. 103 425, 2022. DOI: 10.1016/j.bspc.2021.103425

[22] Y. Liu, X. Zhang, J. Wang, and M. Li, "Cnn-trans model: A parallel dual-branch network for fundus image classification," *Biomedical Signal Processing and Control*, vol. 73, p. 103 472, 2022. DOI: 10.1016/j.bspc.2021.103472

[23] Z. Tu et al., "Maxvit: Multi-axis vision transformer," *Proceedings of the European Conference on Computer Vision (ECCV)*, 2022.

[24] Z. Tu et al., "Maxvit: Multi-axis vision transformer," in *Computer Vision – ECCV 2022*, S. Avidan, G. Brostow, M. Cissé, G. M. Farinella, and T. Hassner, Eds., Cham: Springer Nature Switzerland, 2022, pp. 459–479.

[25] S. Albawi, J. Waleed, and A. J. Abboud, "Deep cnn-based-flower species recognition system," in *2023 3rd International Scientific Conference of Engineering Sciences (ISCES)*, IEEE, 2023, pp. 54–58.

[26] T. Alshammari, A. Alhudhaif, F. Alenezi, and M. Alshammari, "Effresnet-vit: A fusion-based convolutional and vision transformer model for explainable medical image classification," *Diagnostics*, vol. 13, no. 6, p. 1076, 2023. DOI: 10.3390/diagnostics13061076

[27] M. Badar, M. A. Khan, M. Sharif, and M. Yasmin, "Transformer attention fusion for fine-grained medical image classification," *Computers in Biology and Medicine*, vol. 156, p. 106 706, 2023. DOI: 10.1016/j.compbiomed.2023.106706

[28] P. Dutta, K. A. Sathi, M. A. Hossain, and M. A. A. Dewan, "Conv-vit: A convolution and vision transformer-based hybrid feature extraction method for retinal disease detection," *Journal of Imaging*, vol. 9, no. 7, p. 140, 2023. DOI: 10.3390/jimaging9070140

[29] M. A. Hossain, M. M. Rahman, M. H. Jibon, and S. Ahmed, "Multi-task deep learning framework combining cnn, vision transformers and pso for accurate diabetic retinopathy diagnosis and lesion localization," *Computers in Biology and Medicine*, vol. 162, p. 107 011, 2023. DOI: 10.1016/j.compbiomed.2023.107011

[30] C. S. Kameswari et al., "An overview of vision transformers for image processing: A survey," *International Journal of Advanced Computer Science and Applications*, vol. 14, no. 8, 2023.

[31] V.-N. Pham, X. Sun, and H. Choo, "Disease diagnosis on fundus images: A cross-dataset study," *arXiv preprint*, 2023.

[32] M. M. Rahman, M. H. Jibon, M. A. Hossain, and S. Ahmed, "Multi-stage framework using transformer models, feature fusion and ensemble learning for enhancing eye disease classification," *Computers in Biology and Medicine*, vol. 160, p. 106 917, 2023. DOI: 10.1016/j.compbiomed.2023.106917

[33] A. Rezaee and A. Farnami, "Innovative approach for diabetic retinopathy severity classification: An ai-powered tool using cnn–transformer fusion," *Computers in Biology and Medicine*, vol. 158, p. 106 804, 2023. DOI: 10.1016/j.compbiomed.2023.106804

[34] A. Unknown, "A wireless sensor system for diabetic retinopathy grading using mobilevit-plus and resnet-based hybrid deep learning framework," *Applied Sciences*, vol. 13, no. 11, p. 6569, 2023. DOI: 10.3390/app13116569

[35] Z. Wan et al., "A wireless sensor system for diabetic retinopathy grading using mobilevit-plus and resnet-based hybrid deep learning framework," *Applied Sciences*, vol. 13, no. 11, 2023, ISSN: 2076-3417. DOI: 10.3390/app13116569 [Online]. Available: https://www.mdpi.com/2076-3417/13/11/6569

[36] S. Woo et al., "Convnext v2: Co-designing and scaling convnets with masked autoencoders," Jun. 2023, pp. 16 133–16 142. DOI: 10.1109/CVPR52729.2023.01548

[37] F. Alqahtani, M. A. Khan, and M. Sharif, "Explainable transformer-based framework for glaucoma detection using fundus images," *Computers in Biology and Medicine*, 2024.

[38] S. Chakraborty, A. Roy, P. Pramanik, D. Valenkova, and R. Sarkar, "A dual attention-aided densenet-121 for classification of glaucoma from fundus images," *arXiv*, 2024. arXiv: 2406.15113 [eess.IV].

[39] Y. Liu, H. Zhang, L. Wang, and Y. Chen, "Ca-vit: Contour-guided and augmented vision transformers to enhance glaucoma classification using fundus images," *Biomedical Signal Processing and Control*, 2024.

[40] P. U. Pandey et al., "Ensemble of deep convolutional neural networks is more accurate and reliable than board-certified ophthalmologists at detecting multiple diseases in retinal fundus photographs," *British Journal of Ophthalmology*, vol. 108, no. 4, pp. 417–423, 2024. DOI: 10.1136/bjo-2022-322183

[41] C. Playout and F. Cheriet, "Cross-dataset generalization for retinal lesions segmentation," *arXiv*, 2024. arXiv: 2405.08329 [cs.CV].

[42] Rajatha and Ashoka, "Unveiling visionary frontiers: A survey of cutting-edge techniques in deep learning for retinal disease diagnosis," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 33, no. 2, pp. 1261–1272, Jan. 2024. DOI: 10.11591/ijeecs.v33.i2.pp1261-1272

[43] D. Wang, J. Lian, and W. Jiao, "Multi-label classification of retinal disease via a novel vision transformer model," *Frontiers in Neuroscience*, vol. 17, p. 1 290 803, 2024. DOI: 10.3389/fnins.2023.1290803

[44] Y. Yan, L. Yang, and W. Huang, "Fundus-danet: Dilated convolution and fusion attention mechanism for multilabel retinal fundus image classification," *Applied Sciences*, vol. 14, no. 18, p. 8446, 2024. DOI: 10.3390/app14188446

[45] Y. Yang, Z. Cai, S. Qiu, and P. Xu, "Vision transformer with masked autoencoders for referable diabetic retinopathy classification based on large-size retina image," *PLOS ONE*, vol. 19, no. 3, e0299265, 2024. DOI: 10.1371/journal.pone.0299265

[46] S. Akhtar et al., "A deep learning based model for diabetic retinopathy grading," *Scientific Reports*, vol. 15, p. 3763, 2025. DOI: 10.1038/s41598-025-87171-9

[47] M. A. Ali, *Amdnet23: Fundus image dataset for age-related macular degeneration disease detection*, Mendeley Data, V1, 2025. DOI: 10.17632/yj35kjgrv3.1 [Online]. Available: https://data.mendeley.com/datasets/yj35kjgrv3/1

[48] X. Hu, H. Li, Y. Feng, S. Qian, J. Li, and S. Li, "Ccdformer: A dual-backbone complex crack detection network with transformer," *Pattern Recognition*, vol. 161, p. 111 251, 2025, ISSN: 0031-3203. DOI: https://doi.org/10.1016/j.patcog.2024.111251 [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0031320324010021

[49] G. Jayaraman, S. Meganathan, S. Shah, M. Anuradha, R. Sundararajan, and R. Rajakumar, "A hybrid cnn–vit framework with cross-attention fusion and data augmentation for robust brain tumor classification," *Scientific Reports*, vol. 15, Nov. 2025. DOI: 10.1038/s41598-025-28636-9

[50] Y. Liu et al., "Stmf-drnet: A multi-branch fine-grained classification model for diabetic retinopathy using swin-transformerv2," *Biomedical Signal Processing and Control*, vol. 103, p. 107 352, 2025. DOI: 10.1016/j.bspc.2024.107352

[51] D. Lv, Z. Yu, H. Zhang, and J. Xie, "Dual-backbone feature fusion for few-shot specific emitter identification under class imbalance," *IET Radar, Sonar & Navigation*, vol. 19, no. 1, e70081, 2025. DOI: https://doi.org/10.1049/rsn2.70081 [Online]. Available: https://ietresearch.onlinelibrary.wiley.com/doi/abs/10.1049/rsn2.70081

[52] K. Rezaee and F. Farnami, "Innovative approach for diabetic retinopathy severity classification: An ai-powered tool using cnn-transformer fusion," *Journal of Biomedical Physics and Engineering*, vol. 15, no. 2, pp. 137–158, 2025. DOI: 10.31661/jbpe.v0i0.2408-1811

[53] S. Wei, C. Luo, and Y. Luo, "Boosting multimodal learning via disentangled gradient learning," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, Oct. 2025, pp. 22 879–22 888.

[54] H. Yu and X. Dong, "Ensemble-based eye disease detection system utilizing fundus and vascular structures," *Scientific Reports*, vol. 15, p. 19 298, 2025. DOI: 10.1038/s41598-025-04503-5