Fusion-Based Multimodal Deep Learning to Improve Detection of Diabetic Retinopathy and Macular Edema: Integrating Retinal Imaging, Clinical Data and Systemic Biomarkers

by

MD. Raisul Islam
21301406
Samir Yeasir Emon
21301413
Nowshin Sumaiya
21301276
Sakib Khan
21301278
Farhan Labib
21301238

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
October 2025.

Declaration

It is hereby declared that

1. The thesis submitted is my/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:

MD. Raisul Islam

21301406

Samir Yeasir Emon

21301413

Nowshin Sumaiya

21301276

Sakib Khan

21301278

Farhan Labib

21301238

Approval

The thesis titled "Fusion-Based Multimodal Deep Learning to Improve Detection of Diabetic Retinopathy and Macular Edema: Integrating Retinal Imaging, Clinical Data and Systemic Biomarkers" submitted by

1. MD. Raisul Islam (21301406)
2. Samir Yeasir Emon (21301413)
3. Nowshin Sumaiya (21301276)
4. Sakib Khan (21301278)
5. Farhan Labib (21301238)

of Summer, 2025 has been accepted as satisfactory in partial fulfillment of the requirement for the degree of B.Sc. in Computer Science in October 10, 2025.

Examining Committee:

Supervisor:
(Member)

Dr. Jannatun Noor Mukta

Assistant Professor, CSE
Director, BSDS
United International University

Co-Supervisor:
(Member)

Md. Saiful Bari Siddiqui

Senior Lecturer
Department of Computer Science and Engineering
BRAC University

Thesis Coordinator:
(Member)

Dr. Md. Golam Rabiul Alam

Professor
Department of Computer Science and Engineering
BRAC University

Head of Department:
(Chair)

Sadia Hamid Kazi, PhD

Chairperson
Department of Computer Science and Engineering
BRAC University

Abstract

Diabetic Retinopathy, a silent threat to vision, is one of the major causes of vision impairment worldwide, where the retina of the eye is damaged before noticeable symptoms appear. Accompanying DR (Diabetic Retinopathy), DME (Diabetic Macular Edema) frequently develops, stating both are overlapping ocular conditions threatening visual acuity that can be effectively diagnosed by analyzing retinal images. However, relying only on a single modality has proven inadequate accuracy in distinguishing between DME and DR. Traditional diagnostic methods are employed primarily on fundus imaging, OCT (Optical Coherence Tomography), or OCTA (Optical Coherence Tomography Angiography). To date, single modality alone fails to provide a complete contextual understanding necessary for precise classification.This work proposes to offset the limitation by developing deep learning architectures that leverage several image modalities to improve classification performance and yield context-aware outputs. Specifically, the work proposes to develop personalized Convolutional Neural Networks (CNNs) driven mainly by superior fusion methods such as Multi-Head Self-Attention (MSA) Fusion, Gated Fusion, and Feature-wise Linear Modulation (FiLM) Fusion, with model interpretability at each step. The multimodal DR and DME classification strategy proposed architecture fuses two forms of image data or biomarkers so that the model may accommodate both structural and context-specific differences. Our proposed architecture has achieved an impressive accuracy of 95.52% and an F1-score of 0.975, outperforming the existing benchmark. Furthermore, this accuracy is achieved with a lower parameter count of 1.75 million and 2.57 million, with faster inference times of 19.289 ms and 19.843 ms for the two architectures, respectively, setting a state-of-the-art benchmark in the medical field.

Keywords: *Diabetic Retinopathy (DR), Diabetic Macular Edema (DME),Fundus Photography, Optical Coherence Tomography (OCT), Optical Coherence Tomography Angiography (OCTA), Good Health and Well-being(SDG 3), Industry Innovation and Infrastructure(SDG 9), Convolutional Neural Networks (CNNs), Multi-Head Self-Attention (MSA) Fusion, Gated Fusion, Feature-wise Linear Modulation (FiLM) Fusion*

Acknowledgement

To start, all the praise to the almighty Allah for whose help we are able to complete everything in time.

Then, many thanks to our supervisor, Dr. Jannatun Noor Mukta Ma'am and co-supervisor Md. Saiful Bari Siddiqui sir, who have helped us greatly both with ideas and mentally.

We are also thankful to our parents. Without their thoughtful support, it might not have been possible. For their grateful prayers, we are on the edge of graduation.

Table of Contents

List of Figures

Chapter 1

Introduction

1.1 Background

DR and DME have become one of the most critical causes of vision impairment and blindness in the modern world. DR and DME are vascular complications of diabetes. This condition results from the dysfunction of retinal blood vessels caused by chronic hyperglycemia. Approximately 34.6% of diabetic patients' eyes can be affected by DR, with a vision-threatening rate of 10.2% [54]. Statistics [12] reveal that it is the fifth leading cause of global blindness. The World Health Organization (WHO) conducts a survey to determine the number of individuals affected by this particular disease. As per their report from 2013, around 382 million diabetic patients have been affected by this ocular disease with an alarming increasing rate [15]. Also, DME is a severe version of DR, which occurs due to fluid leakage in the retina, which is responsible for blurry vision. Over the years, numerous studies and advancements have been made in the field of DR and DME, which primarily rely on traditional retina imaging, specifically the fundus image of the retina. Among them, there are some works that focus on DR and DME lesions, specifically on the state of the lesions for early disease detection. Additionally, studies related to DR and DME focus on quantitative biomarkers for classifying the severity levels of DR and DME. In short, all these approaches are based on the early detection and severity levels. Currently, further technological advancements are needed in this medical sector.

1.2 Rationale of the Study Motivation

DR and DME have become one of the most significant challenges in global blindness. In a study, it was proven that after 20 years of diabetes, 99% of patients with type 1 diabetes mellitus will develop DR, and 60% of patients with type 2 diabetes mellitus will develop DR [83]. If we can detect this condition early, then we can intervene in a timely manner. Even if it is too late, we can help classify the severity levels and the growth factor of that particular level. With the increasing need, technology is now changing the dimension of this medical diagnosis process. In particular, computer vision has the highest impact on this technological development. Earlier, SVM, random forest, and traditional neural networks had been used. In recent years, deep learning neural networks, such as CNNs, have made significant advances in this process. For feature extraction, several pre-trained models are also

being used. By analyzing fundus images, these models demonstrate good performance with high accuracy and minimal cost. However, relying solely on fundus images may miss important features, resulting in an incomplete evaluation. Moreover, researchers emphasize that models trained using unimodal techniques can lack external validation and exhibit spectrum bias [11]. Furthermore, manual screening of fundus images by ophthalmologists can be biased and complex. These findings suggest that a robust approach is needed that takes other data into account, including fundus images. In our research, we are focusing on fusing the same instances of retinal fundus images with OCT images and OCT quantitative biomarker data to establish a multimodal system for the classification of DR and DME. Since most previous approaches by other researchers focus solely on the retinal fundus image, which has various limitations due to the variability among different patients.

1.3 Problem Statement

Although there have been numerous studies related to DR detection, it still remains one of the major causes of vision impairment and global blindness. Therefore, advancement in this research field remained necessary. Previous methodologies that focused on DR and DME detection have some limitations. For instance, overreliance on fundus images can lead to inaccurate early detection and classification. The reason for this is the variability and diversity of patients. Additionally, fundus images provided by ophthalmologists can be subjective and may miss important signs. Furthermore, there is a huge data gap in the clinical context. There are other important data points that can influence the result, such as Clinical Scores and Biomarker Data. These key factors are often ignored in unimodal approaches. This gap is addressed through our research by introducing a multimodal method. This research utilizes the latest tools and methods, leveraging current advancements in computer vision by introducing a fusion-based custom hybrid architecture. It generates more reliable and accurate results by fusing the outputs with attention mechanisms, such as cross-attention and self-attention networks.

1.4 Objective

The objective of this study is to develop, evaluate, and analyze a multimodal deep learning system that fuses fundus images, OCT scans, and biomarkers for the accurate classification of DR and DME within the specified research timeline. This research intends to provide explainable and interpretable predictions that can assist the medical personnel with decision-making, resulting in medical relevance and improved diagnostic accuracy.

1.5 Methodology in Brief

Using deep learning methods, fundus images, OCT, OCTA, CFP, and ICGA can be analyzed to train the deep models to automate the DR and DME detection and staging process. The most popular deep learning networks are known as CNNs that contain several convolutional fully connected layers. In our research, we are trying

to use retinal fundus images, OCT images (3D images). Alongside image data, we also need tabular biomarker data and clinical data for multimodal fusion methods. We are also trying to extract the DR and DME lesions, which can help us in the decision-making of detection and classification.

The techniques we are going to use are deep learning architectures and transfer learning architectures like ResNet18/50 [44], [59], InceptionV3 [21] and EfficientNetB0/10 [42], and some separate Custom CNN Encoders. Multiple models will be utilized during the project, including a retinal image model, a clinical data/biomarkers model and a fusion model.

1.6 Scopes and Challenges

1.6.1 Scopes

This research helps to improve the accuracy and efficiency of DR and DME detection through a multimodal deep learning framework. Addressing the limitations of previous models and facing challenges like variability in patient data and imaging consistency, we work to make the model more accurate than the existing models. The research uses the latest advancements, including multimodal models that have the ability to process multimodal data efficiently and architectures that allow for seamless integration of textual and imaging data, making them better suited for this research. Our research prioritizes affordability and ease of use so that people can use it at a small cost. As our research is now primarily based on public datasets such as OLIVES for training and testing, in the future, this opens paths for further research. In the future, real-world data can be added to make our research better and has the potential to improve medical image analysis, which helps prevent vision impairment in diabetic patients globally.

1.6.2 Challenges

The main challenge in our research is to integrate or fuse the various data domains into one multimodal network and ensure their availability. If it is available, ensuring patient data privacy and security is important. The availability of the dataset is the main challenge because the difficulty of getting a multimodal dataset is very high, and most datasets that have patients' fundus images have no clinical data provided. Medical records often contain errors and missing values; therefore, data pre-processing requires significant expertise. Combining data from various sources and tracking patients to gather extensive data requires patient cooperation and significant resources. In other words, most of the data is image data focusing on DR and DME fundus images and lesions. On the other hand, even if we manage to get a multimodal dataset or create multimodal data using various datasets, there is high computational complexity in the proposed framework.

1.7 Thesis Contribution

Every researcher wishes to make a significant contribution to their specific area of study. The main purpose is to introduce something new, whether it's a new concept, an improved approach, or improved comprehension of an existing problem.

Our research also has some major contributions. First of all, we work on a fusion-based multimodal deep learning framework that integrates retinal fundus images, clinical patient data, and subsequently relevant biomarkers to detect DR and DME. A simple but effective combination approach is proposed for integrating information from various data sources, allowing the system to understand both the visual as well as clinical aspects associated with the disease. Furthermore, our proposed architecture has achieved an impressive accuracy and F1-score. Our research also highlights the way multimodal fusion could be helpful in more accurate AI-based diagnosis. The study highlights how AI could potentially help doctors in early identification and treatment planning, which could decrease the chance of blindness and improve the condition of patients' eyes.

In short, our research strengthens diagnostic accuracy, makes DR and DME detection at a lower cost, and establishes a solid foundation for future studies.

1.8 Thesis Organization

This research paper is organized into eight chapters. First of all, in the introduction chapter, research background, study motivation, problem statement, research objective, methodology, scopes and challenges are provided. Then, in the literature review chapter, our research preliminaries, reviews of existing papers are provided. This chapter mainly highlights earlier studies and points out the research gap that this study is trying to resolve. We reviewed some survey papers, papers related to the unimodal deep learning approach, as well as the multimodal deep learning approach. Furthermore, in the requirements, impacts and constraints chapter, it has functional requirements, non-functional requirements, constraints, and compliance. Also, this chapter contains societal impact, risk management, environmental impact, economic analysis, and ethical issues.

Additionally, in the dataset chapter, everything about the dataset we used is presented. Also, the concept of data pre-processing steps, dataset splitting, formatting etc. is given. Again, the methodology chapter contains the methodology, ablation study, and details about our proposed architectures. Our proposed architectures are Dual Vision V1/V2 architecture and OCTbiO architecture. Moreover, the architecture configurations and results chapter provides the architecture configuration, Setup, and results of the architectures. Meaning, the results of Dual Vision V1/V2 architecture and OCTbiO architecture are given. Again, the architecture explainability chapter highlights the explainability of the Dual Vision V1/V2 architecture and the OCTbiO architecture. Lastly, the conclusion chapter gives ideas about contributions to the field, limitations of the research and recommendations for future work.

Chapter 2

Literature Review

2.1 Preliminaries

This research focuses on building a multimodal framework that includes retinal images, clinical data, and systemic biomarkers. The theory behind our approach is that there is significant variability among patients, and this variability cannot be addressed or introduced solely by fusing different modalities of the retinal image. In addition to them, we need clinical data and a systemic biomarker to introduce variability. Some research theories, such as lesion-based DR and DME detection, and patient medical records, can improve the possibility of early detection.

In the medical sector, multimodality refers to the concept of integrating different types of modalities within the same class or across multiple classes, which can facilitate more precise and comprehensive diagnosis. All research in DR and DME detection and classification has found different image modalities, which have several limitations due to the lack of information about the unique individuality. For this, we introduce multimodal architectures that focus on normalizing values from different modalities and aligning all metadata corresponding to these modalities.

The models used for the detection and classification of DR and DME include ResNet18/50 [44], [59], the cascade R-CNN model [33], EfficientNetB4 [33], a single pre-trained ConvNet [23], VGG16 [23], NASNet [23], FusionNet [23], [51], Xception [23], Inception [21], and MM-RAD [77]. OCT, fundus retinal images and Electronic Health Records (EHR) using the Long Short-Term Memory (LSTM) model [68] are suitable for sequential data processing, and feature extraction in this research is performed using Local Binary Pattern (LBP) [68], which enhances the retinal features such as blood vessels and other structural abnormalities.

2.2 Review of Existing Research

After determining the problem statement, we review some renowned research papers to better understand this field. These articles can be divided into three parts based on research objectives. Firstly, we review some survey papers to learn about DR and possible existing technologies (old and new) for DR detection. Then, we obtain some papers that utilize the unimodal deep learning approach. Finally, we find

some literature where the multimodal deep learning approach outperforms existing methods.

2.2.1 Survey Papers

A survey article [40] mentions common retinal diseases such as DR and AMD, where traditional methods such as CFP, FAF, OCT, and OCTA, with CFP being the gold standard, and detection approaches including SVM, random forest, and 2-layer neural networks. Models such as AlexNet, VGG16, ResNet, and others are effective for transfer learning, addressing the challenge of insufficient training datasets. Filters such as Gaussian, median, wavelet, and spatial domains are used for data preprocessing and noise reduction. Future improvements could involve integrating AI models into mobile applications to improve accessibility and detection efficiency. Deep learning methods for medical image segmentation, focusing on three categories (FCN, ROI, and weakly supervised methods), are discussed by Qureshi et al. [50] FCNs do not require domain experts, while ROI methods do, as experts are familiar with the regions of interest, and weakly supervised methods are semisupervised. The authors highlight that deep learning algorithms are highly accurate in detecting retinal features and lesions. This survey paper [73] focuses on deep learning techniques for DR using fundus images and highlights the effectiveness of deep learning models such as ResNet, VGGNet, InceptionV3, DenseNet, and ensemble methods in binary and multiclass DR classification. Multiclass classifications, such as ResNet50 and DenseNet, achieve high accuracy with the aid of certain methods, reaching a 95% accuracy for severity classification. Compared to previous and older techniques, deep learning models offer higher sensitivity. Preprocessing techniques, such as image enhancement, argumentation, and GAN-based data generation, improved the performance of the model. The study [61] by Christensen et al. finds that a higher retinal venular length-diameter ratio (LDR) and arteriolar fractal dimension (Fd) predict a lower chance of full treatment response (FTR) in neovascular age-related macular degeneration (nAMD). The study sample size is small and future research should focus on larger, longer-term studies to validate these findings and explore personalized treatment regimens based on retinal vascular markers.

2.2.2 Unimodal Papers

Now that we have gathered ideas about the overall process, we review some papers that implement the models on real datasets. Firstly, numerous studies try to solve the problem by relying on a single modality, be it fundus images, OCT, or UWFFA.

For example, Mohammad T. et al. [41] classify DR using an unimodal approach with OCT/OCTA analysis to reduce the logistic challenges and costs. They make their dataset from the Casey Eye Institute, USA. The SSADA algorithm is implemented to detect the blood flow for each volumetric OCT/OCTA. Motion artifacts are reduced through two raster scans that are merged by an orthogonal registration algorithm. A 3D CNN architecture is employed for DR classification, and batch normalization is applied after each convolutional layer. The model performs slightly better for rDR than vtDR. The overall accuracy for DR classification is 91.52%, and for vtDR, it is 87.39%. Again, in this paper [49], they use the ViT (Vision Trans-

former) model, which is designed for the automatic classification of DR. This model is trained using the FGADR dataset to determine an image's global context, and the model has been optimized with AdamW. This model achieves superior results. The accuracy of this model is 0.825. Lesion splitting is not handled by their ViT model. Furthermore, another research [65] combines Explainable AI (XAI) with deep learning models that analyze retina images to determine signs of DR. Retina pictures are preprocessed to improve quality and avoid noise. The diagnostic accuracy of this model is 94% better than traditional methods. However, the limitation of this research is that XAI's technical addition could increase computational complexity.

A paper [36] introduces an AI-based app for early detection and diagnosis of diabetic retinopathy (DR). The app is built on TensorFlow deep learning, helps patients predict DR severity levels, and provides results via email. The app achieves an accuracy of 91.44%, 91.35% specificity, and 92.51% sensitivity. Abbood et al. [35] introduce an Algorithm to improve the quality of retina images by reducing the noise and improving the contrast. This model has 2 main stages: crop images to remove irrelevant details, conduct a circle crop, and then apply Gaussian blurring. CNN is used here to diagnose diabetic retinopathy and improve retinal pictures. This proposed work is 92% accurate on the Kaggle dataset and 93.6% accurate on the MESSIDOR dataset. However, this method fails to solve the illumination problem. On the other hand, in this article [27], Al-Antary et al., alongside their teammate, design a computer program MSA-Net (Multi-Scale Attention Network) to observe eye images, in big pictures with small details to find signs of DR. They implemented it on two popular accessible datasets, APTOS as well as EyePACS. The accuracy of EyePACS is MSA-Net with MM & without MM, respectively 0.799, 0.775. Kappa score of EyePACS is MSA-Net with MM & without MM, respectively 0.878,0.842. Furthermore, the accuracy of APTOS is MSA-Net with MM & without MM, 0.846, 0.821. Kappa score of EyePACS is MSA-Net with MM & without MM, respectively 0.896,0.857.

Additionally, researchers [34] show that during the initially applied treatment of DR, the 7-standard area picture using ultra-wide-field fundus scanning beats both the optic disc as well as macula-centered picture statistically. Their ETDRS 7SF based on photos DR identification technique establishes accuracy of 83.38 $\pm$ 0.47. On the other hand, the ETDRS F1–F2 based on photos DR identification technique establishes accuracy of 80.60 $\pm$ 0.54.The limitation of their study is that, in the UWF photography, they come up with a ROI based on the entire retinal region's DR recognition by the ETDRS 7SF. Some unimodal approaches have even proven themselves better at training with fewer images. Sahlsten et al. [21] focus on five grading systems as well as clinical methods used by doctors. In this paper, they build a model using the Inception-v3 deep learning framework. It needs fewer images to train compared to older models but can perform better as this model has a sensitivity of 89.6% and specificity of 97.4% for detecting DR, which is better than models from previous studies. Jabbar et al. [59] implement a deep learning-based and severe state lesions approach for the classification of DR severity levels by using GoogleNet and ResNet models based on an adaptive particle swarm optimizer (APSO). APSO is utilized for enhanced feature extraction and refined model tuning to improve the model's performance. They use the hybrid models to extract features

from lesions, learn the lesion's pattern, and feed it into different machine learning classifiers like the Decision Tree, SVM etc. EyePACs dataset is used and all images of this dataset are preprocessed into (224 x 224 x 3). Their radial SVM overall performance accuracy is 94%, precision is 97%, and F1-score is 96%. Ultra-widefield fluorescein angiography (UWF-FA) is also useful for identifying the severity level of DR. In another research, Attiku et al. [43] examine the difference between a complete UWF picture of non-stereoscopic UWF pictures and the DR level of severity that is established when only the ETDRS 7-field region has been taken into consideration. By analyzing 157 patients, they discovered that the severity of DR can be influenced by looking at the retina as an entire thing instead of only the ETDRS 7 regions.

Alyoubi et al. [26] proposed a system that combines CNN512 and YOLOv3 models and shows promise in advancing DR diagnosis. They present a fully automatic system for diagnosing DR using deep learning models, achieving great accuracy. The need for sizable datasets for training, which might not be easily accessible in all clinical settings, and the possibility of decreased performance on images with different lighting and quality conditions are some of this study's limitations. There is a scope to enhance the system's robustness by incorporating more diverse datasets and improving lesion localization accuracy to assist ophthalmologists in clinical practice. There are also several types of research to detect the severity level of DR using a single modality. A related work [57] has used UWF-FA, which helps them identify and quantify relative biomarkers related to DR severity. The biomarkers they identified, nonperfusion (NP) and neovascularisation (NV), play a pivotal role in DR severity. These biomarkers enable the detection of DR severity based on the location and size of the NP and NV, identification of high-risk patients, and monitoring of DR progression. For instance, this study [76] uses the Multi-frame Medical Image Distillation (MMID) method, which significantly improves multi-label classification accuracy without requiring additional annotations. However, the research acknowledges several limitations, including the reliance on a single-direction scan image that cannot accurately diagnose certain diseases and the need for experienced ophthalmologists to interpret OCT images. Furthermore, Chen et al. [53] study the association between the retinal age gap (chronological age) and the risk of developing DR. It is validated that retinal age is one of the non-invasive biomarkers for aging. Even after adjusting for the confounding factor, this study found there is a 7% increase in the risk of DR in every additional year. Therefore, for a profound and robust analysis, feature fusion strategy or multimodal approaches can be more reliable which take these factors into account.

2.2.3 Multimodal Papers

We then review some research to learn how the introduction of multimodality changed the dimensions of this field.

El Habib Daho et al. [44] approach a multimodal framework that uses two different image modalities, which are OCTA (3D image) and UWF-CFP (2D image) with the help of the Resnet50 model, 3D Resnet50 model, and Squeeze-and-Excitation (SE) blocks for extracting relevant features to predict DR severity. Also, they use

Manifold Mixup to avoid overfitting. This paper [23] proposes a new feature blende feature extraction method that is a multimodal approach (features from different CovNets like VGG16, NASNet, Xception, and Inception ResNetV2). For model training, for DR identification, they have used binary cross-entropy, and for DR classification, they have used categorical cross-entropy. Their model performed very well in the APTOS2019 dataset, achieving an accuracy of 97.41% for DR identification and an accuracy of 80.96% for DR severity classification. This researcher Bodapati et al. [23] use both unimodal and multimodal approaches in which they introduce a single pre-trained ConvNet. VGG16, NASNet, Xception and Inception ResNetV2 models are used for transfer learning, where Xception and Inception ResNetV2 outperform other models. DNN with dropout is introduced to handle overfitting. VGG16-fc2 has a better kappa score of 70.2. This model is suitable for lesion extraction. For DR detection, the multimodal blending of VGG16-fc1 and VGG16-fc2 has a kappa score of 91.89 in max pooling, which outperforms others.

This article [33] studies multimodal deep learning techniques for multiple retinal vascular disease detection and anti-VEGF treatment requiring case identification using retinal images: fundus photography, OCT, FA with/without ICGA. The method of multimodality is introduced in DME, nAMD, mCNV, BRVO, and CRVO detection. This research uses a cascade R-CNN model for the detection phase. For classification, they use EfficientNetB4 as CNN. The AUCs of the model are 0.987 and 0.969, respectively, for retinal vascular diseases and treatment-requiring disease prediction. Bidwai et al. [51] focus on DR detection enhancement, where optical coherence tomography Angiography (OCTA) and fundus images are used. These two types of images are then fused, and the feature map from OCTA images is generated using RESNET101. Then, the fused images are processed by an AI-based Light Gradient Boosting Machine (GBM), which classifies the image as normal or abnormal retina. In this paper, the Dynamic Knowledge Sharing Algorithm (DKSA) and LightGBM classifiers are used.

Wardhani Kasim Erianda Hassan et al. [68] use a fusion approach using Optical coherence tomography (OCT), fundus retinal image & EHR using the LSTM model are suitable for this sequential data processing, and feature extraction of this research is done by LBP, which enhances the retinal features such as blood vessels and all other structural abnormalities. In this case, the multimodal fusion approach performs well using all these methods; the AUC is 0.99, demonstrating excellent generalization as well as sensitivity. Moreover, another researcher Manikandan et al. [47] improves the accuracy of DME detection using Retinal Images such as fundus Images and OCT images. They found that OCT images have a sensitivity of 96% whereas Fundus Images have a sensitivity of 94%. After a combination of both types of methods, accuracy is high and improved. However, the study has limitations in image quality and dataset size. This paper [67] reviews the use of deep learning and artificial intelligence in multimodal ophthalmology. Their study highlights the advantages of multi-modal AI, which provides more details and comprehensive information for diagnosis. Some models, such as FusionNet and MM-RAF, perform significant improvements, achieving high accuracies, such as AUC scores of 0.97 for glaucoma detection and 96% for AMD classification using modalities. AlFahdawi et al. [56] use a publicly available OIA-ODIR dataset from Shanggong Medical Tech-

nology Co. The CLAHE method with a median filter is used to improve contrast and noise reduction. For feature extraction, HRNet block and the Attention block are used. For quality and robustness improvement, the SENet block is used. For classification tasks, the DRBM model associated with the softmax layer is introduced. For the evaluation process, they use F1-score, kappa score, and AUC. Off-site test sets had an accuracy of 88.5%, 88.9% and 99.7% respectively. On-site test sets have an accuracy of 89.1%,88.9% and 99.8%.

In this paper [32], the researchers have implemented an incremental learning approach known as knowledge distillation, mainly focusing on DR lesions. They have improved the segmentation accuracy and efficiency because a very minimal amount of annotations is being used, which overcomes the challenge of catastrophic forgetting. This research paper [77] focuses on the improvement of the medical anomaly detection and localization (MADL) such as retinal diseases, more likely retinal artery occlusion (RAO), by using OCT images and fundus images. Then they improve a model called MMRAF, which combines information from both types to give good accuracy and results. ANoGAN and f-AnoGAN only work on single types of images, whereas MMRAD is better, gives accurate results, and is better at finding problem areas. This paper [37] from Information Fusion states a novel approach for detecting retinal microaneurysms (MAs) using deep neural networks (DNNs) and generative adversarial networks (GANs). Their methodology aims to provide a straightforward detection method that processes complete raw images without preprocessing the data. Although the performance depends on the availability of high-quality multimodal images for pre-training, while the method shows promising results, its generalisability to other datasets is to be evaluated considering low computational complexity. A journal [19] of healthcare engineering presents a machine learning algorithm that combines OCT and color fundus images to diagnose glaucoma with high accuracy. It has the potential to detect early glaucoma. The algorithm achieved a 10-fold cross-validation AUC of 0.963. The study has its limitations, including a relatively small sample size and samples collected from one ethnicity, which will affect the generalizability of the results.

2.2.4 Fusion Model Papers

Ghouali et al. (2022) [36] introduce an AI-based app for early detection and diagnosis of DR. The app, built on TensorFlow deep learning, helps patients predict DR severity levels and provide results via email. The app achieves an accuracy of 91.44%, 91.35% specificity and 92.51% sensitivity. Research [52] tries to identify technical failures in DR screening, and the main finding is that technical failures occur for upgradability and delay diagnosis of retinal images. They use the method of a logistic regression model, primarily and then secondarily making a subset of specific cases and analyzing them separately to identify specific causes. They find that on average the TF rate is 4.9, with a decreasing rate in recent years. The significant agents for TF they find are patient age, disease duration and presence of cataracts. 52% of cases are for cataracts. Again, a research [72] discusses extracellular vesicles (EVs), which play a vital role in DR by mediating cellular communication and disease progression through their specific cargo. Monitoring EV quantity and content offer can be a potential for DR diagnosis. EVs, particularly those from mesenchymal

stem cells (MSCs) show promise as therapeutic agents for retinal and other complex diseases in preclinical studies.

Rivera-De-la-Parra et al. (2023) [64] discuss the relationship between uric acid (UA) levels and sight-threatening DR in type 2 diabetes patients aged 18-70. UA is analyzed as a continuous and dual variable, with $\geq$ 7.8 mg/dL being the excellent threshold, and results show that higher UA levels are associated with referable DR. The study [54] discusses the relationship between blood vitamin A levels and DR risk in participants aged 40 and older. The study found that higher serum vitamin A levels (0.51–0.64) reduce the risk of DR, with great results observed in males and individuals under 60. Vitamin A's anti-angiogenesis and anti-inflammatory properties help prevent DR progression. Hasan et al. (2023) [45] propose a self-supervised segmentation-driven classification framework and focus on retinal lesions for retinopathy grading. Compared to previous or existing methods the framework achieves notable results in mean intersection over union F1-scores. Validation on seven public datasets highlights its potential for advancing autonomous retinal screening systems.

In this research [70], Gated fusion makes sure the model always concentrates on the most helpful spatial patterns by automatically adjusting how much information to extract from Graph Convolutional Network (GCN) and Graph Attention Network (GAT) features. The combined details are then sent to the temporal modeling block following gated fusion, which improves how the traffic changes over time. Again, in this article [18], for the purpose of improving low-resolution and blurry images, the gated fusion module determines how much should be taken from the "deblur" and "super-resolution" branches. Clear images are then produced by passing the combined features through an image reconstruction module. Furthermore, in order to improve medical images, this model employs two paths: a Transformer to figure out the overall structure and a CNN to collect small details. By cleverly combining both outputs, the gated fusion technique generates medical images that are sharper and more precise [80].

In this research [84], features derived from two spectroscopy techniques, such as Raman and FTIR, are combined using bilinear fusion. It helps to diagnose benign and malignant thyroid tumors more accurately. Additionally, this paper [46] chooses bilinear fusion because it makes the feature representation more powerful, allowing the system to distinguish finger veins more accurately. It improves recognition accuracy by capturing subtle correlations between smaller details and the overall vein structure.

This paper [62] emphasizes the correspondence between images of different modalities of the same patient by implementing cross modal attention (CMA) mechanism. The dataset contains CFP and OCT images of patients with retinal disease. Self-attention mechanisms on multimodal data can result in huge computational overhead and the ability to leverage relevant information from different modalities is often absent. They used CRD-net architecture that uses two independent CNNs for feature extraction and then performs an interactive fusion by CMA. This technique helps to focus on those features that cause disease such as affected lesions. On the

MMC-AMD dataset, this architecture outperformed the existing self-attention approach on multimodal data by achieving 84.66% F1-score, 79.57% kappa score and 85.62% acc.

This research[69] fuses CFP and IFP images by implementing dual system architecture Cross-Fundus Transformer (CFT) for more accurate DR grading. CFP and IFP are considered as highly correlated and often complementary data in DR detection. IFP remains clear by gray fog and CFP provides better color information. So a multimodal interaction between these two modalities can be useful for diagnosis. They designed a CFA (Cross Fundus Attention) module to integrate the patch embedding representation of the two modalities. This research performed several experiments on CIDR dataset by single modalities, self-attention of CFP and IFP and finally cross-attention of CFP and IFP. The final result concludes that the cross attention outperforms all other approaches by achieving a kappa score of 84.44%, acc of 73.47% and F1-score of 65.51%. Thus, this proves the superiority of this attention mechanism for DR grading.

Oahidul et al. worked on research [58] for retinal image classification using multihead self-attention by using Vision transformer models. To achieve an accurate diagnosis in lesser time without dependency on degeneration, their VIT model proved to be applicable in this task. A single matrix multiplication makes the input dimensions smaller and generates a series of patches and goes to the transformer encoder with positional embedding. This proposed methodology ensures that the spatial connections among several patches are retained. They achieved an accuracy of 96.13%, F1 rating of 0.93, ROC-AUC of 0.969 and recall of 0.95. Future research should work with a variety of training datasets and create more efficient iterations of vision transformers.

Dulal et al. worked on cattle identification using the MSA feature of CNN and transformer [75]. Their primary goal was to capture the connection of different fusing features while retaining the originality unchanged. Their proposed system improves the accuracy in addition and concatenation. Their research found the optimal performance of fused features by combining the Q,V for transformers and K for CNN. For feature extraction, they used a modified model of ResNet50 for local features and ViT for lower layers. Feature fusion through multihead self-attention was conducted for a single and unified representation of features from different sources. Multihead self-attention helped to focus on the feature maps specifically. Across the spatial domain, their model helped to focus on the important spatial regions. This model was validated by 2 datasets flower12 and CIFAR10 and they got an accuracy of 99.76 and 99.46. And proves the usage of MHSA technique to be beneficial to capture the complex dependencies of CNNs and vision transformers.

This work [30] enhanced the understanding of collaborative multi-head attention layers that learn shared projection by using multiple heads. They conducted an experiment that concludes the possibility of re-parametrize of a pretrained multihead attention layer into the collaborative layer. Their model reduced the size of the projection in one fourth while keeping the same performance and speed. A typical attention mechanism concatenates the heads to obtain multihead, while this work

improved the capacity of the model by combining the heads. The proposed system learns key/value projection for every head at a single time and after that the heads use re-weighting of the projections. This methodology introduced a parameter and computation efficiency. This method optimized the redundant learning of key/query representations of concatenated multihead attention models.

2.3 Summary of Key Findings

After reviewing the existing research papers, we find some interesting patterns. Firstly, relying on a single modality for classifying and staging DR is not feasible. So, a multi-modal fusion process will be a good solution to build a more accurate, feasible, and robust system. Secondly, using a dataset containing multiple types of imaging data alongside the patient's age and medical history, instead of only imaging data will give more reliable output. Finally, introducing transfer learning will be a smart approach to solving our problem.

The following table 2.1 summarizes the key findings of the existing research, which highlights author and year published, modality type, method used, dataset and results.

Author and Year Published	**Modality Type**	**Methods/ Models Used**	**Datasets**	**Results**
Nazih et al.(2024) [49]	Unimodal	ViT(Vision Transformer), optimizer: AdamW.	real-world FGADR dataset	F1-score, accuracy, precision, recall is 0.825 and balanced accuracy, AUC, and specificity is respectively 0.826, 0.964 and 0.956
Jabbar et al.(2024)[59]	Unimodal	GoogleNet, ResNet model, adaptive particle swarm optimiser (APSO), Decision Tree, SVM	EyePACs dataset	accuracy: 94% , precision: 97%, and F1-score: 96%
Zang et al.(2022)[41]	Unimodal	SSADA algorithm, orthogonal registration algorithm, DR classification : 3D CNN architecture	Casey Eye Institute, USA	rDR classification : 91.52% , vtDR was 87.39%.

Author and Year Published	**Modality Type**	**Methods/ Models Used**	**Datasets**	**Results**
Shahzad et al.(2024)[65]	Unimodal	Explainable AI (XAI) with deep learning models	Diagnosis of Diabetic Retinopathy By CNN (PyTorch)	accuracy of this model is 94%
Sahlsten et al.(2019)[21]	Unimodal	Inception-v3 deep learning framework	Digifundus Ltd, DR screening and monitoring services in Finland.	sensitivity of 89.6% and specificity of 97.4%
Daho et al. (2023)[44]	Multimodal	Resnet50 model, 3D Resnet 50 model, Squeeze-and-Excitation (SE) block, Manifold Mixup	(EviRed) project	mild NPDR (0.8566), moderate NPDR (8037), severe NPDR (0.7922), PDR (0.8820)
Bidwai et al.(2024)[51]	Multimodal	RESNET101, Light Gradient Boosting Machine (GBM),Dynamic Knowledge Sharing Algorithm (DKSA)	Natasha Eye Care and Research Institute database of Pune Maharashtra	(AUC) of 0.911
Bodapati et al.(2020)[23]	Multimodal	Single pre-trained ConvNet, VGG16, NASNet, Xception and Inception ResNetV2	APTOS 2019	VGG16-fc1 and VGG16-fc2 had a kappa score of 91.89

Author and Year Published	**Modality Type**	**Methods/ Models Used**	**Datasets**	**Results**
Wardhani et al.(2024)[68]	Multimodal	Long Short-Term Memory (LSTM), Local Binary Pattern (LBP)	Diagnosis of Diabetic Retinopathy (Kaggle), DRYAD dataset, Diabetic Retinopathy Retinal OCT Images“ by Gholami, Roy, and Lakshminarayanan (2018).	OCT and LPB: AUC of 0.98, Fundus and LPB: AUC of 0.96, Multimodal: AUC of 0.99
Kang et al.(2021)[33]	Multimodal	Cascade R-CNN, EfficientNetB4 as CNN	fundus photography, OCT, and FA/ICGA (2013 - 2019) collected from Chang Gung Memorial Hospital, Linkou Medical Center	retinal vascula r diseases AUC : 0.987 treatment-requiring disease prediction AUC: 0.969
Li et al.(2024)[77]	Multimodal	Segment Anything Model(SAM), MMRAD	DRIVE, AIDE, OCT	MMRADAUC: 0.825

Author and Year Published	**Modality Type**	**Methods/ Models Used**	**Datasets**	**Results**
ALFahwadi et al.(2024)[56]	Multimodal	DRBM model with softmax layer	OIA-ODIR dataset Shanggong Medical Technology Co	offsite Test Set: (F1-scores : 88.56 %, Kappa scores : 88.92 %, AUC : 99.76%, final scores: 92.41 %) and onsite test set: (F1-scores : 89.13 %, Kappa scores : 88.98 %, AUC : 99.86 %, final scores: 92.66 %)

Table 2.1: Summary of related works on unimodal and multimodal

Chapter 3

Requirements, Impacts and Constraints

3.1 Final Specifications and Requirements

The research objective is to develop an effective and explainable multimodal fusion model for classifying DR and DME. The system uses three types of data: Fundus images, OCT scans, and biomarker data. By combining these different data sources, the model aims to outperform single-modality models. The combination of structural and functional retinal data through this combination ought to present a more comprehensive and credible diagnosis that can be utilized in clinical practice.

3.1.1 Functional Requirements

- **Data Handling:** The system must be able to load and pre-process the data and must be capable of handling the noisy and missing data. Additionally, the system should be capable of undergoing contrast enhancement, normalization, and resizing.
- **Model Development:** The system should be able to train the deep learning model by extracting and learning from each feature. Also, the models should optimize parameters using a defined loss function.
- **Model Evaluation:** System shall evaluate model performance by the standard metrics such as accuracy, F1-score, AUC, specificity, sensitivity, precision, recall, and so on.
- **Inference and Deployment:** The System should accept unseen data and give a prediction for those. The system also generates a visual representation of the prediction.
- **Reproducibility:** The configuration of experiments, the training history records, and the performance metrics are stored in the system.

3.1.2 Non-Functional Requirements

- **Scalability:** The model should be flexible enough to include new data sources or imaging types in the future with minimal changes.

- **Efficiency:** Training and testing time should be optimized so that the model can be applied in real clinical settings.
- **Generalizability:** The model must be effective in patients of various backgrounds and images of various devices.
- **Interpretability:** The system must clearly show how it reaches its decisions, helping doctors trust and understand the results.
- **Security and Privacy:** Training should be done anonymously with all information about patients. The system must adhere to the regulations on data protection and local medical ethics.

3.1.3 Constraints and Compliance

One of the main challenges of this project is the limited availability of paired multimodal datasets. Fundus, OCT, and biomarker information should be aligned in each patient record. The need decreases the overall quantity of available data and is potentially relevant to the generalization of the model. This study aims to utilize the time series quality of the dataset to manage this problem.

3.2 Societal Impact

Our suggested multimodal fusion model for DR and DME classification is an important advancement toward more dependable and accessible ophthalmic diagnosis.

Early and proper detection of DR, DME may drastically decrease the rate of vision loss and blindness in diabetic patients. Automated, multimodal AI systems can help diagnose patients in hospitals and eye screening initiatives, especially in areas where professional ophthalmologists are relatively few. This immediately benefits public health by increasing both the availability and affordability of advanced eye disease identification.

3.3 Environmental Impact

Our research has minimal noticeable environmental impact in this stage because it produces no physical products, chemicals or waste from the laboratory. But, in the future, this study could potentially have a good environmental impact. Early and precise identification of DR and DME might prevent excessive imaging tests, hospitalizations, and treatments. This reduces the total environmental effect of medical diagnostics, the overall amount of energy used in healthcare systems, and the need for medical equipment.

3.4 Risk Management

There are several risks involved in developing and running a multimodal medical artificial intelligence system. These risks should be carefully evaluated. The first and foremost challenge is the possibility of data being incorrect and imbalanced.

Image quality, acquisition settings, and modality availability can vary across multimodal datasets. Incorrect results could be obtained through the fusion method whenever one modality is absent or performs poorly. In order to tackle these difficulties, researchers make sure that strategies that allow the model to gracefully handle missing modalities are utilized along with strong pre-processing, normalization, and data augmentation procedures.

Another key component of risk management is ensuring the security of personal data and confidential information. It is crucial to guarantee that each patient's data is handled properly since healthcare AI systems are trained on sensitive data. Our recommended approach doesn't involve private data, such as patient names, ages, BMI, and so on. Only medical data related to eye health, such as fundus images, OCT scans, and biomarker readings, are utilized. Our method guarantees the highest level of privacy of the patient's personal data while maintaining ethical and regulatory information security requirements.

3.5 Economic Analysis

In ophthalmology, the use of a multimodal artificial intelligence technology has tremendous economic benefits. Automated AI evaluation techniques allow hundreds of images to be evaluated quickly and accurately. Additionally, it significantly reduces the workload for ophthalmologists. This reliability enables clinical professionals to focus on more complex problems that require human judgment. So, it raises the healthcare organization's productivity. Again, by identifying the progression of the illness early on, patients can take advantage of timely treatment. Also, it minimizes the need for expensive procedures like intravitreal injections and laser therapy.

From a technological point of view, establishing multimodal AI frameworks may additionally promote the expansion of the medical technology industry, opening up new opportunities for emerging companies, research collaborations, and AI-powered health products. Thus, economically, our suggested method not only lowers the long-term financial burden of diabetes blindness but also encourages the establishment of sustainable, cost-effective health service ecosystems.

3.6 Ethical Issues

Ethical considerations are the fundamental basis of all medical AI research. Our suggested multimodal system has been developed and executed, complying with the principles of privacy, fairness, and transparency. Our suggested approach doesn't demand any personal information; it simply needs medical data on eye health, such as fundus images, OCT scans, and biomarker readings. Another ethical factor is transparency and human monitoring. In medical diagnostics, doctors need to know how an AI system makes its decisions before relying on it. Importantly, the AI model wasn't developed to replace ophthalmologists, but rather to supplement their decision-making by providing an additional viewpoint that improves diagnosis quality.

Additionally, the ethical responsibilities include ensuring equal opportunity. Our plan is, advantages of multimodal AI need to be spread beyond advanced healthcare centers to rural and underprivileged areas as well. Ensuring affordability and availability can help to prevent a technological gap in medical services, which will unlock AI's true potential for benefiting everyone.

Chapter 4

Dataset

4.1 Introduction

The OLIVES dataset, or Ophthalmic Labels for Investigating Visual Eye Semantics, is a rich medical dataset developed to advance research in diagnosing and treating eye diseases, particularly DR and DME. It is a joint project between Georgia Tech scientists and ophthalmologists of Retina Consultants of Texas. In this paper, the OLIVES dataset is introduced as a structured set of data that promotes clinical research in computer vision and machine learning for ophthalmic diagnosis [39]. The data were generated from two real clinical trials, PRIME and TREX-DME, conducted between 2013 and 2021. What is special about OLIVES is the magnitude of data it presents in a single location. It has fundus photos, OCT scans, and other clinical measurements with expert annotations that doctors would refer to when evaluating eye health. The eyes of each patient were followed during numerous visits, in some cases as many as 16 in a year or more. Images and test results were gathered at each visit, and then graders (with training) added labels. This combination of images, clinical scores, and disease labels renders OLIVES one of the first datasets to correlate all of these aspects within a single, cohesive entity.

4.2 OLIVES Dataset MAP

The OLIVES dataset offers a multimodal resource on retinal diseases, complementing high-resolution imaging with a structured clinical background and expert manual biomarker annotations. A visit is enhanced using carefully marked out 16 biomarkers, including fluid accumulation, structural disruptions, and other pathological signs, totaling 150,528 labels. The combined approach of quantitative and qualitative validation of disease indicators would facilitate both the conventional statistical study of retinal pathology and the construction of more complex models. The dataset combines three types of data: OCT scans, fundus images, clinical data, and biomarker labels. The following writings provide more information about these three types.

4.2.1 Optical Coherence Tomography (OCT)

The number of OCT scans that were added per visit was approximately 49 per eye for each patient's weekly visit. It reached a total of 78,189 OCT scans in the dataset.

OCT is the most data-rich imaging modality in this research, since it offers accurate cross-sectional viewing of the retina.

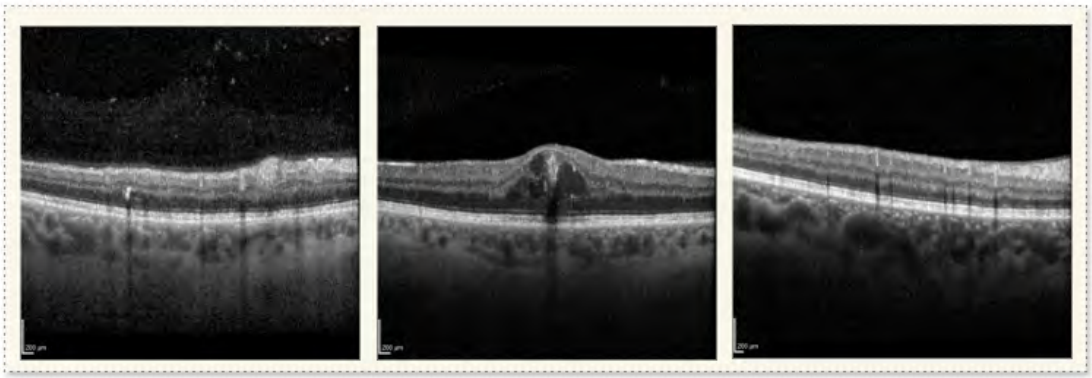

Figure 4.1: Examples of OCT slice images

4.2.2 Fundus Photography

There were 2 fundus images per eye (OD and OS) for each patient visit. The dataset comprises a total of 1,268 fundus images. Fundus images would add a superficial examination of the retina to the structural information available in OCT.

Figure 4.2: Examples of fundus images

4.2.3 Biomarker and Clinical Data

There were multiple biomarkers related to the OCT images of each patient. 49 OCT slices correspond to their OCT Biomarker Tabular values. The Biomarker tabular values consist of 13 unique biomarker binary presences and 2 unique clinical eye scores. A total of 15 biomarker and eye score values, each strongly correlated to the OCT image modality and the Disease labels. The biomarkers are listed below with their details:

Biomarker Name	**Correlation with OCT Images**	**Dataset Value**
IR hemorrhages	Blood Spots inside retinal layers which are small	0/1
IR HRF	Bright Spots inside retinal layers which are small	0/1
Partially attached vitreous face	The Gel situated at the back surface of the vitreous face	0/1
Fully attached vitreous face	Attached vitreous face to the Retina	0/1
Preretinal tissue/hemorrhage	Blood and Abnormal tissue which is located at the frontal part of the retina	0/1
Vitreous debris	Found in the vitreous face as floating particles	0/1
VMT	disorder where the vitreous pulls on the macula (central retina), causing vision distortion and edema.	0/1
DRT/ME	A widespread enlargement of the retinal tissue that typically results from fluid leakage	0/1
Fluid (IRF)	fluid pockets (cyst-like areas) that form inside the layers of the retina.	0/1
Fluid (SRF)	Fluid buildup beneath the retina	0/1
Disruption of RPE	RPE layer disruptions or anomalies	0/1
PED (serous)	Fluid accumulation beneath the RPE layer causes it to peel away from the underlying Bruch's membrane.	0/1
SHRM	An abnormal substance under the retina that shows up as light on OCT, such as blood, scar tissue, neovascular membranes, or fibrin	0/1

Table 4.1: Biomarker feature table

Eye Score Name	**Correlation with OCT Images**	**Dataset Value**
CST	With corrective lenses (glasses or contacts), a patient can attain the finest vision possible.	Min score = 156 Max score = 715
BCVA	Measured on OCT, the average retinal thickness in the middle 1-mm region (macula) is used to track edema.	Min score = 41 Max score = 97

Table 4.2: Visual eye score feature table

The OLIVES dataset is a combination of two large long-term eye studies named PRIME and TREX-DME. The PRIME study involved a sample of patients with DR of 40 patients. These patients have visited the clinic for two/three years, also these visits were designated as W0, W4, W8, etc., up to W104. Images of both eyes (OD for right, OS for left) were obtained by the doctors at every visit. The number of visits exceeded 20 for some patients, allowing us to make a profound observation of how their eye condition evolved over the years.

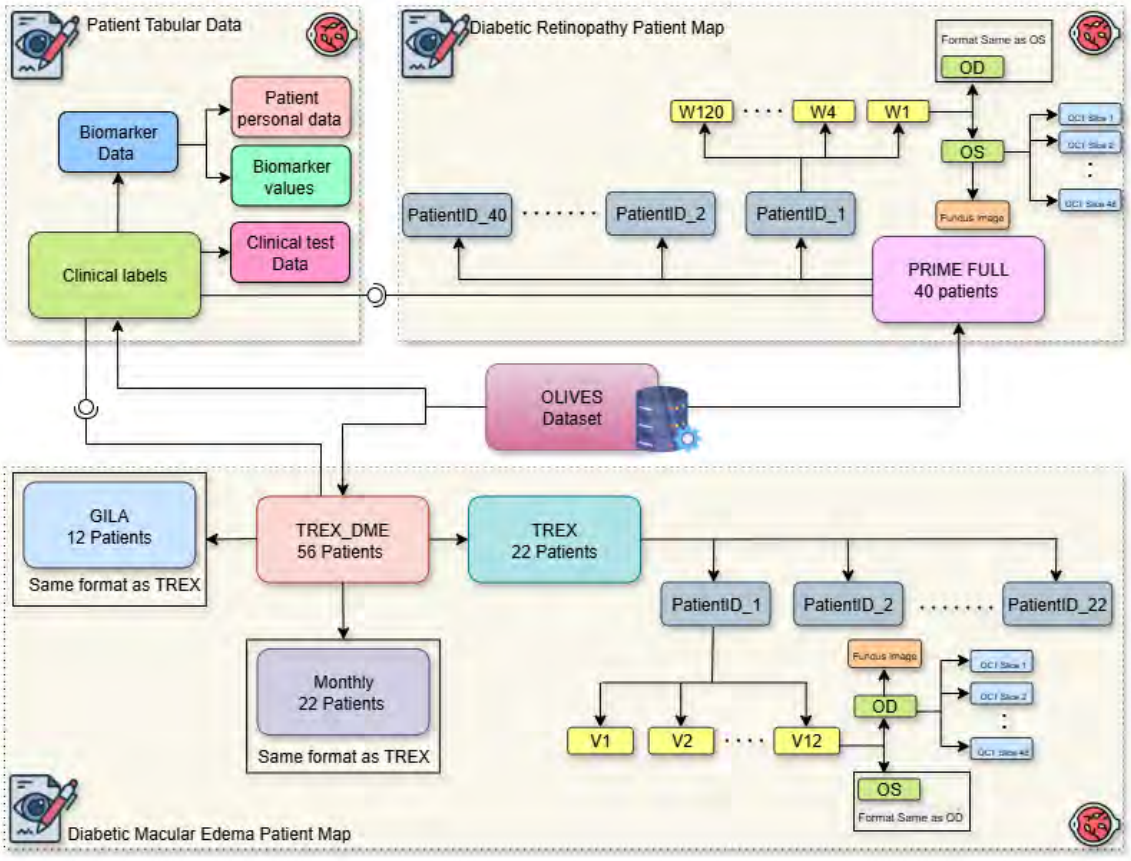

Figure 4.3: OLIVES dataset structure and mapping

TREX-DME was a study of individuals with DME and involved 56 patients. These patients were divided into three smaller groups: TREX, GILA and Monthly. Their check-ups were coded as V1 to V25, and they came in almost every month. This provided extremely precise information over time, which assisted researchers in examining both short-term and long-term changes in the eye.

4.3 Single Patient Image Availability

In our dataset, each patient is categorized into two major folders: the right eye (OD) and the left eye (OS). In each of the folders, there are two modalities of retinal imaging. The former is a fundus image, providing a global overview of the retina that encompasses surface-based anatomies. The second modality is OCT B-scans, the cross-sectional images performed across the retina in sequence. In each volume of OCT, there are 49 slices, ranging from 0 to 48, that span the two edges of

the macula. Thus, one OD/OS folder contains a total of 50 images, i.e., one fundus picture and 49 OCT slices.

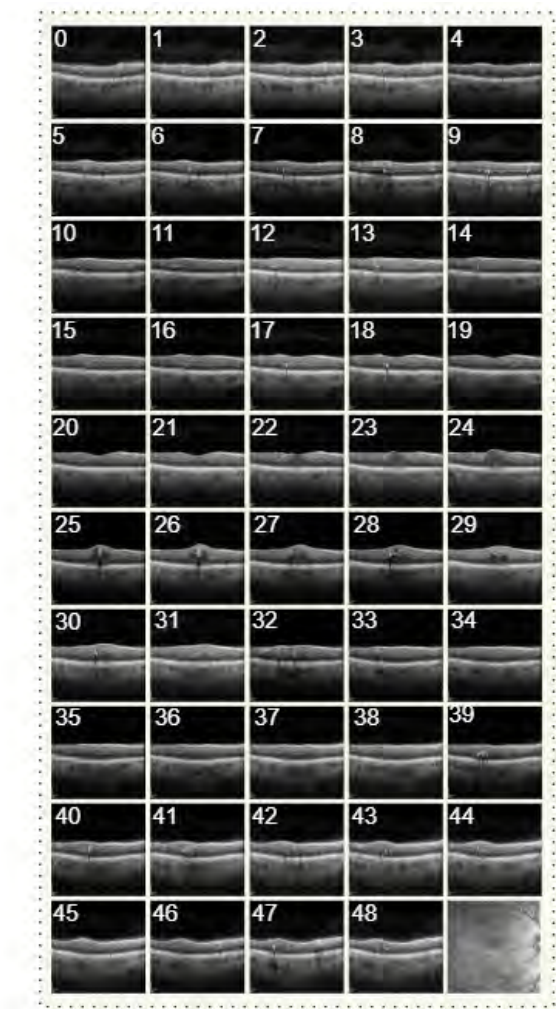

Figure 4.4: Full OCT volume with 1 fundus

4.3.1 Formatting 1

For our Dual Vision Architectures, we propose an approach that utilizes only a subset of the 49 OCT slices for each instance. Instead, we format our OCT data in such a way that it will generate approximately the same quality of fused data. Meaning the fused OCT image will extract features with a decent result. This approach will overcome our computational limitations. To achieve this goal, we built our data structure in 8 different ways. First, we worked with only 3 OCT slices out of 49. Then we used 5,7,9,11,13,15 and 24 OCT slices. For example, for the 5 OCT structure, we used 5th, 15th, 25th, 35th and 45th OCT images. The overall structure is the same except for this minor change. After formatting our dataset in this approach, we observed the changes in our results. Surprisingly, we encountered a minor, gradual improvement in our model's performance and from this, we proposed a hypothesis that if we did not encounter processing limitations, our model would perform well.

4.3.2 Formatting 2

For our OCTbiO architecture, we need to consider only each patient's first week and last week either OD folder data or OS Folder data. So for a single patient we get combined first and last week's total 98 OCT slices and each 98 OCT slices has a tabular 98 sets of Biomarker values shown in the Biomarker Feature table. Each set has 15 unique Biomarker Presence values. But we are not using the fundus image in any OS or OD 50 image folder sets because this formatting is only based on the OCT slices and their biomarker values.

4.4 Data Pre-processing

Before training the model, we clean up the images to make them as stable as possible. OCT and raw fundus scans often exhibit significant noise, uneven lighting, and distracting backgrounds. For this reason, we execute a sequence of pre-processing steps. These are used to slice in around the central part of the eye, improve contrast, reduce noise, and adjust brightness. We also put all images to the same size so the model can only see the useful parts and learn more effectively.

4.4.1 Fundus Pre-processing

Once the fundamental pre-processing stages were set up, we focused on the fundus images and applied specific techniques to remove unwanted background and highlight the most important retinal features, enabling the model to learn more efficiently.

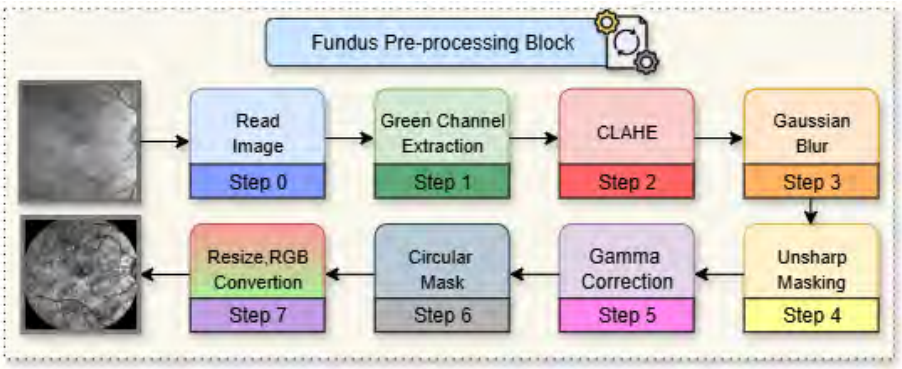

Figure 4.5: Step-by-step pre-processing pipeline for fundus images

- **Green Channel Extraction**

  The green channel in fundus images usually shows better contrast for the retinal tissue absorbs green light, which helps make the blood vessels clearer [7], [81]. On the other hand, the RGB channel often has more noise and is usually weaker, so the green channel is preferred in ophthalmology centers. In our process, we focus the green channel to filter out extra information and make vessels more distinct than in RGB image.

- **Contrast Limited Adaptive Histogram Equalization (CLAHE)**

  CLAHE helps to fix low contrast and uneven lighting in retinal scans by equalizing small regions and preventing overexposure [1], [25]. It can highlight exudates, vessel boundaries, and microaneurysms, making it suitable for DR screening and vessel segmentation. We applied an 8x8 grid and a clip limit of 2.0, which improved vessel patterns and lesion visibility while reducing noise for subsequent steps.

- **Gaussian Blur and Unsharp Masking**

  Gaussian blur reduces noise in retinal scans, and unsharp masking enhances vessel edges by using a blurred copy of the image [24], [79]. This process smooths the background while making vessels and lesion boundaries more identifiable without adding artificial effects. To improve lesion margins and vessel visibility, we applied a 55x55 Gaussian blur with weighted addition.

- **Gamma Correction**

  Gamma correction helps adjust brightness in retinal scans by using a nonlinear mapping to improve mid-tone contrast and reveal small details [10], [28]. It is widely used in medical imaging to balance illumination and make it easier to detect tiny vessels or lesions. In our work, we applied 1.2 to preserve sharp edges and enhance mid-tones, highlighting vascular structures.

- **Circular Masking and Cropping**

  It removes the dark background in fundus photos by keeping only the retinal field of view [2][3], [8]. It prevents the model from non-retinal parts or learning noise and improves classification. We applied a centered mask and cropped tightly to include the retina region.

- **Fundus Resizing and RGB Conversion**

  For all fundus images, resize them to the same dimensions so that the model, which receives a fixed input, also does not misclassify [4], [9]. It helps to make a balance between detail preservation with computational cost and is standard in retinal image processing pipelines. For our work, we resized all images to 224x224 px and converted them to 3-channel RGB by replicating the greyscale channel.

4.4.2 OCT Pre-processing

After finishing the pre-processing steps for the fundus images, we moved on to the OCT scans. These scans show the detailed structure of the different layers in the retina, but they often come with some challenges such as the scans are typically noisy, having low contrast, and varying in resolution. To overcome these problems,

we implemented a custom cream artillery pipeline specifically adapted for OCT slices.

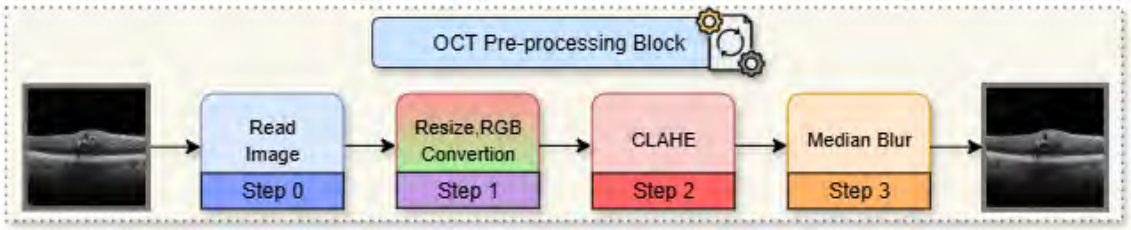

Figure 4.6: Step-by-step pre-processing pipeline for OCT images

OCT Resizing

If we don't normalize OCT scans, it can make training unstable because they come in different sizes and shapes [6], [17]. Resizing needs to preserve fine anatomy visibility while maintaining consistent inputs for CNNs. For our work, we resized each OCT slice to 224x224 px so small features remain detectable with manageable computation.

Contrast Limited Adaptive Histogram Equalization (CLAHE)

OCT scans often have uneven lighting and low contrast, which makes it harder to see retinal layers and detect disease signs [6]. Applying this improves local contrast by enhancing small tiles while limiting over-bright areas and noise. We applied CLAHE to make retinal boundaries, structures, and lesions clearer for automated detection and segmentation.

Median Blur Denoising

OCT images often show speckle noise from the light source, which makes it difficult to see thin retinal layers and small problems [5]. To reduce this, we used a median blur, where each pixel is replaced with the middle value of its neighbors, which smooths the noise while keeping the edges sharp. We applied a 3x3 median filter in OpenCV, which removed much of the speckle while keeping the retinal borders clear.

4.5 Time Series Data

A time series is a series of data observed or measured at a series of time intervals, usually spaced at fixed intervals, to study patterns and relationships. In time series data, the data is collected sequentially at different time steps with some temporal dependencies. [16], [38], [55].

Now, the formatted data we are using in our model can be categorized a time series data. Because data were collected on a weekly basis for each patient visit to monitor the gradual progression of the patient. Therefore, for different time steps, the labels differ. We proposed utilizing this data collection pattern in our preferred manner. We have sufficient data, but the number of unique patients is very low

in this dataset. So, we need to increase the uniqueness to get a good result. The data collection pattern of the OLIVES dataset helped in our proposed approach. We considered one weekly data point as a single data point consisting of 50 images. This way allowed us to treat the weekly images of the same patient as a unique point. Thus, one patient contributed by providing multiple data points. Therefore, we addressed the issue of a smaller number of unique patients by focusing on the weeks as the key to uniqueness.

4.5.1 Methodological Implementation and Output Validation

Extraction and Structuring

This final dataframe contains 3,029 time series data points. A data point represents a special combination of patient eye visits and is captured at a specific week. DR and DME are dynamic and progressive diseases. Therefore, the clinical features and imaging components will naturally differ over time. Time series construction in this work was performed using a carefully designed data extraction pipeline, as confirmed by the results of Jupyter notebooks. The procedure follows:

- Goes through each patient directory (patient ID) followed by each available visit per patient and eye (OD, OS) by week number.
- On each visit-eye tuple, aggregates all image and metadata data and admits only ones that have full records (one fundus, all 40 slices of an OCT scan).
- Stores results as rows of a consolidated DataFrame: each row is a visit-eye, each column a tabulated feature (e.g., OCT1path, label), in strictly-temporal order per patient.

4.5.2 Output-Driven Verification

- The DR patient's statement attests to the resulting number of 1,281 complete time series values, as many as are independent visits to an eye of an eye patient.
- It is also correct, displaying the temporal, multimodal, and labelled organisation of every record and the fact that key attributes (visit order, patient ID) are being stored as such in the downstream sequential model.

4.5.3 Scientific Advantages

By organising the ophthalmic clinical and imaging data into time series, it is possible to:

- **Disease trajectory models:** This is used to predict retinopathy severity or time to treatment need (or negative outcome) using the complete record of the disease.
- **Dynamic risk estimation:** Traditional risk profiles, which have received new visits, are updated.

- **Personalised medicine:** Letting algorithms learn using specific, temporal interactions in people, instead of only using cross-sectional comparisons that are fixed in advance in the cross-section [38], [55].

This methodology supports current models of automated progression modelling in diabetic retinopathy (e.g., the Deep DR system) and has been viewed as the most cutting-edge in predictive medicine and clinical science [55].

Concluding, while multimodal, visit-specific ophthalmic records of patients and the eye are being converted to a time series format, all the disease trajectories of individual patients and eyeballs are captured, which makes it feasible to enable granted time initiative operations along with forecast modelling, which are all based on rigorously verified output of data engineering principles [16], [38], [55].

4.6 Dataset Splitting

The dataset is divided into two primary categories: the Test set and the Train + Validation set. Inside the Train + Validation set, another step is applied, which is stratified K-fold cross-validation. This technique divides the Train + Validation set into five equal groups: Split1, Split2, Split3, Split4, and Split5.

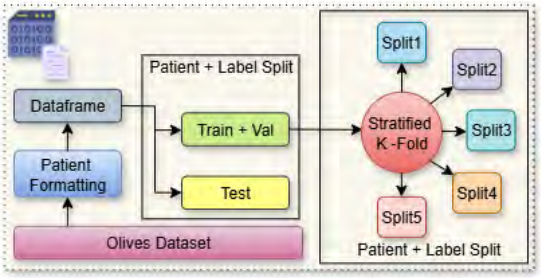

Figure 4.7: Train-Val and Test splitting

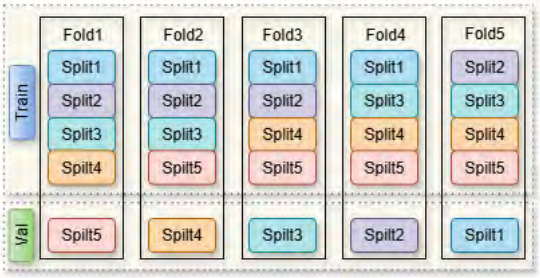

Figure 4.8: K-Fold splitting

One fold is chosen as the validation set, and the other folds are utilized as the training set in each training loop. For every loop, it selects training data (K-1 splits) and validation data (1 split). So, each fold has one opportunity to be used as the validation set during the K times of this process.

Through this rotation, the model is trained and validated multiple times, and the performance is averaged across all folds. After training is complete and the best-performing model is chosen, the completely untouched test set is used to measure the model's true accuracy and generalization ability.

After applying K-fold, our data distribution across the training folds in Dual Vision V1,V2 architecture is displayed in the Table 4.3 below:

Fold	Number of data
fold1_train	2108
fold2_train	2160
fold3_train	2092
fold4_train	2164
fold5_train	2064

Table 4.3: Number of data in Dual Vision training folds

Data distribution across the validation folds in Dual Vision V1,V2 Architecture is displayed in the Table 4.4 below:

Fold	Number of data
fold1_val	539
fold2_val	487
fold3_val	555
fold4_val	483
fold5_val	583

Table 4.4: Number of data in Dual Vision validation folds

After applying K-fold, our data distribution across the training folds in OCTbiO Architecture is displayed in the Table 4.5 below:

Fold	Number of data
fold1_train	6396
fold2_train	6397
fold3_train	6397
fold4_train	6397
fold5_train	6397

Table 4.5: Number of data in OCTbiO training folds

Data distribution across the validation folds in OCTbiO Architecture is displayed in the Table 4.6 below:

Fold	**Number of data**
fold1_val	1600
fold2_val	1599
fold3_val	1600
fold4_val	1600
fold5_val	1600

Table 4.6: Number of data in OCTbiO validation folds

After splitting our data by the K-fold method, we checked if any patient data overlapped for each fold. We verified that there are no overlapping patients. We also verified if there are any duplicate rows across the different folds. Again, we got a positive outcome. For each fold, we save the performance metrics (such as accuracy or loss). Ultimately, we calculate the average score across all folds. This is a good choice because the model is tested more fairly; it reduces the chance of getting a biased result from a single lucky split, and every data point is used for validation only once.

Chapter 5

Methodology

In our initial phase of the research, we reviewed various surveys and modality-wise papers related to ocular diseases, identifying the types that are particularly crucial for eye healthcare. After researching, we found that DR and DME are much more vision-threatening and crucial due to their overlapping patterns, and we began to look for datasets related to DR and DME. Most previous research focuses on a single modality or multimodal approaches, but not specifically on DR with DME.

For classifying DR and DME, we utilize the OLIVES dataset, which contains only patients with DR and DME. Since we plan to work with multimodality, OLIVES data is the only option because we need to trust more than one modality for an accurate diagnosis. This research aims to distinguish these two ocular conditions by using a multi-modal fusion-based approach. In our ablation study, we employ a dual-branching design that combines each patient's fundus data and OCT data using concatenation fusion. For this Dual Branching, we have pretrained CNN models for each modality, such as ResNet18/50, EfficientNet-B0/10, and Inception-V3. However, there are some limitations to using pretrained models with concatenation fusion, and to overcome these, we aim to use custom CNN encoders for each modality with various kinds of communication-based fusion mechanisms.

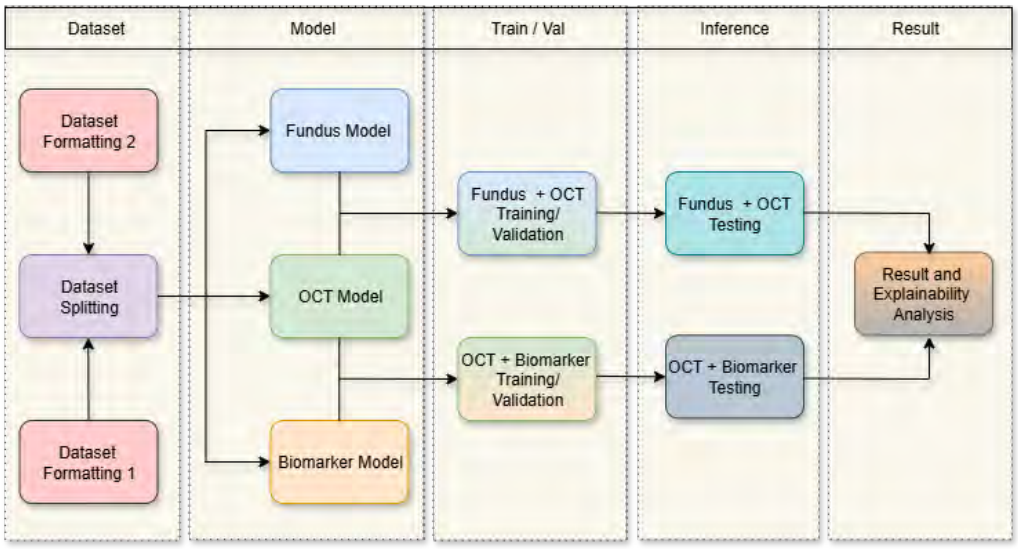

Figure 5.1: Methodology of our research

Following our ablation study and key insights of Dual Branching results, we now aim to create two distinct architectures, each focusing on a different format of the dataset. Dual Vision architectures focus on utilizing the dataset format 1, which includes fundus and OCT slices of the same patient. On the other hand, OCTbiO architectures focus on utilizing the dataset format 2, which includes OCT slices and their respective biomarkers of the same patient. In the ablation study design, we have the same pre-processing method for both modalities, but after analyzing the key differences of each modality, we have different pre-processing approaches for each image modality (fundus and OCT) after several trials of pre-processing techniques.

In Figure 5.1 for the fundus modality, we have the fundus model, which is a CNN encoder, and for the OCT modality, we have an OCT model, which is also a CNN encoder. On the other hand, we have a biomarker encoder for each OCT's respective biomarker. We aim to train, validate and test the fundus encoder and the OCT encoder together for dual-vision architectures with different fusion mechanisms such as gated fusion and MSA fusion. On the contrary, we also aim to train, validate and test the OCT encoder with the biomarker encoder with different fusion mechanisms such as gated fusion, MSA fusion and FiLM fusion. After the end-to-end classification, we plan to perform a result verification explainability check and visualization for each of our architectures to see how much each modality performs by itself and after fusion.This visualization can help us find key insights into how each modality interacts and how each modality contributes to the final classification decision. We will compare our each model and architecture metrics, which include parameters, FLOPS and inference time. In addition, we will compare our qualitative and quantitative results with dataset benchmarks.

5.1 Dual Branching Design (Ablation Study)

For multimodal approaches, dual-branched designs perform well because they introduce richer understanding and flexibility [60], [66], [78]. In our Dual Branching Design, we have a fundus branch and an OCT branch. Later, we have a fused branch to fuse the output of the dual branches. Each branch has a separate workflow depending on the image modality. In our dataset (OLIVES), we have approximately 40 selected patients among all instances. Each patient has fundus image data and OCT image data from multiple weeks. Below, we will be focusing on all branches of our preliminary approach:

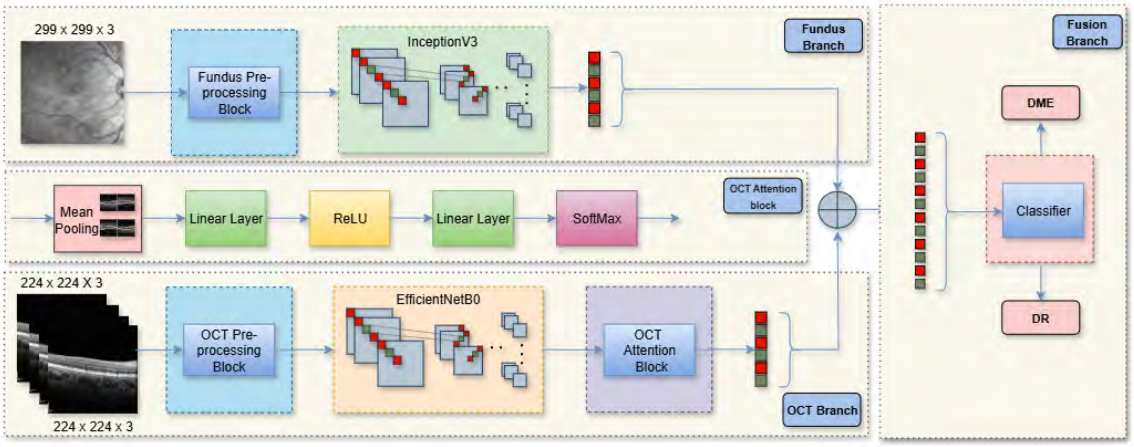

Figure 5.2: Dual Branching diagram

- **Fundus Branch:** For each patient, we have a grayscale fundus image. Each fundus image indicates a certain label. So, if a patient has gone through 10 weeks of treatment, we have 10 images of his/her left eye fundus image and 10 images of his/her right eye fundus image. All images of that particular patient indicate the same label. So, the fundus image is pre-processed by the fundus pre-processing block and then fed into the Inception-V3 architecture. Then it creates a feature vector, which is the output of the fundus branch.

- **OCT Branch:** For each patient, we have a grayscale OCT image. Each OCT image corresponds to a specific label. So if a patient has gone through 10 weeks of treatment, then we have 40 images of his/her left eye OCT image and 40 images of his/her right eye OCT image, considering 4 OCTs per left or right eye. Here, the OCTs are being stacked on each other. That being said, all images of that particular patient indicate the same label. So, the OCT image is pre-processed by the OCT pre-processing block and then fed into the EfficientNetB0 architecture. Then it creates a feature map which enters the OCT Attention Block. Then it creates a feature vector which is the output of the OCT branch.

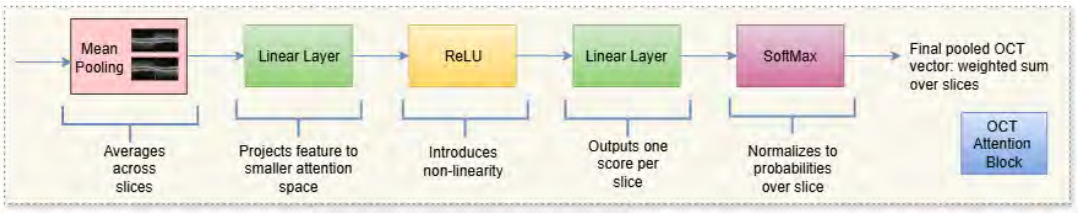

Figure 5.3: OCT attention block diagram

The OCT Attention Block is for highlighting the most important and informative OCT features between the 4 slices of OCT images. It helps to focus on the most relevant features and picks them for later classification.

- **Fused Branch:** The output of the Fundus branch and OCT branch which is their respective feature vector is later fused in the Fused Branch and passed to the classifier for the classification of DR severity level. Here, for fusion, we are using concatenation-based fusion. In other words, we are applying late fusion.

But this architecture has overfitting tendencies as from the very first epochs it starts with high training and validation accuracy. Also, the validation loss was less than the training loss from the very start at each progressing epoch, which is also a major issue. There could be several reasons behind this. First, we are using the same pre-processing for both OCT and fundus images. Then the OCT Attention Block might cause this issue or concatenation-based fusion can result in memorization rather than picking on the actual pattern. So, based on this, we have a totally different approach for our final proposed architectures.

In our new proposed fusion-based architectures, we have different fusion modifications. These modifications are based on our previous Dual Branching proposed architecture (ablation study) results and methods. Our main objective was to design the architecture based on the data modalities we have and what type of data each image modalities offer. Most ocular disease based architectures are built in a way that only find patterns over the ocular surface, but for an architecture to generalize over an ocular disease, we need surface-level information alongside layer-level information. On the other hand, OCT slices contain valuable biomarker information. Since we have each OCT slice biomarker tabular values that indicate if that specific biomarker is present or not in that specific OCT slice. Keeping in mind these methods and modalities, we have proposed two multimodal fusion based architecture that introduces variability between two types of modalities that combine both layered, surfaced information of an ocular image as well as their biomarker values.

5.2 Dual Vision V1 Architecture

Our Dual Vision V1 architecture has five parts. Each part has a certain objective and in each part we are introducing something new from the previous design. The parts are each patient image modality part, pre-processing part, CNN models (fundus model, OCT model) part, fusion mechanism part and lastly a classification part. Despite the number of OCT images, our architecture remains pretty much the same. We are experimenting with a variable number of images per patient and got different types of results which will be discussed later. Below we have our main proposed Dual Vision V1 architecture:

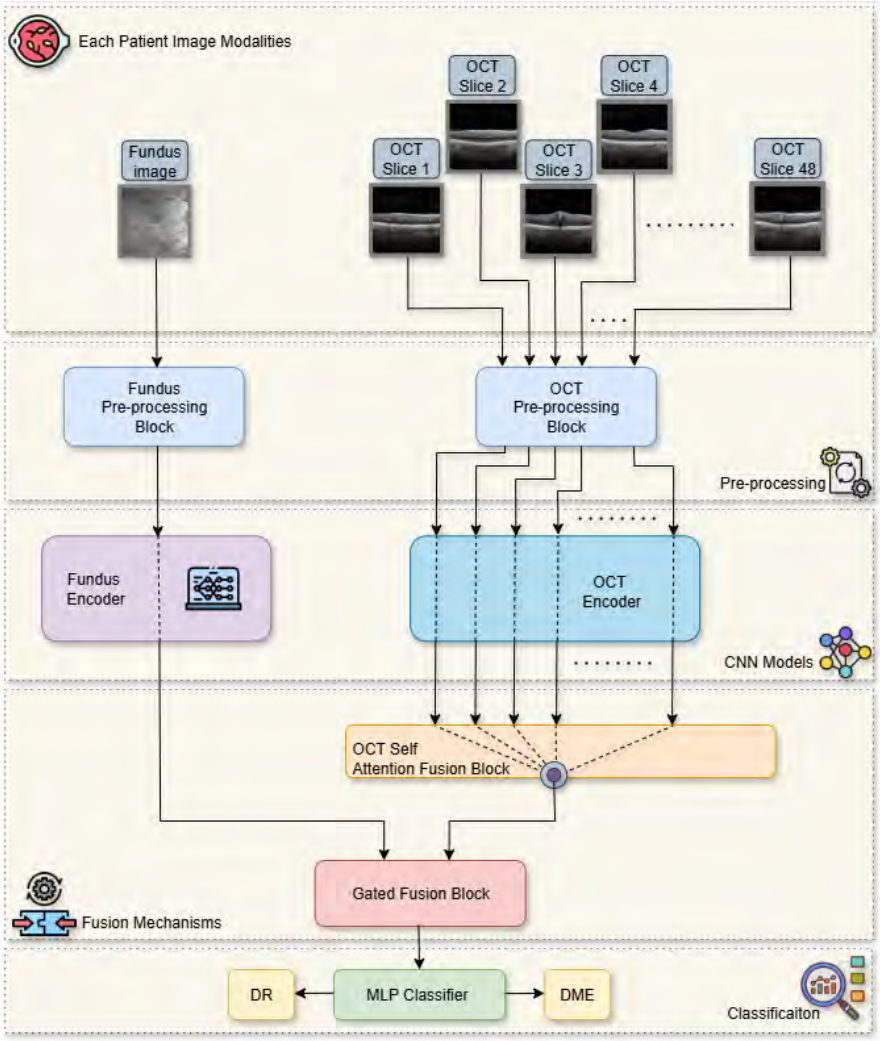

Figure 5.4: Dual Vision V1 architecture diagram

In our Dual Vision architecture, we have one fundus image per patient and multiple OCT images per patient. Here we are trying at least 3 OCT images per patient with a certain interval or gap. As we already mentioned, we have a maximum of 49 OCT images. An OCT scan of a certain eye is not a single image. All 49 slices together are considered a single OCT image. In other words, all 49 slices make a 3D OCT volume and these slices are 2D slices. Now, as already mentioned, if a patient had 49 slices, we are taking a minimum of 3 slices in 15 intervals. In other words, we took slice 15, slice 30 and slice 45. If it is 4 slices per patient, then slice 0, slice 15, slice 30 and slice 45. So, for a single patient, we have a minimum of 4 images (1 Fundus image, 3 OCT slices) and a maximum of 50 images (1 fundus image, 49 OCT slices). So, in each patient modalities part, the data format should be like this to work with our fusion-based architecture. Now, in the pre-processing part, we have a separate pre-processing block for the fundus image and OCT slices. Previously, we had the same pre-processing strategies for both image modalities. That worked well with OCT images, but the fundus image became much noisier for the model to catch the pattern. Now we have a fundus pre-processing block for the fundus image and OCT pre-processing block for the OCT slice. Each block has different steps for each modality to identify their pattern well. We have already discussed how each pre-processing block works and its outputs in the dataset pre-processing section.

5.2.1 Encoders and Attention Blocks

Now we have two custom CNN models that are totally separate. One is for fundus images and one is for OCT images. The reason we had two different CNN models is the imaging capabilities they offer. Fundus images mostly offer a retinal surface-level overview, whereas OCT images offer layered retinal thickness. So, there is mostly disjoint but correlated information between the fundus image and the OCT slice. We later dive into the medical level model interpretation based on the image modality. So, for fundus images, we have a fundus encoder and for OCT slices, we have an OCT encoder. For a single sample, a fundus image goes into the fundus encoder and each multiple OCT slice goes into the OCT encoder.

Fundus Encoder

Our Fundus encoder requires an image size to be 224 x 224 (height x width). So, at the pre-processing last step, we resize our fundus images into this required size for the model. Here (B,3,224,224) means (Batch size, Input Channels, Height, Width). Even though the input images were grayscale, which is a single channel, we converted them into 3 different channels because most of the CNNs require all three channels to be present. So, finally, after resizing and RGB channel conversion, all the images enter our fundus encoder. Below the fundus encoder diagram is given:

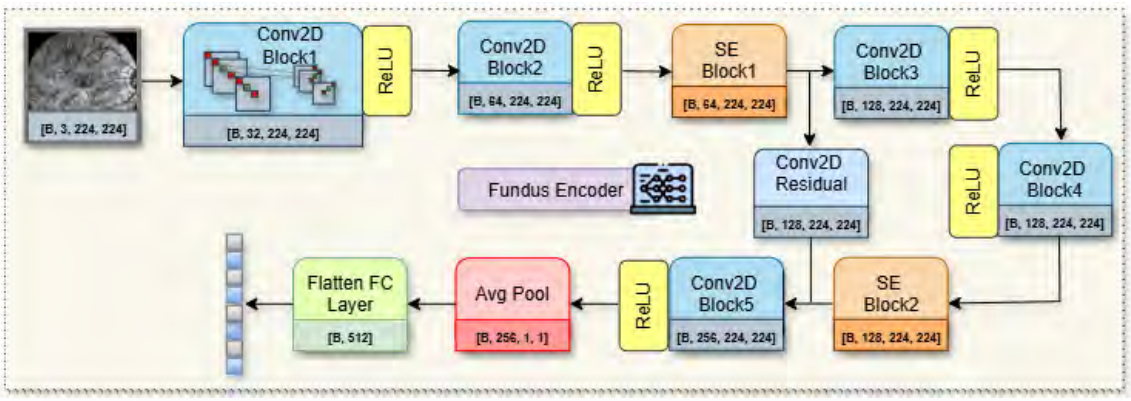

Figure 5.5: Fundus encoder diagram

In our fundus encoder, we have multiple sequential convolutional blocks like Conv2D (Input channel, output channel, kernel, padding) blocks with attention blocks. The input shape is (B,3,224,224). Firstly we have Conv2D Block1 (3, 32, 3x3, 1) followed by a ReLU layer. The output shape of this is (B,32,224,224). Then a Conv2D Block2 (32, 64, 3x3, 1) followed by a ReLU layer. The output shape of this is (B,64,224,224).

Then, we have a SE Block1 (Squeeze and Excitation Block 1). Here, the output of Conv2D Block2 enters SE Block1. Here, the input and output shape of the SE BLock1 is the same because the Squeeze and Excitation attention method does not change the spatial resolution. So, the output shape of this SE Block1 is

(B,64,224,224). Then the copy of output is saved as SEBlock1Output and the real output enters Conv2D Block3 (64, 128, 3x3, 1) followed by a ReLU layer. The output shape of this block is (B,128, 224, 224). Then the output of Conv2D Block 3 enters Conv2D Block 4 (128, 128, 3x3, 1), followed by a ReLU layer. The output shape of this consecutive block is (B,128, 224, 224).

Then, after this, the output enters a SE Block2. This block acts the same as the SE Block1. Now the output of SE Block2 is residually connected with the copied output from before, which is SEBlock1Output. Here, the SEBlock1Output dimension is (B, 64, 224, 224), but the SEBlock2Output dimension is (B, 128, 224, 224). But for residual adding and matching the channels, we need to match the dimension size. So, we have a Conv2D Residual Block. The SEBlock1Output enters this Conv2D Residual Block, and the output size becomes (B, 128, 224, 224).

Now SEBlock1Output and SEBlock2Output are residually connected and passed to Conv2D Block5 (B, 256, 224, 224). The output of Conv2D Block5 is (B, 256, 224, 224). These are all our convolutional blocks of fundus Encoder. Now after the Conv2D Block5 the output of it enters a Adaptive Average pooling layer which outputs a (B, 256, 1, 1) dimension. Then it enters a Flatten layer which outputs a 1D feature vector per patient. The output shape is (B,256). Then it enters an FC (Fully Connected) layer, which increases the dimensionality to (B,512). In other words, it expands the dimensionality for global embeddings. This is the output of our Fundus Encoder. There are total of 20 layers in our fundus encoder.

Squeeze and Excitation Block

Squeeze and Excitation Attention methodology is used for enhancing features and giving attention to them channel-wise [20]. Squeeze and Excitation explicitly demonstrates the relationships between channels to adaptively tune feature values at the channel level helps finding patterns for DR or DME detection. There are 3 parts of a SE block and each part has particular motives. Below, a SE Block diagram is given:

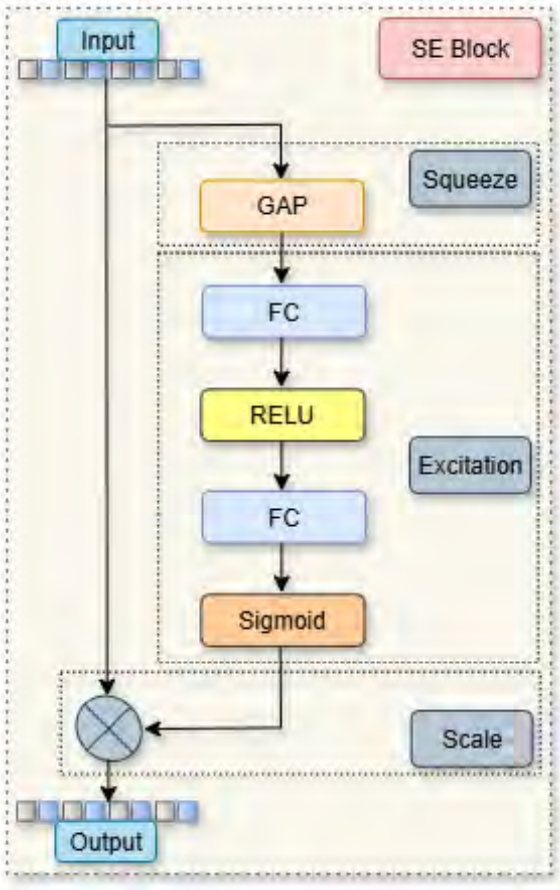

Figure 5.6: SE block diagram

The first part is the Squeeze step. In this part there is only one layer which is GAP(Global Average Pooling). In order to reduce spatial dimensions, we averaged them. In other words, it averages over all the spatial positions and compresses them. So, we obtain each channel's global description. The output of the GAP layer is (B x 1 x 1 x C). Then comes the second part which is the Excitation. Excitation has four steps which are FC Layer 1,ReLU layer, FC Layer 2, Sigmoid Layer. FC Layer 1 is the reduction layer which reduces the channel dimension. Before it was C then after it became C/r where the r is the reduction ratio. So, the output of the FC Layer 1 is (B, C/r, 1, 1). Then comes the ReLU layer that helps the model to learn complex dependencies by introducing non-linearity to the model. After ReLU we have FC Layer 2. The objective of FC layer 2 is to expand the dimension back to C. So the output of the FC layer is (B, C, 1, 1) again. Then, lastly the output of FC Layer 2 goes into a Sigmoid Layer which compresses the values between [0,1]. Now we have per channel weights. Since we have the weights our excitation part is done and now comes the last part which is to Scale. In the Scale part, we re-calibrate the channel features. We have the channel features at the first and after we get the excitation weights we multiply the weights with the channel features. Here, after getting multiples, we only get the features of each channel that are important and suppress the less important ones. Also, the input and the output of the SE block remains in the same dimensionality which is (B, C, H, W). Since DR or DME lesions can appear anywhere on the Fundus image and in any shape, GAP ensures global detection. Then the FC bottleneck in the SE block ensures feature suppression. If

hard exudates are large and appear more than once, it can ignore the noise of the thin vessels.

OCT Encoder

Our OCT Encoder requires image size to be 224 x 224 (height X width). So at the OCT pre-processing, we resize our OCT images into this required size for the model. Here (B,3,224,224). So, if there is a patient who has (3 + n) number of OCTs where n = 0,1,2...46 we resize every one of them to the same size.

For each OCT slice per patient, we are not creating a copy of the OCT encoder. To clarify, if a patient has 5 OCT slices, we are not creating a single OCT encoder for the exact OCT slice. Rather than all OCT slices going through a single shared Encoder. The reason behind it was, if we make a copy of the OCT encoder for each OCT slice, then the total model parameters will also increase with the number of increased OCT slices per patient. So, a patient who has 5 OCT slices will have 5 OCT encoders and a patient who has 49 OCT slices will have 49 OCT Encoders, which seems to be redundant and heavy. It increases the model parameters by a lot. So, to resolve this problem, we have a shared OCT model, and for each patient, no matter how many OCT slices, each one goes through the single shared OCT encoder. So, as the number of slices increases the model parameters stay the same for each format. There are total of 12 layers in our OCT encoder.

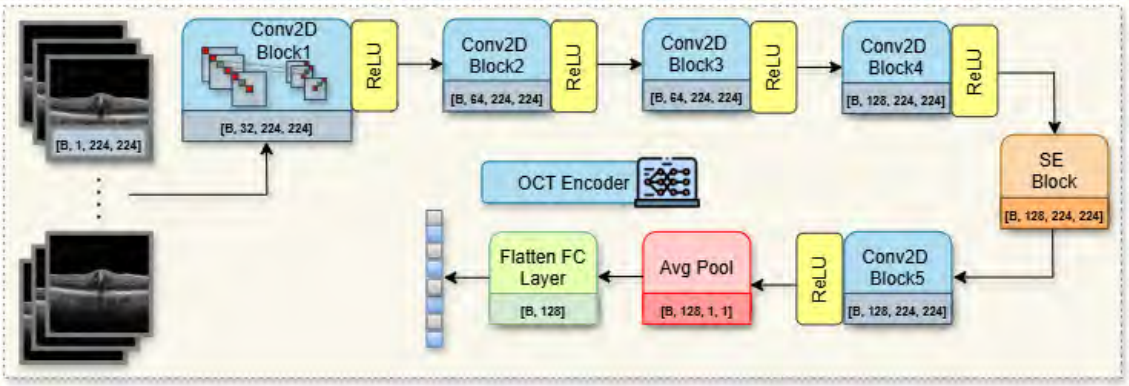

Figure 5.7: OCT encoder diagram

In our OCT encoder, we have multiple sequential convolutional blocks like Conv2D (Input channel, output channel, kernel, padding) blocks with attention blocks. Here like the fundus encoder, we do not have any skip connection/residual connection from one layer to another. We are keeping the OCT encoder a little bit simple. Firstly, we have Conv2D Block1 (3, 32, 3x3, 1) followed by a ReLU layer. The output shape of this is (B, 32, 224, 224). Then a Conv2D Block2 (32, 64, 3x3, 1) followed by a ReLU layer. The output shape of this is (B, 64, 224, 224).

Now after this, we have a Conv2D Block3 (64, 64, 3x3, 1) again followed by a ReLU layer. The input and output dimension sizes are the same for this block. The output shape of this is (B, 64, 224, 224). Here, the output shape does not change between Conv2D Block 2 and 3. Then a Conv2D Block4 (64, 128, 3x3, 1) followed by a ReLU layer. The output shape of this is (B, 128, 224, 224).

After 4 consecutive convolution layers, we have the SE Block, like the fundus encoder. Here, the input and output shape of the SE BLock1 is the same because the squeeze and excitation attention method does not change the spatial resolution. So, the output shape of this SE Block1 is (B, 128, 224, 224). This SE block acts the same which is to highlight the strong and important features and suppress the less relevant ones or noises. SE block is pretty much necessary here because OCT slices are very noisy. So in order to remove or decrease noise after pre-processing we need this SE Block.

Then a Conv2D Block5 (128, 128, 3x3, 1) followed by a ReLU layer. Here, the output shape does not change between SE Block and Con2D Block 5. Now after the Conv2D Block5 the output enters an Adaptive Average pooling layer, which outputs a (B, 128, 1, 1) dimension. Then it enters a flatten layer, which outputs a 1D feature vector per patient. The output shape is (B,128). Here, for the n number of OCT slices of that certain sample, we get n number of 1D feature vectors that layer goes into a fusion mechanism.

5.2.2 Fusion Mechanisms

Now, In our fusion mechanism Part we have two different types of fusion. In computer vision methodologies, fusion mechanisms are used to connect different modalities (Image, text, sound etc). But here we are not using sound based data in this architecture. We only have relevant image data and tabular data. Since we have two different types of images per patient, which are fundus Image and OCT slice, we have done a different type of fusion between these two in order to connect their feature map for better understanding. We have tabular values of each patient but we can connect them with OCT slices only for image-tabular multimodal fusion. That's why our Dual Vision V1 architecture is based on image-image multimodal fusion.

Multihead Self-Attention Fusion

Since we have multiple OCT slices per patient, we are doing this fusion between the OCT Slices. Assuming that a patient has n OCT Slices, each of them goes into the OCT encoder and outputs n different features maps. So, we are performing MSA Fusion between these n features maps. This fusion methodology's main architecture was actually described in this study [14] where they integrated the MSA Fusion block in their Transformer architecture.

In our code, we have a stack consisting of n OCT slices. Our target is to fuse them through a self-attention technique. Meaning, from the feature vectors extracted from the OCT slices, we will apply a self-attention mechanism to examine which feature vector should focus on which other OCT slices for making a decision. Later we will apply a gated fusion with the fundus encoded feature vector and the fused OCT we got from the MSA fusion. This is how we can use the knowledge of multimodality to know how much focus or attention we should put. In this fusion method, we have 7 Steps for the fusion to work on n OCT Slices. The steps diagram are given below with explanation of each step:

Step 1 and 2: Q,K,V Projection and Multiple Head Generation

For simplicity we assume each patient has n OCT slices. Then after each slice goes into the OCT encoder we get n different feature vectors (x1,x2,x3,x4 ...xn). Here, we consider each feature vector as a Token meaning each token carries their feature vectors.

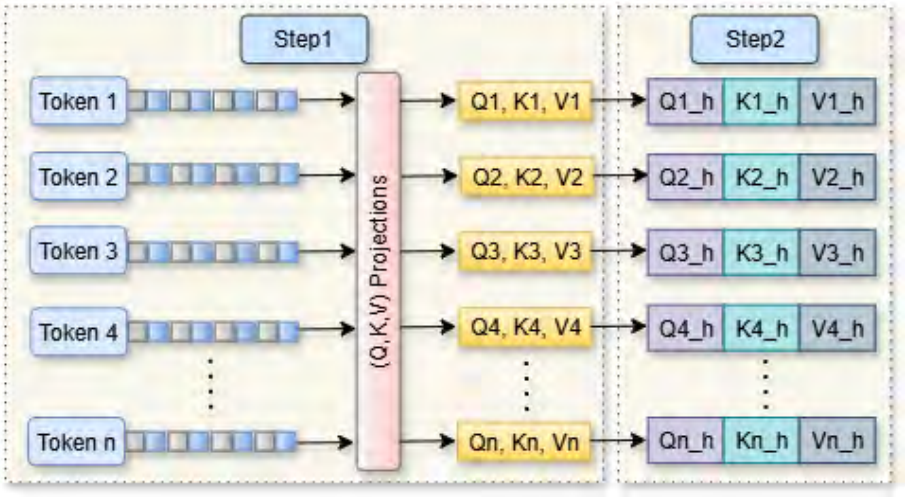

Figure 5.8: Step 1 and 2 diagram

For step 1, X = [x1, x2, x3, x4 ...xn] = [Token 1, Token 2, Token 3, Token 4 ...Token n]. Now each token goes into the (Q, K, V) Projection layer like shown in Figure 5.8. Here Q means Query, K means Key and V means Value. Q means what the token is asking. K means what the token offers what others do not possess and V means what the token actually has. Now we have Wq, Wk and Wv which are randomly initialized by pytorch and are the learnable weight matrices. Also, we have bq, bk and bv which are the bias terms of nn.Linear Layers. The output of (Q, K, V) projection is calculated by the formula below:

The projection layer has three separate linear layers, each with its own weight matrix + bias:

- **Query projection:**
$$Q_1 = x_1 W_Q + b_Q$$
- **Key projection**
$$K_1 = x_1 W_K + b_K$$
- **Value projection**
$$V_1 = x_1 W_V + b_V$$

So, for x1, we calculate (Q1, K1, V1). To simplify, Q1 is what the exact token1 (x1) is asking for, K1 is what token1 (x1) can offer as an identifier and V1 is what the token1 (x1) actually has.

For **step 2**, we declare num_of_heads = 4 since we are doing MSA Each head might focus on different sets of perspectives. Since we have n tokens for head 1, we will

have n different Q,K,V projections of head 1. To clarify, we will have (Q1_h1, K1_h1, V1_h1) for token 1 (Q1, K1, V1) to (Qn_h1, Kn_h1, Vn_h1) for token 1 (Qn, Kn, Vn). Then for head 2 we will have n different Q, K, V projections of head 2. Like, we will have (Q1_h2, K1_h2, V1_h2) for token 1 (Q1, K1, V1) to (Qn_h1, Kn_h1, Vn_h1) for token 1 (Qn, Kn, Vn). So for each 4 heads we will have the same calculation. Now head wise Q,K,V projections are ready, then we will proceed to the next step.

Step 3: Generating Score Values

As we already know, we have 4 heads. For each head, we have to calculate scores separately. So for simplicity, let's assume we are working with head1. For head1 we have QKV_head1 matrix: [(Q1_h1, K1_h1, V1_h1), (Q2_h1, K2_h1, V2_h1), (Q3_h1, K3_h1, V3_h1), (Q4_h1, K4_h1, V4_h1),.....,(Qn_h1, Kn_h1, Vn_h1)]. Now, from using the values of Q and K, we generate scores.

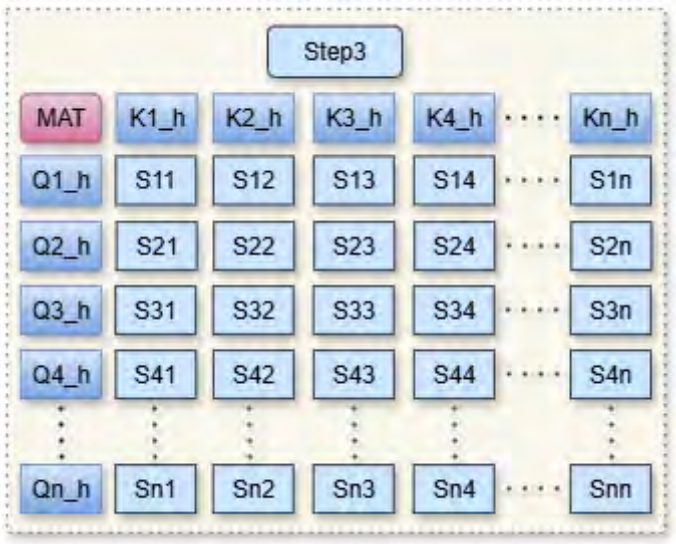

Figure 5.9: Step 3 diagram

For generating score values, we need to perform a dot product between a Query and a Key. Here, from QKV_head1 matrix we have(Q1_h1, Q2_h1, Q3_h1, Q4_h1 ... Qn_h1) and (K1_h1, K2_h1, K3_h1, K4_h1 .... Kn_h1). We need to perform dot products between each one of these pairs as shown in the Figure 5.9. The score dot product formula is given below:

$$d_h = \frac{D}{H}.$$

For each head, compute the similarity between queries and keys:

$$\text{score}[i, j] = \frac{Q_i \cdot K_j}{\sqrt{d_h}}.$$

- **Input:** $Q, K \in (B, H, n\_\text{tokens}, d\_h)$
- **Output:** Scores $\in (B, H, n\_\text{tokens}, n\_\text{tokens})$

Here, there is a score generated for (Q1_h1 and K1_h1) and also for (Q1_h1 and K1_h1). We are doing it because we want to know how much each token or feature should attend to another token. Also, these score values are a communication between each token. Each Token's query value is attending another n number of tokens key value. So, if we have a 4 token system, we will generate 16 score values (S11, S12, S13, S14, S21,... S43, S44). Since, we have n number of tokens where n can range between [3,48]. We will have a minimum of 9 score values and a maximum 2304 (S11..Snn) score values. All these score values are the primary interaction between each token.

Step 4: Score to Softmax Attention Weights

Since we have n number of scores we have to turn them into softmax attention weights. Here, before going to softmax, we have to arrange each score to their particular query number. To determine which score belongs to which query number (Q1, Q2, Q3, Q4, ...,Qn), we need to get back their score calculation query contribution.

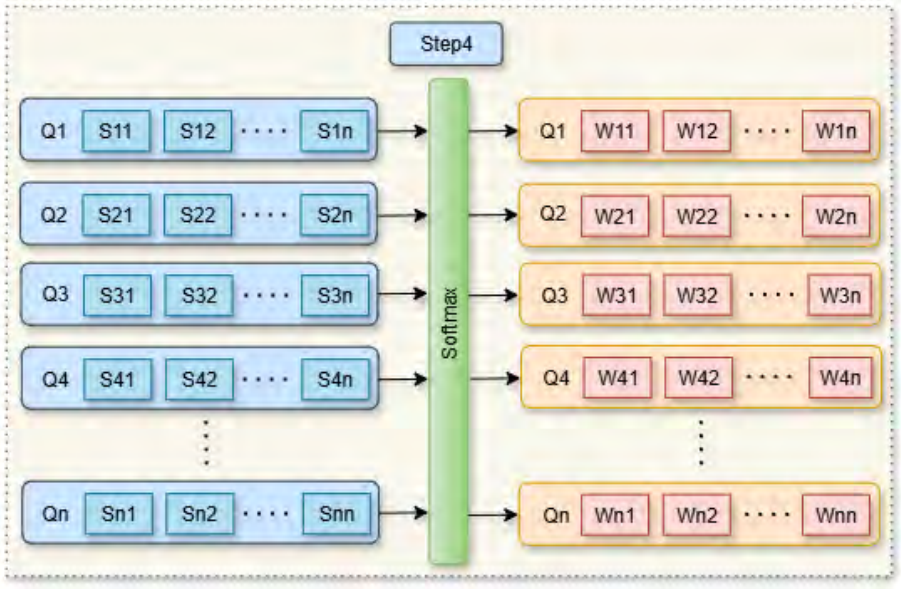

Figure 5.10: Step 4 diagram

So, if a score is calculated by Qn value shown in figure 5.9, then that Qn is the query contributor of that score. To elaborate, for S11 to be calculated, we performed a dot product between Q1_h1 and K1_h1. So Q1_h1 has a contribution meaning for token 1, Q1_h1 paying attention to K1_h1. Same for Snn, meaning Qn_hn has a contribution. Qn_h1 paying attention to Kn_h1. This way we arrange each score by its particular Query number.

$$W_{i,j} = \frac{\exp\left(S_{i,j}\right)}{\sum_{k=1}^{n} \exp\left(S_{i,k}\right)}$$

Now, we insert row wise score values into the softmax equation like shown in the softmax equation. This softmax equation turns them into probabilities . The shape

of this softmax probability matrix is the same as the score matrix of the Q1. Now we have (W11, W12, W13, W14 .. W1n) weights for Q1 and for Q2 we will have (W21, W22, W23, W24 .. W2n) and for Qn will have (Wn1, Wn2, Wn3, Wn4 .. Wnn) like shown in the figure 7. Here, between (W11,W12,W13,W14. . . W1n) if W14 has the highest score that means after softmax equation Q1 will put much higher weight to the key associated with that score which is K4_h1.

Step5: Each Head Output Generation

Now for each head we have (Q1, Q2, Q3, Q4 .. Qn) sets of weights. For example for Q1, (W11, W12, W13, W14 .. W1n) set of weights. Now since we have matrix weights per query, we have to perform a matrix multiplication with the Vales(V1_h1, V2_h1, V3_h1, V4_h1, ...Vn_h1). We are performing row-wise multiplication. Each row wise multiplication will produce a Output value (O1, O2, O3, O4, ...On)

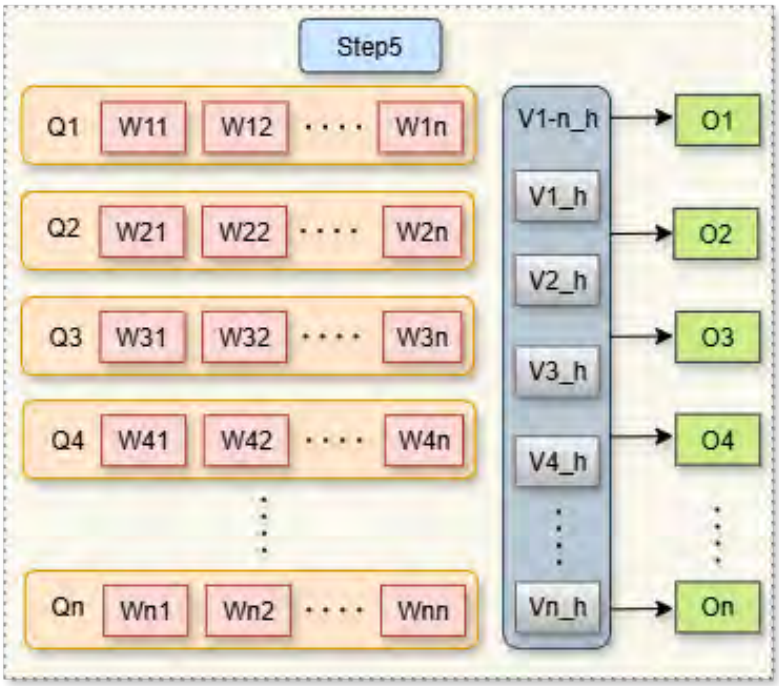

Figure 5.11: Step 5 diagram

$$\text{For each head:} \quad O_i = \sum_{j=1}^{L} W_{ij} \cdot V_j$$

Here, (O1, O2, O3, O4, ... On) is per per-head output. Here, the purpose of this matrix multiplication is to blend the information of the value with the weights. The weights here are how much a token_i is asking of token_j and values are what the token_j actually has. Their Matrix multiplication gives us spatially relevant embedding of token_i of other tokens.

Step 6 and 7: Head Wise Output Concatenation and Projection

At the start of step 6 we have each head output sequence. For example, (O1,

O2, O3, O4, ...On) is the output of head 1. Same for other heads , we will have the same sequence of outputs which will be (O1, O2, O3, O4,.... On). We know that each head will learn something different from each other. So, if head 1 is local lesion context aware where head 2 and others might be global lesion context aware. Each head needs to share their perspectives.

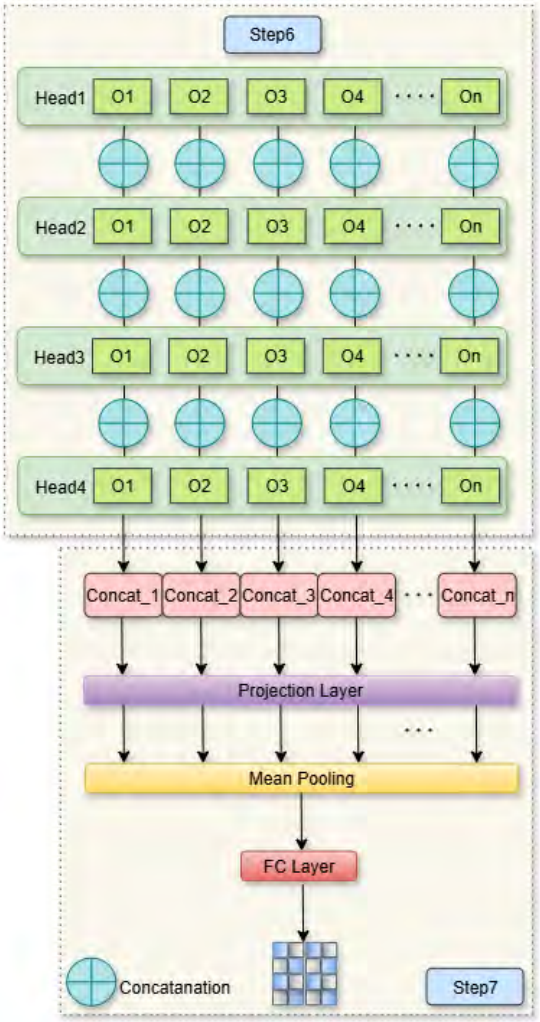

Figure 5.12: Step 6 and 7 diagram

For each token i,

$$O_i = \left[O_i^{(1)} \parallel O_i^{(2)} \parallel O_i^{(3)} \parallel O_i^{(4)} \right]$$

For each head we are concatenating their ith output. So, all O1 from each head will be concatenated like figure 5.12. Then the same for O2, O3, O4 ...On. After all concatenations, we will get n number of concat values (concat_1, concat_2, concat_3, concat_4 ...concat_n). These are the final headwise perspective of the fusion mechanism so far. After this, in step 7 we have to project the concatenation values through a liner projection layer. After the projection layer it goes into a pooling layer followed my a FC layer which outputs a single fused vector consisting of multiple head point of view of the shared token information.

Now, the reason why we choose this fusion strategy rather than other fusion strategy

at this point is that we knew each OCT slice did not have to carry enough information. Let's assume among 3 OCT slices the 2nd OCT slice had 60 percent relevant information in it but other relevant information were in 1st (30 percent) and 3rd (10 percent) OCT slices. So communicating and sharing information between OCT slices, as what they have and what they can offer to each other is a medically correct approach. Also, another reason behind it was OCT images are normally a little bit noisier than other image modalities. So, using only one OCT per patient was riskier and that's why using MSA fusion can help us rely on more than one option [30].

Gated Fusion

After the fundus encoder, we have a fundus feature vector and after the OCT encoder followed by MSA fusion we have a fused OCT fused feature vector. Now since we have two different image modality feature vectors we have to fuse them using gated fusion. A gate (learnt weights ranging from 0 to 1) determines, per feature dimension, how much information to extract from each input modality in the feature-combination method known as the Gated fusion. [82] So, now using this fusion idea we can determine between the fusion modalities, which modality is more trustworthy or they have equal importance. Normally, for classifying DR or DME, one can use any of these images but together they provide us different context and depending on the sample, sometimes we might depend on only one modality or partially on both.

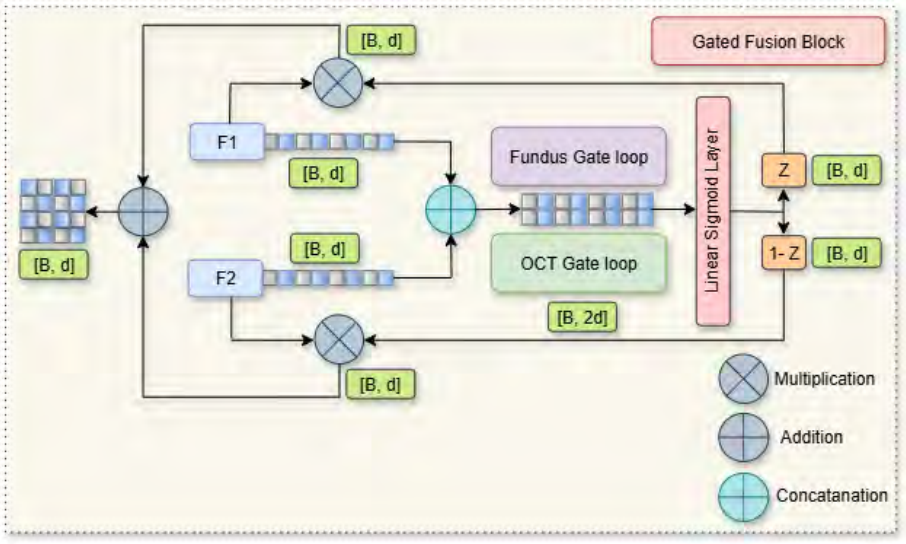

Figure 5.13: Gated fusion diagram

Here, F1 and F2 in Figure 5.13 are two feature vectors and in our case F1 is fundus feature vector and F2 is final OCT feature vector which is MSA fused before. Each F1 and F2 has a shape of [B,d]. Here, B = Batch Size and d = Dimension of feature. Firstly, we need to concatenate them. The concatenation helps the gate to focus on each modality side by side. After the concatenation, the output size becomes [B,2d]. Then this concatenated feature vector is passed into a Linear Sigmoid Layer.

$$\sigma(x) = \frac{1}{1 + e^{-x}}$$

$$z = \sigma(W \cdot fusion + b)$$

Here in this formula of z value, the fusion variable is the concatenated feature vector of Fundus and OCT. Now, (W,B) are the learnable parameters. Sigma (σ) is the softmax variable function. Sigma is dealing with the z value because each gate value needs to be represented as a weight to be between F1 and F2. The z value is a vector of length d consisting of all z values of the whole fusion mechanism. So, the Z value shape is [B,d]. Since we have two modalities of image, for the fundus feature vector, we are going to represent the value z for it and for the OCT feature vector we are going to represent the (1 - z) value. Then the z value is multiplied by the actual fundus feature vector which is from before the Linear Projection layer. Also, the (1- z) value is multiplied by the actual OCT feature vector which is from before the Linear Projection layer. The output shape of both the fundus feature multiplication and OCT feature multiplication is [B,d]. Now, since we have fundus feature multiplication matrix and OCT feature multiplication matrix we have to add them up and the output will be a Fused feature map. The actual formula from z value to fused feature map is given below:

$$\text{Fused} = z \odot F1 + (1 - z) \odot F2$$

- $F1$ = Fundus features (B, d)
- $F2$ = OCT features (B, d)
- $z \in (0, 1)^d$ = gate values (per feature dimension)

Now In this formula if a gate value, z for fundus feature is 1 then (1 - z) for OCT feature becomes 0 meaning we have to completely depend on fundus features. Also, if a gate value, z for fundus feature is 0 then (1 - z) for OCT feature becomes 1 meaning we have to completely depend on OCT features. But if the expected z value per batch is [0.1,0.3,0.5,0.7,0.9] which is all fraction values, we can interpret like this:

- For $z = 0.1$, $1 - z = 0.9$ (mostly have to trust OCT feature)
- For $z = 0.3$, $1 - z = 0.7$ (more trust on OCT feature)
- For $z = 0.5$, $1 - z = 0.5$ (equal trust on both features)
- For $z = 0.7$, $1 - z = 0.3$ (more trust on fundus feature)
- For $z = 0.9$, $1 - z = 0.1$ (mostly have to trust fundus feature)

So, the z value determines which modality to trust more and also pro-vides us that for a single sample which modality is contributing and dominating for the exact decision. That's why we are using the gated fusion at this point because we might not for a single sample which modality is trustworthy. There could be a situation where for a sample the OCT slices are more noisy and extract-ing information from even after proper pre-processing is hard, then the fusion has to either focus on fundus image more or only. Also, There could be another situ-ation where the Fundus image pattern could not be figured out by the encoder for any single sample, so that feature map is not trustworthy and we have to rely on OCT features then. The most perfect scenario is that both modality z values range between [0.4,0.6] where we can trust both modality and the decision comes through both of their feature patterns.

5.2.3 Classifier Head

Last Part is the classifier part where after the fused fundus with OCT feature patterns enter the MLP (Multi Layer Perceptrons) classifier . MLP classifiers are computer vision networks which are fully connected, In other words it is a feed forward Neural Network which will output hard labels (0 or 1) for a fused feature vector.

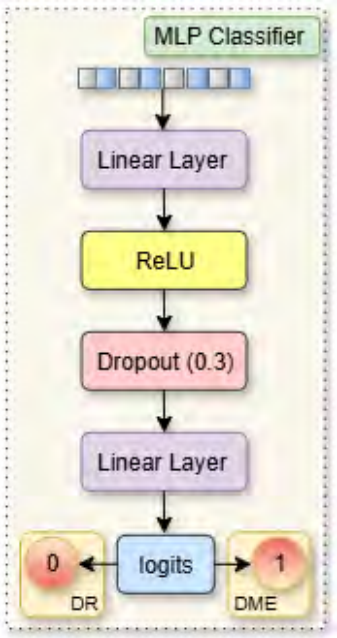

Figure 5.14: MLP classifier diagram

The input shape of this MLP classifier is [B,512] meaning a fused feature vector having 512 D. So, the first linear layer takes it as input and compresses it into [B,256]. Then it enters a ReLU block where it introduces non-linearity for complex decision boundaries. After the ReLU the shape remains the same. Then a Dropout layer for regularization. Here, Dropout(p = 0.3) meaning 30 percent of the neurons will be disabled which is a moderate regularization value. This dropout is for tackling overfitting during training. After dropout, the shape remains the same and it enters another linear layer which outputs a shape [B,2]. These are the logit scores and this enters the softmax layer from the CrossEntropyLoss function which outputs softmax probabilities for the predicted classes which is in our case, DR vs DME.

5.3 Dual Vision V2 Architecture

Our Dual Vision V2 architecture has the same five parts as Dual Vision V1 architecture. In Dual Vision V1 architecture, if a patient has n OCT slices then, we perform an MSA fusion among the n OCT slices and here in V2 we are performing a gated fusion between n OCT slices. Then in V1, we are performing Gated fusion between fused OCT feature vector and fundus feature vector but here in V2, we are performing an MSA fusion between fused OCT feature vector and fundus feature vector. So, the core difference between the two versions is that at which point we are using a fusion mechanism between modalities. Rather than the swapping of the fusion mechanism and the structure of the fusion mechanism, everything else in this

architecture version is pretty much the same. Below the Dual Vision architecture V2 diagram is given below:

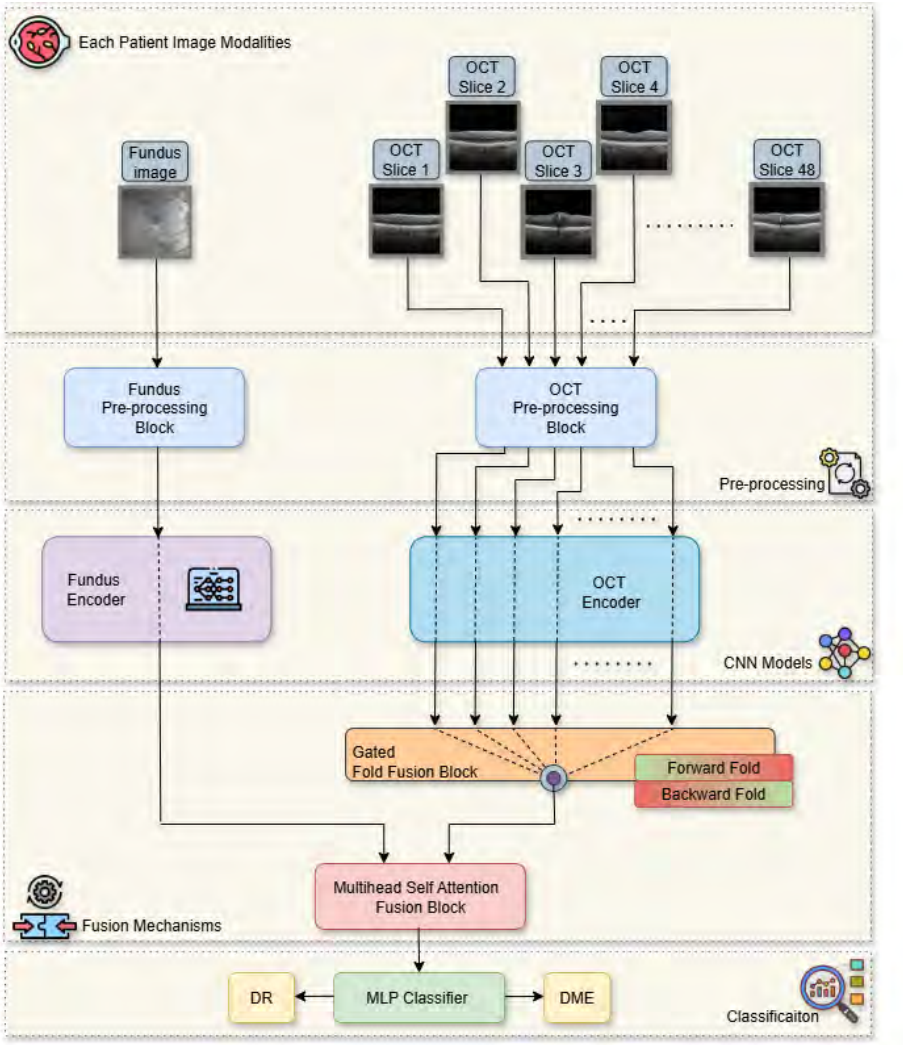

Figure 5.15: Dual Vision V2 architecture Diagram

Here, we can see that each patient modality part, pre-processing part, CNN encoders (Fundus encoder, OCT encoder) part and classifier Part are the same but the only part which is different is fusion mechanism Part.

5.3.1 Fusion Mechanisms

For n OCTs , we are performing a gated fold fusion which is different from the normal gated fusion we have described before. The fusion mechanisms between these two modalities are set this way so that we want to see which fusion at what point lets the both modality communicate better between them. Gating helps when we cannot fully trust a single modality, where MSA helps when we can totally trust each modality. Here, by trusting we are saying that our architecture will be biased towards a single modality for decision making. Even gating mechanisms do not silence a particular modality, each fused feature vector has at least a little amount of influence so that we can capture the true essence of multimodality.

Gated Fold Fusion

This fusion strategy is a bigger and more complex version of the previous gated fusion mechanism [82]. The Previous gated fusion mechanism we are using in Dual Vision V1 architecture was to fuse only two inputs/modalities (OCT fused feature and fundus features). But here we are trying to perform a gated fusion between variable OCTs meaning if a patient has n number of OCT slices, then we have n number of OCT feature vectors. So, the number of inputs becomes n for the gated fusion. But the actual gated fusion shown in Figure 5.13 has Gate1 (Z value) and Gate2 (1 - Z) value meaning we can take only 2 inputs at a time and perform a gated fusion among the pair.

So, to deal with this issue, we are constructing a looping gated fusion first. Meaning if a patient has n OCT slices feature vector we are going to perform n-1 time gated fusion between them. To clarify if a patient has 5 OCT slices then after OCT encoder we get 5 OCT slice feature vectors and at first perform a gate fusion between OCT slice 1 feature vector and OCT slice2 feature vector and get a OCT slice12 feature vector. Then we again perform a gated fusion between OCT slice12 feature vector and OCT slice 3 feature vector and get an OCT slice123 feature vector. Then we again perform a gated fusion between OCT slice123 feature vector and OCT slice 4 feature vector and get an OCT slice1234 feature vector. Lastly, we again perform a gated fusion between OCT slice1234 feature vector and OCT slice 5 feature vector and get an OCT slice12345 feature vector. So, we had to perform the fusion 4 times. But there is an issue with this looping fusion strategy. Here, the earlier slices will be more dominated than the later slices meaning slice 1,2,3 will be focused more than 4,5. Since all OCT slices together make a full OCT volume, we do not know which slices are the most effective one. The effective ones may occur last or in the middle with certain or uncertain slice intervals. Again, another issue is if the number of OCT slices is less than it may not be a problem, but for larger numbers of OCT slices, it will be a huge problem because later OCT slices will be directly neglected due to the dense earlier fused slices.

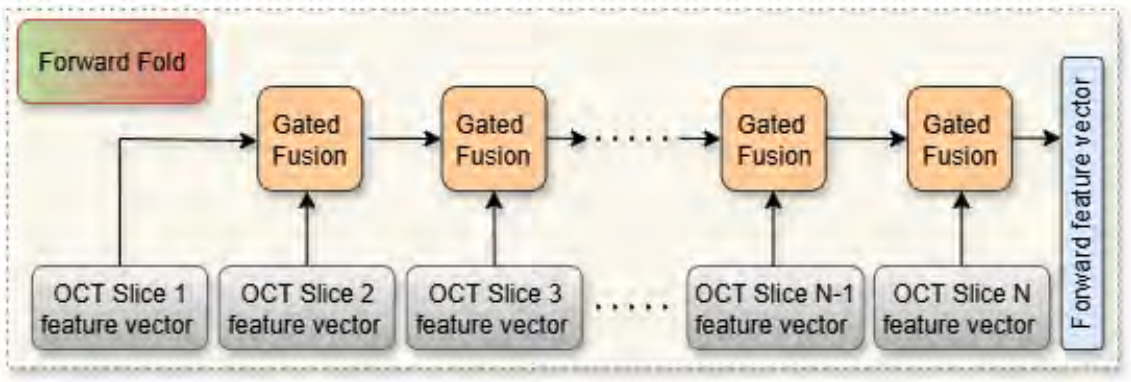

Figure 5.16: Forward fold diagram

To address this problem we have designed a new Gated Fold Fusion method which has a bi-directional point of view of OCT slices. To elaborate, rather than just forward traversal, we also rely on backward traversal as well. So, we are dividing the fusion mechanism into two folds. The forward fold which traverses forwardly

and a backward fold which traverses backwardly. Now, the forward traversal starts with earlier OCT slice 1 and 2 feature vectors and gradually moves to slice 3,4,5. . . . to later n-1,n feature vectors.

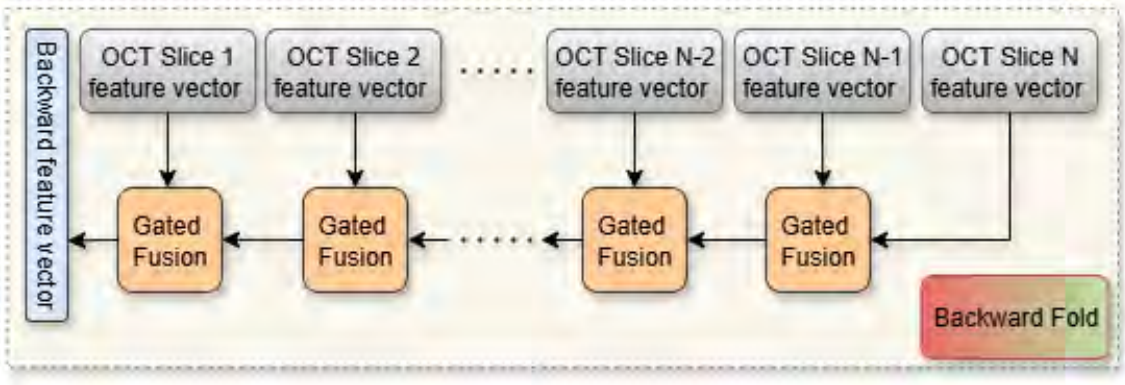

Figure 5.17: Backward fold diagram

On the other hand, the backward pass starts with later OCT slice n and n-1 feature vectors and gradually moves to earlier slice 3,2,1. We are giving the model the positional understanding of each OCT slices in both forward and backward folds to understand the arrival order of the OCT slices. To elaborate, we have a pos forward variable which determines forward traversal position value of each slice . Its is calculated by the given formula:

$$\text{pos}^{\text{fwd}}(i, N) = \frac{i}{\max(1,\ N-1)}$$

Here, N is the total number of slices and i is a value from [0,N-1]. N-1 means this number of calculations is going to happen for each gated fusion in a single passing fold. The max(1,N-1) used for division by zero cases. Now the value of pos_forward will range from 0.0 to 1.0 for any variable number of slice systems. So, if we have each patient 5 OCTs then the value list will be [0.0, 0.25,0.5,0.75.1.00]. Each value in the list corresponds to their position. But the calculation of Backward is different because we want the fusion to give importance on the later slices more because of the earlier slice gated fusion density. For this we have a pos_backward variable which determines forward traversal position value of each slice . Its is calculated by the given formula:

$$\text{pos}^{\text{bwd}}(i, N) = \frac{(N-1)-i}{\max(1,\ N-1)} = 1 - \text{pos}^{\text{fwd}}(i, N)$$

So, i is a value from [0,N-1]. N-1 means this number of calculations is going to happen for each gated fusion in a single passing fold. The max(1,N-1) used for division by zero cases. So in other words, pos_backward = 1- pos_forward. Then we put the pos_forward and pos_backword into a logit calculating function which calculates the score values of each slice position. Here, in the calculation of logits or l_t, we have a Weight Matrix W, Bias b and Recency_strength. Initially, W is b is set 0 and 1 respectively. Later the W and b changes dues to models learning progresses on each epoch. On the other hand the Recency_strength is a scaler value that determines for each fold traversal how much the gate prefers the arriving slices or

later slices. This value changes per sample traversal. Recency_strength is initialized with value 0.3 and it changes in model back propagation. So, if Recency_strength grows mean that the later slices are helpful. On the other hand, if Recency_strength decreases meaning later slices are not that helpful and lastly if Recency_strength is equal to 0 meaning the order or position does not matter. The formula is given below:

$$\underbrace{\ell_t}_{\text{logits}} = W\,[h_{t-1},\, x_t,\, pos_t] + \mathbf{b} + \text{recency\_strength} \cdot pos_t$$

$$z_t = \sigma(\ell_t), \qquad h_t = (1 - z_t) \odot h_{t-1} + z_t \odot x_t$$

Later these logit values enter a sigmoid function and produce z_t and using z_t value we calculate the h_t. Now for 5 OCT slice cases, the number of h_t values are 5 (h_0, h_1, h_2, h_3,h_4). Each h_n value is calculated with the help of their z_t and previous h_n-1 value. So for forward fold we have h0 to h_n values and the last h_n value is the final feature vector of the forward fold. Then doing the same for the backward fold, for 5 OCT slice cases, the number of h_t' values are 5 (h_0', h_1', h_2', h_3', h_4'). So for backward fold we have h0' to h_n' values and the last h_n' value is the final feature vector of the backward fold. After the two types of traversals, we have a forward feature vector and backward feature vector like shown in the Figure 5.14 and Figure 5.15. But now we have two different point of view feature vectors , one from forward and one from backward. To help the model to rely on either one point of view or partially on both we need to perform a simple gated fusion one last time between them. Below is the diagram and formula:

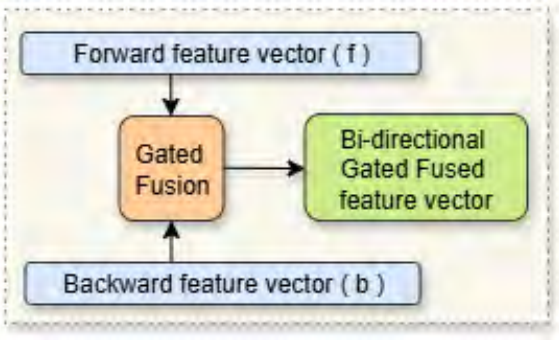

Figure 5.18: Gated fold diagram

$$\text{fused} = g \odot f + (1 - g) \odot b.$$

Here, like the traditional gated fusion we have a g value which according to Figure 5.13 is z value. This value is calculated by a linear projection followed by a sigma function. Here, f is the actual forward feature vector and b is the actual backward feature vector. This formula calculates the final feature vectors which determine the forward and backward dependencies separately or together. So, this the new gated fold fusion mechanism which helps the model to understand bi-directional point of view of OCT slices occurrence.

Multihead Self - Attention Fusion

Here, like the traditional MSA fusion we are using in the Dual Vision V1 architecture, does the same. Just the number of input tokens becomes two only which are the fundus feature vector token and fused OCT slices token. Otherwise the number of heads remains the same so the calculation steps are also the same for Dual Vision V2 architecture. The reason behind performing an MSA fusion between OCT images and Fundus images is, fundus images provides us a surface level point of view and OCT slices provides us a layered point of view of that Fundus image. Surface level lesion strongly corresponds to the disruptions among multiple layers in OCT slices. So that the model understands the relationship between each modality dynamically. In simple words, because of the query,key, value methodology we can get the context of exactly which part of the Fundus image directly communicates with which part of the OCT image.

5.4 OCTbiO Architecture

In Our OCTbiO architecture, we are focusing on how to use the OCT slices with their biomarker values. As far as we know, having 96 unique patients in our whole dataset . The dataset potential based on image was far much very good. But, When it comes to using biomarker values, we are using less amount of images. The reason behind this is the limitations of multimodal datasets on ocular diseases.

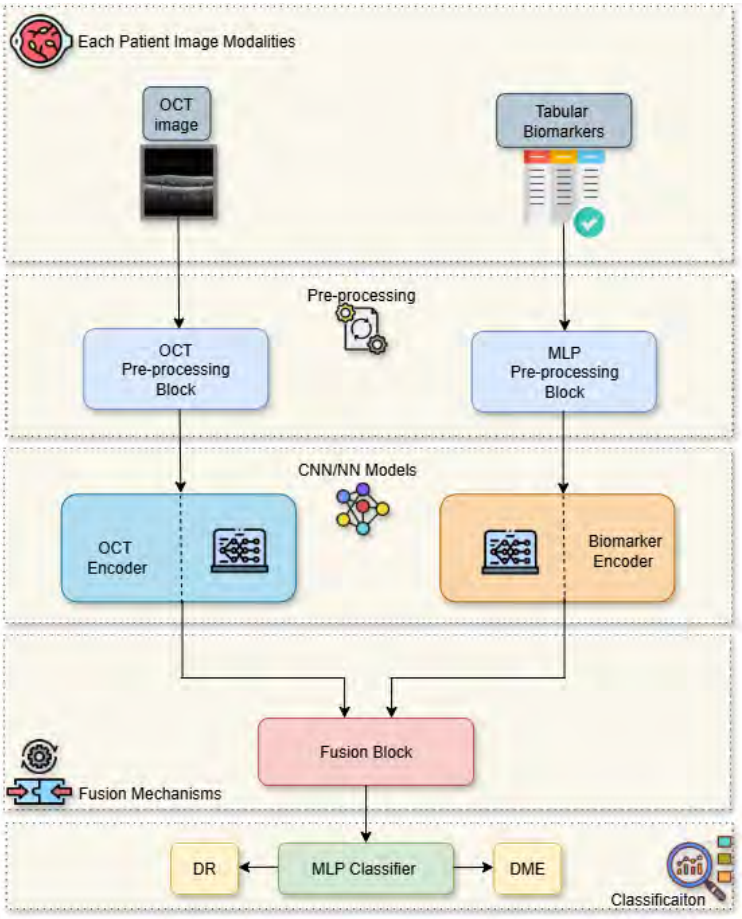

Figure 5.19: OCTbiO architecture diagram

In this proposed architecture, we have per row one OCT slice and its respective biomarkers related to that DR or DME. The biomarker tabular values not only contain the biomarker presence indication but also have two types of score values. These are the BCVA (Best Corrected Visual Acuity) and CST (Central Subfield Thickness) scores. These are the test scores used in ophthalmology which are strongly correlated to the Fundus and OCT slice modality.

This architecture is quite similar to our Dual Vision V1,V2 Architecture but here we are applying FILM fusion strategy. For preprocessing we have again used the same OCT pre-processing block from the Dual Vision architecture since exactly that flow of OCT pre-processing steps are effective for the model to understand the pattern. On the other hand, for biomarkers we have a Biomarker Pre-processing block which normalizes biomarker test scores, handles null values with imputation and checks correlation between the biomarker values.

5.4.1 Encoders and Attention Blocks

Here, we have a custom CNN model for OCT slices. This Custom model is the same as the OCT encoder we are using in the Dual Vision V1,V2 architecture. So, for each sample, an OCT image slice enters the OCT encoder and outputs an OCT feature vector. Then, respective to that OCT slice, we are passing its biomarker presence tabular values, CST and BCVA scores. How the images are passed to the OCT encoder of the Dual Vision architecture and OCTbiO architecture is different because for each sample or datapoint, the Dual Vision architecture OCT encoder requires n number of OCT slices together. Here, n is minimum 3 and maximum 49. In other words , all n OCT slices shared the same OCT encoder for a single sample. But here for each sample we have one OCT slice for OCTbiO architecture OCT encoder. The SE Block used here has the same design as the OCT encoder of the Dual Vision architecture to capture the mid level relevant patterns. On the other hand, the MLP classifier is same as the Dual Vision V1,V2 architecture.

Biomarker Encoder

Now for each OCT slice we have their respective biomarker presence values. For the biomarkers we have made a simple MLP encoder which is named Biomarker encoder in the OCTbiO architecture. We have totally 14 biomarkers and 2 score columns so before entering the Biomarker encoder, for each patient we have a 16 column long tabular values which are preprocessed. The Biomarker encoder diagram is given below:

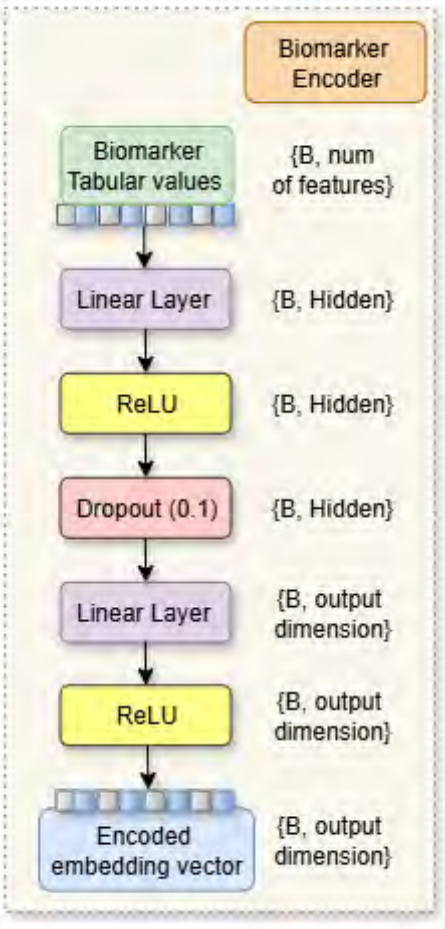

Figure 5.20: Biomarker encoder diagram

Firstly we have a linear layer which has an input shape of B, number of features . Here, B is the batch size and number of features means the tabular biomarker number of columns. For example:

tab_cols = [“IR hemorrhages”, “IR HRF”, “Partially attached vitreous face”, “Fully attached vitreous face”, “Preretinal tissue/hemorrhage”,“Vitreous debris”, “VMT”, “DRT/ME”, “Fluid (IRF)”, “Fluid (SRF)”, “Disruption of RPE”, “PED (serous)”, “SHRM”, “BCVA”, “CST”]

Each column is a feature. So , in our case the number of features is 15. The output shape is [B, hidden]. Here, hidden is the size of the hidden layer which is by default 128. This enters a ReLU layer for non-linearity where the output size does not change. So the ReLU output shape is [B, hidden]. Then we have a dropout layer where Dropout = 0.1 and the output size remains the same . Dropout helps to prevent overfitting. After Dropout we have another Linear layer which has an output size of [B, output_dimension]. Here, output_dimension is by default 64. In other words each row of our biomarker dataset transforms into an encoded dense embedding vector of size 64. Then it enters another ReLU for non-linearity and this last ReLU layer outputs the final encoded embedding vector of each Biomarker row sample.

5.4.2 Fusion Mechanisms

We are trying three different Fusion mechanisms. To clarify, we are not using all of these three fusion mechanisms altogether. In the Dual Vision V1 architecture, we are using MSA fusion on OCT slices and then later gated Fusion between fundus image and the output of MSA fusion on OCT slices and in V2 we have the reverse option. So, we are using double fusion together one by one. But here, for each try we are using only one single fusion mechanism among three fusions. Each mechanism is based on how OCT slices are connected to their respective biomarker values and how they depend on each other for the model to understand the relationship between ocular image and ocular tabular data. In our OCTbiO architecture, we are using gated Fusion, MSA fusion and FiLM fusion.

FiLM Fusion

FiLM means Feature-wise Linear Modulation. This fusion strategy is used to modulate one modality with the help of another [71]. Here, our two modalities are OCT images and OCT Biomarker tabular values. So, OCT slices are being modulated by their OCT biomarkers tabular columns. Traditional fusions like concatenation Fusion, gated Fusion, or MSA fusion usually work with the final output of the CNN models or encoders. But FiLM fusion techniques help with conditional modulations with mid CNN layers. Here, by modulating we mean that biomarker tabular features that enter the Biomarker encoders and output a dense encoded embedding vector , they help to shift and scale the OCT Encoder mid layers. Here, shifting and scaling is the conditional representation learning of the OCT clinical values.

In our FiLM fusion block, we apply it after two of our OCT encoder Layers. The first FILM Block is placed after the Conv2D BLock3 followed by the ReLU layer. The output of the ReLU layer, which is the mid-layer OCT Slice feature map, enters FiLM Fusion Site1. We represent the OCT slice feature map as X3. Now, on the other hand, we have that exact OCT biomarker tabular encoded embedded vector which enters the FiLM 1 Generator Block. The embedded vector shape is [B,d]. Here, B is the batch size and d is the dimension. In the FiLM 1 Generator Block we have a linear projection which outputs a shape [B,2C]. Here for Conv2D Block3 , C = 64 so the output shape becomes [B,128]. Now, we have to split this shape into two parts. So after splitting, the shape becomes [B,C] = [B,64] again. These two parts are gamma3 and beta3, each having the same shape. After that, we have to reshape gamma3 and beta3 by unsqueezing in order to broadcast them. Now we have the gamma3, beta 3, and X3, which is the OCT Slice feature map in hand, we just need to modulate it.

$$\text{FiLM}(F \mid c) = (1 + \gamma(c)) \odot F + \beta(c)$$

- F = OCT feature map
- $\gamma(c)$ = **per channel scale factors** (multiplicative modulation)
- $\beta(c)$ = **per channel shift factors** (additive modulation)

- $\odot$ = channel-wise multiplication

Here, the gamma value represents the scale factor per channel and beta value represents the shift factor per channel. In other words we have each channel modulation values. To define some gamma and beta value scenarios:

- **Case1:** For gamma = beta = 0, the channel remains the same, meaning that at that specific layer the biomarker tabular values have no impact on their OCT slice, which is the safe case.
- **Case2:** For gamma > 0 , the channels are amplified or boosted. So, if any of the biomarkers indicate hemorrhages, the channel's strength of detecting the hemorrhage area increases than before.
- **Case3:** For gamma < 0 , the channels are suppressed or inverted meaning for a case where the clinical scores are normal then the channel will give less attention to the OCT noise areas.
- **Case 4:** For beta > 0, the channels bias will increase
- **Case 5:** For beta < 0, the channels bias will decrease
- **Case 6:** For gamma $= 0$ and beta $\neq 0$, then the channel will only get a spatial constant shift
- **Case 7:** For gamma $\neq 0$ and beta $= 0$, then the channel will only get a positive or negative reactivation.

Depending on the gamma and beta value, gamma3, beta3, and X3 enter a FiLM Apply block, which calculates the new channel-wise feature map X3', which immediately enters the next layer. Below The FilM fusion diagram is given with each site:

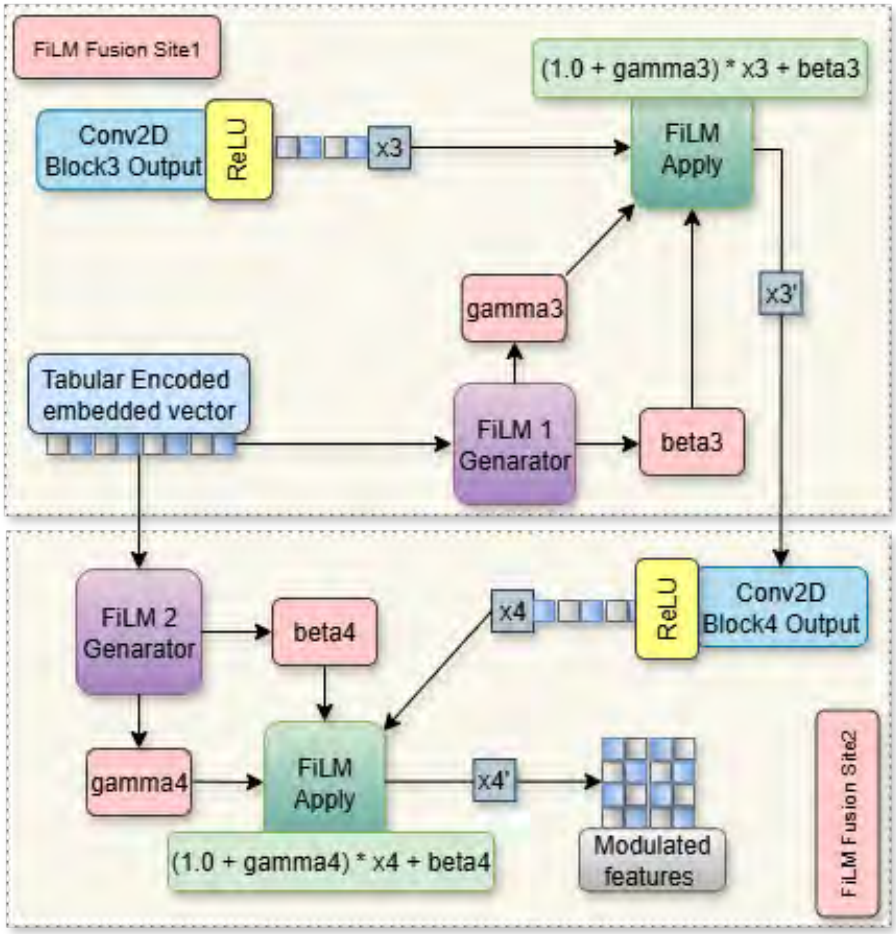

Figure 5.21: Film fusion diagram

According to the diagram, we have already discussed the FiLM Fusion Site 1. The output of the FiLM Fusion Site 1 is X3' and it enters the next convolutional layer which is Conv2D Block 4 followed by the ReLU layer. The output of the ReLU layer which is the mid-final layer OCT slice modulated feature map enters FiLM Fusion Site2. We represent the OCT slice feature map as X4. Now on the other hand we have that exact OCT biomarker tabular encoded embedded vector which enters the FiLM 2 Generator Block. Here, the FiLM 2 Generator Block and the previous FiLM 1 Generator Block do not have the same modification because of their dimension size, and they do not produce the same gamma and beta value. Now, the embedded vector shape is [B,d]. Here, B is the batch size and d is the dimension. In the FiLM 1 Generator Block we have a linear projection which outputs a shape [B, 2C]. Here for Conv2D Block3 , C = 128 so the output shape becomes [B,256]. Now, we have to split this shape into two parts. So after splitting the shape becomes [B, C] = [B, 128] again. These two parts are gamma4 and beta4, each having the same shape. After that we have to reshape gamma4 and beta4 by unsqueezing in order to broadcast it. Now we have the gamma4, beta 4 and X4 which is the OCT slice modulated feature map in hand, we just need to modulate it. Depending on the gamma and beta value, gamma4, beta4 and X4 enters a FiLM Apply block which calculates the new channel wise feature map X4'. Here, X4' is the new modulated mid-final level OCT feature map and again it is passed into later OCT Encoder final layers. After that we are not using any FiLM methodology in any other layers because X4' enters

a SE Block followed by the last convolution block.

Gated Fusion

This fusion strategy is the same one that we are using in the Dual Vision V1 architecture. In Dual Vision V1 architecture, we are gated fusing each sample fundus feature vector and fused OCT slices vector because we know there could be chances that for a single sample we might need to either binary trust any modality or fractionally trust both modality. Here, we are following the same method. We are projecting the OCT biomarkers as the same as the OCT slices dimensions for appropriate gated fusion.

There could be a situation where for a sample the OCT slices are more noisy and extracting information from even after proper pre-processing is hard then we need to wholly trust the biomarker tabular features. But since for OCTbiO architecture, we have each patient only 2 weeks of biomarker data meaning each patient has first week 49 OCT slices and last week 49 OCT slices. So total 98 OCT slices with 98 sets of biomarker tabular features. Even if all are from a single patient but these are not same week data these are time series data with a large time gap and patient medical treatment. And each OCT slice of 49 ones from a certain week is different, there could be a chance of repeated biomarker tabular features. That's why we are using a gated fusion so that even if there is very much less chance of repeated biomarkers, they cannot harm the decision boundary because gated fusion will nullify their repeated presence. As a result the model will be less likely to memorize any biomarker values.

Multihead Self - Attention Fusion

This fusion strategy is the same one that we are using in the Dual Vision V1,V2 architecture. In Dual Vision V1 architecture, we are performing between multiple OCT slices since each patient had 49 OCT slices so that the OCT slices could communicate between them. Here, the fusion is happening between each OCT slice and its respective biomarker tabular values. We are projecting the OCT biomarkers as the same as the OCT slices dimensions for appropriate MSA fusion.

The OCT slice vector itself is a token and its respective biomarker embedded encoded biomarker is a token. Both OCT token and Tabular Biomarker token size is [B,D] where B is the batch size and D is the OCT encoder output dimension size which is 128. These two tokens enter the MSA block and output. Here, num of heads = 4 meaning each of our tokens is divided into 32 size dimension sub parts in order to get multiple points of view to the fusion. The actual fusion diagrams and steps are shown in Figure 5.8 to Figure 5.12.

Chapter 6

Experiment Configurations and Results

6.1 Experiment Configuration and Setup

Based on three different architectures, they have nearly identical training configurations. The LR (Learning rate) is initially set to $1 \times 10^{-4} = 0.0001$. For updating the model's weight, the LR helps the optimizer to adjust the step size. WD (Weight Decay) is initially set $1 \times 10^{-5} = 0.00001$. This WD is a hyperparameter for L2 regularization. It helps the model penalize larger weights, preventing the model from overfitting. For the Dual Vision V1 and V2 architectures, the batch sizes are 4 and 8. Then for OCTbiO architecture, the batch size is 8 and 16. The reason behind the two different batch sizes is that the Dual Vision V1 and V2 architectures have many more images compared to the OCTbiO architecture. That's why, for the NVIDIA Tesla P100 GPU efficiency, a smaller batch size is used in the Dual Vision V1 and V2 architectures.

Moreover, the criterion variable is initialized with a loss function known as Cross-Entropy Loss. In most classification tasks, this loss function is employed. It determines how much a single sample (x)s, the actual distribution y, and its probability distribution y' matches [48]. The true label distribution here is binary or one-hot coded, meaning 0 or 1, but label smoothing prevents the model from being overconfident in its decision. So, if the true label distribution is for a 3-class system [class1, class2, class3] = [1, 0, 0], then after label smoothing, it becomes [0.95, 0.025, 0.025]. This highlights that the label smoothing, which distributes 0.05 from class 1 among classes 2 and 3, is preventing the model from becoming overconfident in predicting class 1. Thus, if we have a batch size = N, with label smoothing = ε and total K classes, the smoothed target distribution formula is given below:

$$y_{ik}^{(\text{smooth})} = \begin{cases} 1 - \varepsilon & \text{if } k = c_i \text{ (correct class)} \\ \frac{\varepsilon}{K-1} & \text{if } k \neq c_i \end{cases}$$

Here, c_i is the actual true class of a sample i in that batch. So, using this value, we can calculate the smoothed Cross Entropy Loss function with the given formula below:

$$\mathcal{L}_{\text{LS}} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{k=1}^{K} y_{ik}^{(\text{smooth})} \cdot \log(\hat{y}_{ik})$$

Here, the loss is slight when a model predicts a class that is correct with high probability, and the loss is high otherwise. In our case, the label smoothing value is set to 0.05. For optimization, we are using the Adam Optimizer and passing the LR and WD to it. The Adam Optimizer is one of the popular optimizers for weight updating [13]. The Adam Optimizer formula is given below:

1. Gradient computation
$$g_t = \nabla_w \mathcal{L}(w_t)$$
2. Exponential moving average of gradients (momentum)
$$m_t = \beta_1 m_{t-1} + (1-\beta_1) g_t$$
3. Exponential moving average of squared gradients (RMS scaling)
$$v_t = \beta_2 v_{t-1} + (1-\beta_2) g_t^2$$
4. Bias correction
$$\hat{m}_t = \frac{m_t}{1-\beta_1^t}, \qquad \hat{v}_t = \frac{v_t}{1-\beta_2^t}$$
5. Weight update rule
$$w_{t+1} = w_t - \eta \cdot \frac{\hat{m}_t}{\sqrt{\hat{v}_t} + \epsilon} - \eta \cdot \lambda w_t$$

Where:

- m_t = moving average
- w_t = current parameter (weight or bias)
- $\eta = lr$ = learning rate
- $\lambda = weight\_decay$ (L2 penalty term)
- β_1, β_2 = decay rates (default: 0.9, 0.999)
- ϵ = stability constant (1e-8)

As declared, the LR initially 1×10^{-4}, it needs to be controlled later as epochs go by. For that, a scheduler is used that controls the LR over later epochs. The reason behind this is that the model need to converge at every point, and that's why the LR needs to be reduced a little at each step to avoid any local minima being stuck. In the scheduler, we are passing the optimizer (Adam) because its LR needs to be updated based on the model training and validation loss. The factor parameter is set to 0.2, as when the factor is triggered, then the current LR will be multiplied by 0.2 to make it smaller than before. The minimum LR is set as 1×10^{-7}, meaning the LR will be reduced to a minimum of 1×10^{-7} and not exceed it.

Next, for each training fold, the default number of epochs is 30 for Dual Vision V1, V2 architecture. The number of epochs is 50 for the OCTbiO architecture, meaning there will be at most 30 or 50 epochs for each fold. After various attempts, we observed that most of the folds are overfitting by an average of 30 epochs. To prevent extensive overfitting, we are using early stopping. The early stopping value is set to 5, meaning that while training and validation, the training loop will be stopped if the validation loss per epoch does not decrease.

6.2 Experiment results

For each architecture, K-fold validation is used, resulting in a total of 5 folds. The same training and validation configuration is done for all folds. For each fold, we initialize our model, and that model is trained using that exact fold's train set and validation set. For testing the data, we are using all the fold models for prediction. For instance, we are assembling the prediction knowledge from every fold model. Since we have 5 folds, we have model_fold1.pt, model_fold2.pt, model_fold3.pt, model_fold4.pt and model_fold5.pt. Instead of relying on a single fold model's decision, we can utilize the soft average, which incorporates every fold's decision. The test F1-score, recall, precision, support, AUC and ROC (Receiver Operating Characteristics) curve formula.

- **Precision:**

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

- **Recall:**

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

- **F1-Score:**

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

- **Support:**

$$\text{Support}_{\text{class}=c} = \text{Number of samples with true label} = c$$

For binary:

$$\text{Support for class } 1 = TP + FN$$

$$\text{Support for class } 0 = TN + FP$$

- **True Positive Rate (TPR):**

$$\text{TPR} = \frac{TP}{TP + FN}$$

- **False Positive Rate (FPR):**

$$\text{FPR} = \frac{FP}{FP + TN}$$

- **ROC Curve:**

$$ROC = \{(\text{FPR}(\tau),\ \text{TPR}(\tau)) \mid \tau \in [0, 1]\}$$

- **AUC:**

$$AUC = \int_0^1 \text{TPR}(\text{FPR})\, d(\text{FPR})$$

6.2.1 Dual Vision V1 Architecture Results

The 5 fold based (Train vs Validation) accuracy and loss with DR and DME classification report, ROC curve and confusion matrix are shown in figure 6.1:

(Train vs Validation) Acc and Loss:

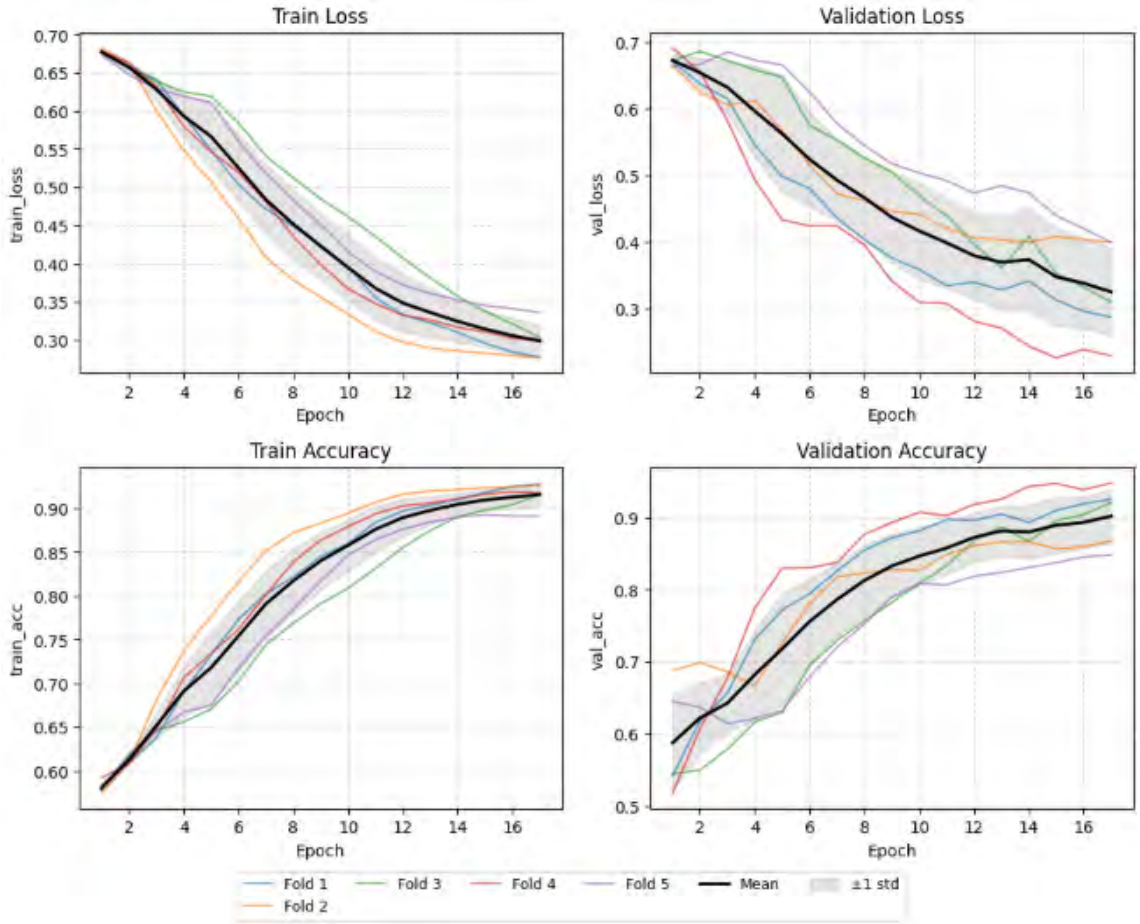

Figure 6.1: Mean loss and accuracy curve among 5 folds (Dual Vision V1)

From the training and validation curves of the Dual Vision V1 architecture, we can see strong generalization as the curves show effective convergence. The loss curve decreases steadily. We can notice the sharpest decline in the first 6-8 epochs before stabilizing. In the accuracy curves, we can observe a consistent improvement, resulting in high accuracy at the end. There is a minimal gap between training and validation accuracy curves, indicating that weight decay, label smoothing and LR scheduler controlled the overfitting. However, we showed this curve until 16

epochs because, as we triggered early stopping to prevent overfitting, the number of epochs is not the same for every trial. Therefore, for the sake of comparative analysis, we used the minimum number of epochs obtained, which is 16. However, the smooth mean curve confirms that the Dual Vision V1 architecture is stable and well-generalizing.

Classification Report:

	precision	**recall**	**F1-score**	**support**
DR	0.95	0.95	0.95	170
DME	0.96	0.96	0.96	212
accuracy			0.96	382
macro avg	0.95	0.96	0.95	382
weighted avg	0.96	0.96	0.96	382

Table 6.1: Classification report of Dual Vision V1 architecture

From the classification report of the Dual Vision V1 architecture, we can see a strong predictive performance for both classes. For DR and DME classification, the model achieves a precision, recall and F1-score of 0.95 and 0.96, respectively, for 382 data points in total. The overall accuracy is 96% with macro average and weighted average maintaining similar percentages. This proves the effectiveness of classification.

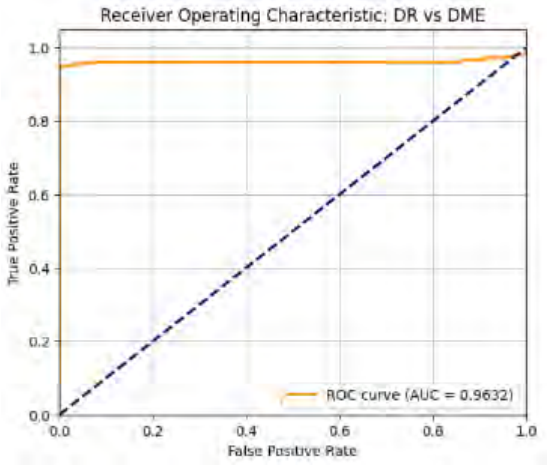

Figure 6.2: ROC curve of Dual Vision V1 architecture (DR vs DME)

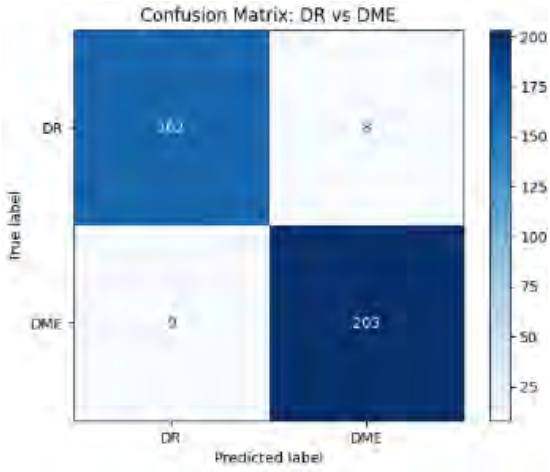

Figure 6.3: Confusion matrix of Dual Vision V1 architecture (DR vs DME)

From the confusion matrix, we can see that, for a balanced support (170 vs 212), the data is not biased towards one class. Among the 170 DR classes, the model correctly predicted 162 classes as DR. Similarly, among the 212 DME classes, the model correctly predicted 203 classes as DME. 8 DR classes and 9 DME classes

were misclassified as false positives. The ROC curve visualizes how well the model separates the DME class from the DR class. From the ROC curve, we can see the model achieves an AUC of 0.9632. The curves are close in the top left corner, indicating a high sensitivity for class distinction. The high AUC value indicates clinical reliability of this architecture.

6.2.2 Dual Vision V2 Architecture Results

The 5 fold based (Train vs Validation) accuracy and loss with DR and DME classification report, ROC Curve and Confusion matrix are shown in figure 6.4:
(Train vs Validation) Acc and Loss:

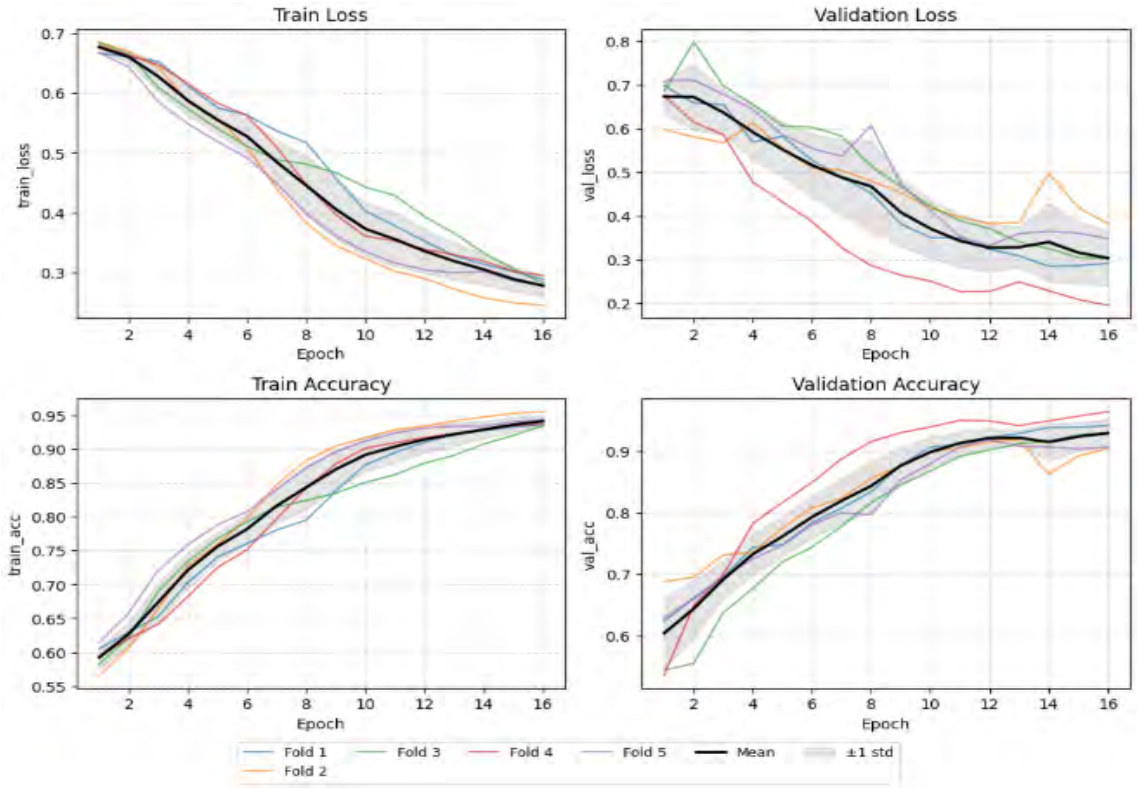

Figure 6.4: Mean loss and accuracy curve among 5 folds (Dual Vision V2)

For the Dual Vision V2 architecture, the loss and accuracy curves for training and validation show consistent convergence across all five folds. This indicates the model is learning and the effectiveness of the Adam and Scheduler optimizer. The narrow standard deviation band indicates the robustness across fields. The overfitting is also controlled as the validation curve closely follows the training curve. This shows the effectiveness of weight decay, label smoothing, and the LR scheduler. Overall, the loss and accuracy curves suggest that the model is competent for DR vs DME classification.

Classification Report:

	precision	recall	F1-score	support
DR	0.95	0.99	0.97	171
DME	0.99	0.96	0.98	212
accuracy			0.97	383
macro avg	0.97	0.98	0.97	383
weighted avg	0.97	0.97	0.97	383

Table 6.2: Classification report of Dual Vision V2 architecture

According to the classification report for the Dual Vision V2 architecture, we observe strong predictive performance for both classes. For DR classification, the model achieves a precision of 0.95, a recall of 0.99, and an F1-score of 0.97, resulting in a total of 171 DR classes, which can be considered excellent performance. Again, for DME classification, we can see an outstanding performance of the model, with precision, recall, and F1-score of 0.99, 0.96, and 0.98, respectively, for 212 support. The overall accuracy is 97%, with both the macro average and weighted average maintaining a high performance of 0.97 for a total of 383 classes.

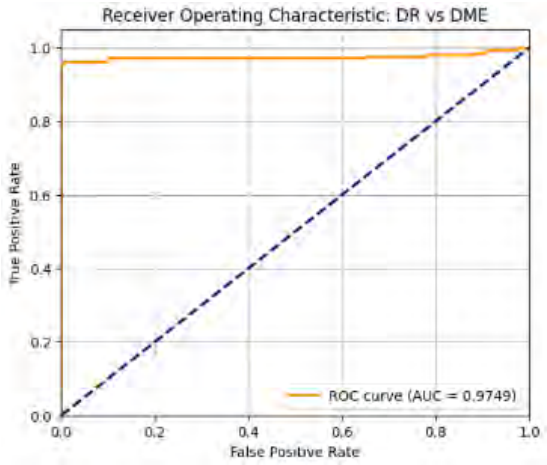

Figure 6.5: ROC curve of Dual Vision V2 architecture (DR vs DME)

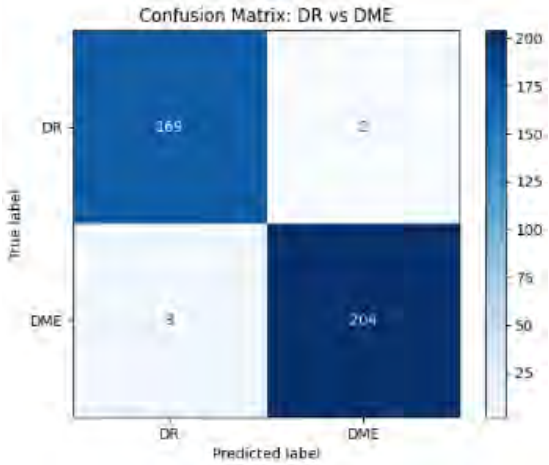

Figure 6.6: Confusion matrix of Dual Vision V2 architecture (DR vs DME)

From the ROC curve of the Dual Vision V2 architecture, we achieved an AUC of 0.9749 which can be considered as an outstanding performance of the model. This indicates that the model can distinguish between the DME class and the DR class with very high accuracy, as the curve approaches the top left corner of the ROC. From the confusion matrix, we can see that for 383 data points, the model correctly predicted 169 classes as DR with a misclassification for only 2 cases. And for DME, the model correctly predicted 204 courses with a misclassification of 8 cases. The lower rate of misclassification indicates that this architecture is highly reliable, and the changed fusion mechanism actually has a positive effect on the model's overall performance.

6.2.3 Dual Vision V1,V2 Architecture Hypothesis

As our Dual Vision V1, V2 architecture is focusing on using one Fundus image and n OCT slices per patient, for each case, we are using different OCT count examples. For each case, we are adding two more OCT slices. So, our experiment cases in this architecture are:

Fundus images	**OCT slices**	**Total Images**
1	3	4
1	5	6
1	7	8
1	9	10
1	11	12
1	13	14
1	15	16
1	24	25

Table 6.3: Fundus + n OCT slice experiment cases

We are experimenting with a minimum of 3 OCT slices and a maximum of 24 OCT slices per patient. For each system, all patients are in the same configuration. The OCT images taken for 3 OCT systems are the same as the OCT images taken for 5 OCT systems, meaning the previous OCT slices are reserved with two new slices. The reason behind this methodology is that we wanted to prove a hypothesis that more OCT images can not necessarily lead to better accuracy.

- **Null Hypothesis (H0):** A larger number of OCT slices does not necessarily lead to better accuracy.
- **Alternative Hypothesis (H1):** A larger number of OCT slices always leads to better accuracy.

So, based on our Null hypothesis and Alternative hypothesis, we have tried all Fundus + n OCT slice Experiment Cases table 1 and got the results below:

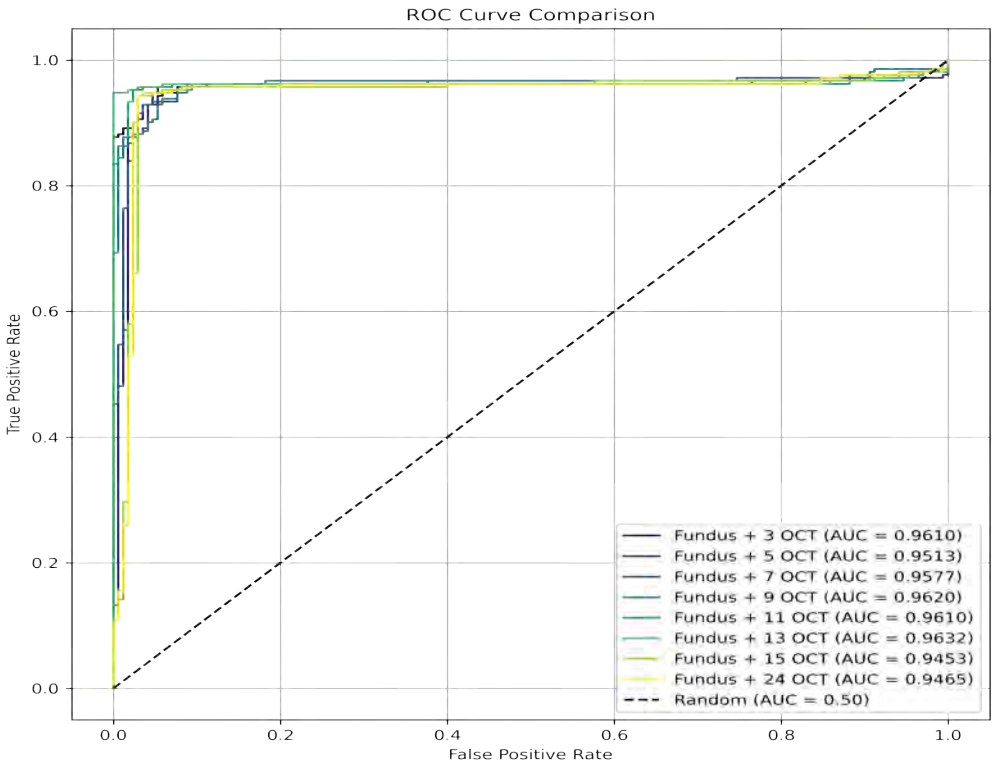

Figure 6.7: All experiment ROC curves

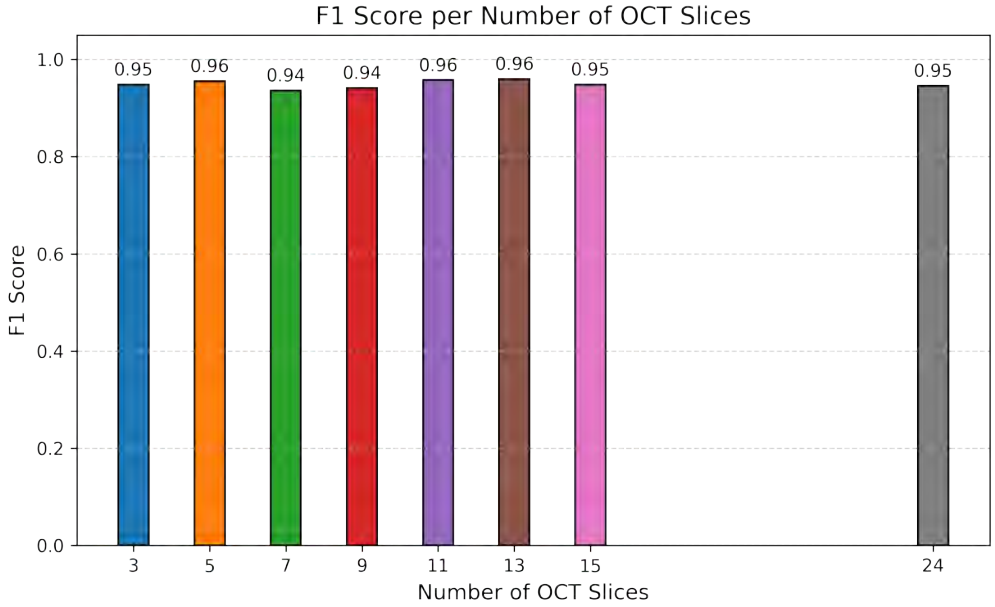

Figure 6.8: All experiment F1-scores

Here, by looking at Figure 6.7, the AUC value does not have an increasing trend, meaning a system where fewer OCT Slices are integrated can give a better AUC value than a system where more OCT Slices are integrated. In figure 6.8 we observe the same trend for F1-scores across all systems, which also do not exhibit any increasing trend, indicating that we do not require more OCT slice integration for

improved results. Since we are choosing each slice at an interval, choosing the correct interval to pick the right OCT slice can lead to better results than integrating all without any reason. Here, the training and validation times also increase as the number of slices increases, and we cannot use larger batch sizes due to GPU memory restrictions. Based on the results, we can support our null hypothesis, which states that a larger number of OCT slices does not necessarily lead to better accuracy.

In real-life scenarios, having more OCT slices means having more points of view for the models, which theoretically results in better accuracy. Since we are picking OCT slices in an interval, there could be a scenario where we might integrate such OCTs, which are part of the volume but noisier, and that affects our accuracy. Since we are performing the same pre-Processing for all OCT slices, there could be a situation where any one or two of the OCT slices are pre-processed negatively, which can also affect our accuracy, potentially leading to our Null hypothesis being proven.

6.2.4 OCTbiO Architecture Results

The 5 Fold based (Train vs Validation) Accuracy and Loss with DR and DME classification report, ROC Curve and Confusion matrix are shown in Figure 6.9:

(Train vs Validation) Acc and Loss:

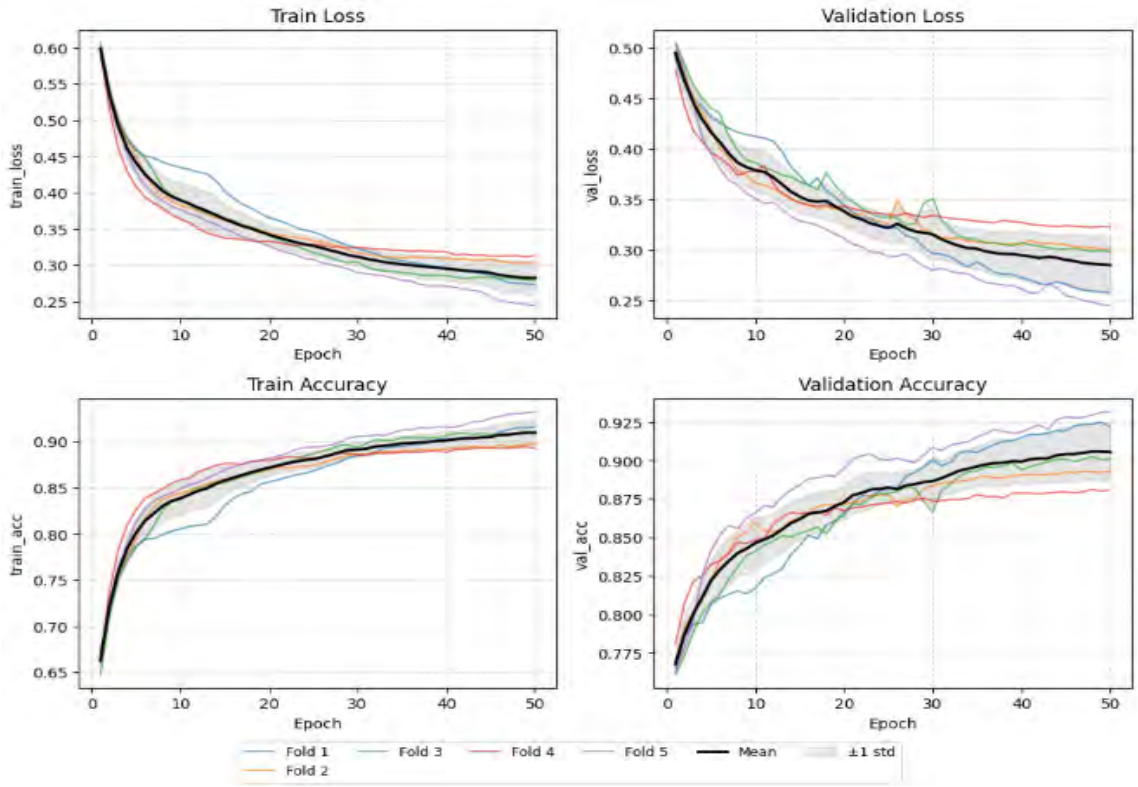

Figure 6.9: Mean loss and accuracy curve among 5 folds (Gated Fusion)

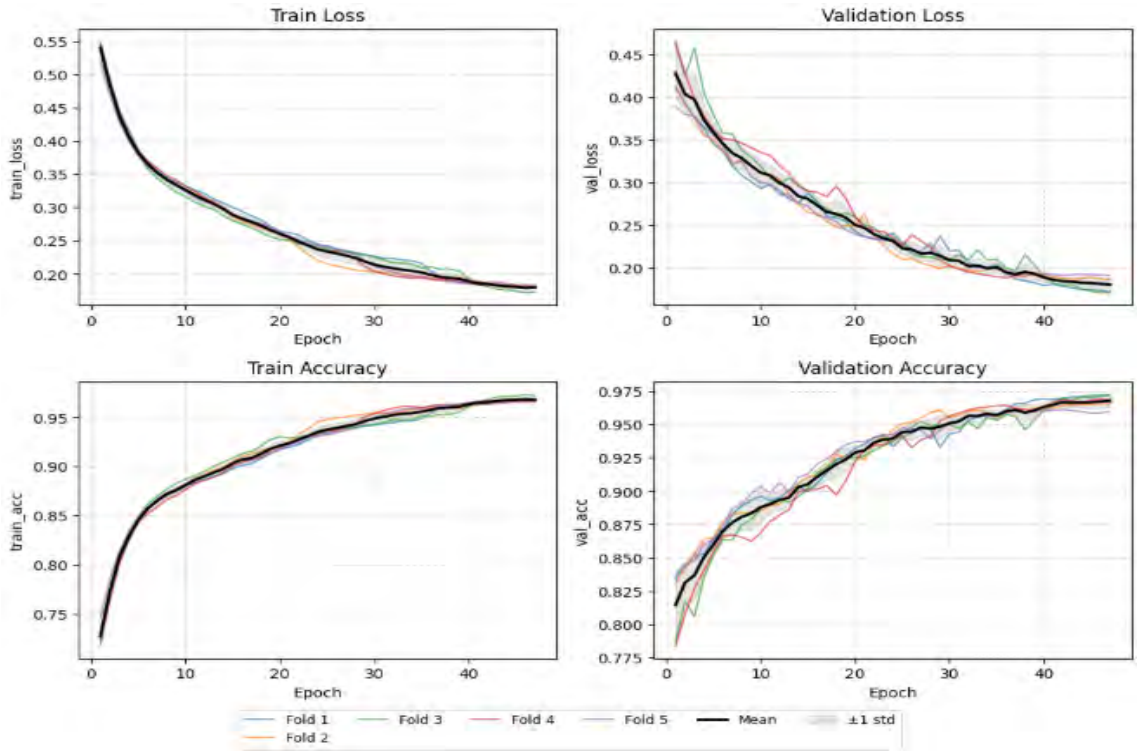

Figure 6.10: Mean loss and accuracy curve among 5 folds (MSA)

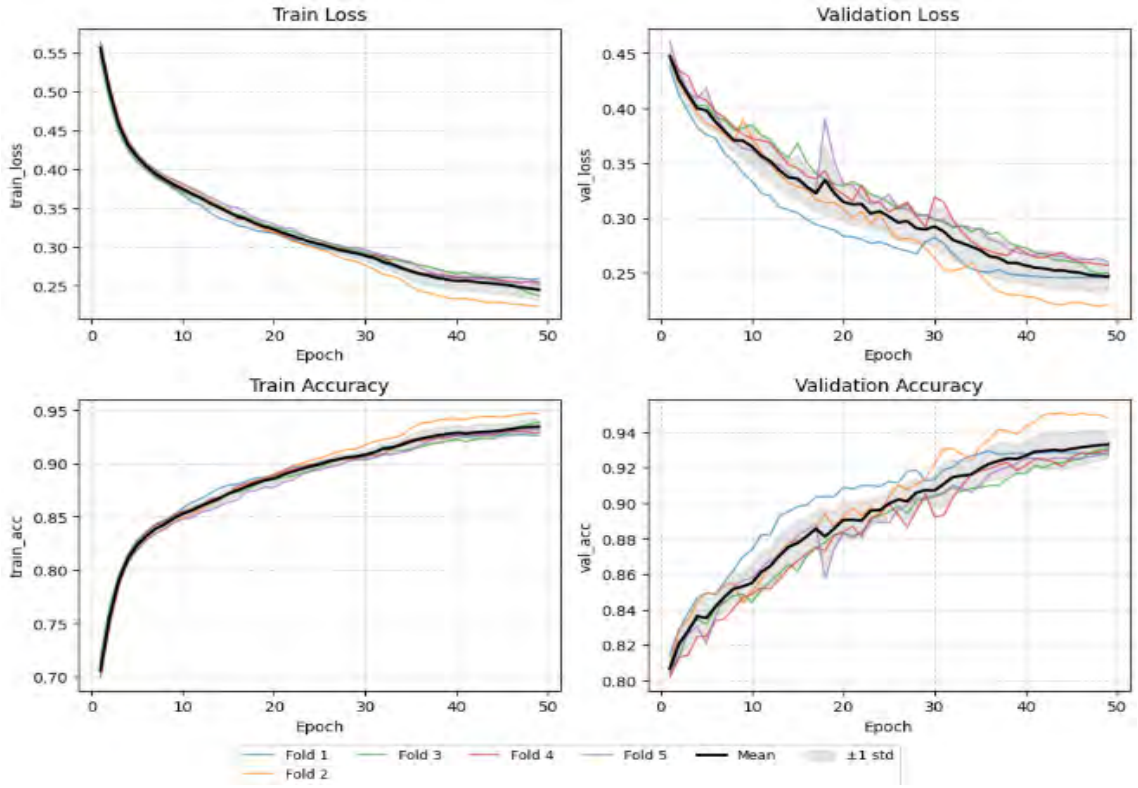

Figure 6.11: Mean Loss and Accuracy Curve Among 5 Folds (FiLM Fusion)

From the above mean loss and accuracy curves of the OCTbiO architecture, which fuses OCT images using gated fusion, MSA and FiLM fusion, we can observe

that MSA exhibits a smooth and steady decrease in both loss curves, with minimal fluctuations. Additionally, the accuracy curve rises rapidly, indicating rapid convergence, strong generalization, and minimal overfitting. The architecture using FiLM fusion also shows a steady decline, but a bit noisy validation loss and accuracy curve. This indicates a slightly slower learning compared to multihead self-attention. In contrast, the architecture using gated fusion shows a higher loss in training and validation with slightly lower accuracy.

Classification report:

	precision	**recall**	**F1-score**	**support**
DR	0.86	0.96	0.91	588
DME	0.97	0.89	0.93	824
accuracy			0.92	1412
macro avg	0.92	0.93	0.92	1412
weighted avg	0.93	0.92	0.92	1412

Table 6.4: Classification report of gated fusion (DR vs DME)

	precision	**recall**	**F1-score**	**support**
DR	0.96	0.98	0.97	588
DME	0.98	0.97	0.98	824
accuracy			0.97	1412
macro avg	0.97	0.97	0.97	1412
weighted avg	0.97	0.97	0.97	1412

Table 6.5: Classification report of MSA fusion (DR vs DME)

	precision	**recall**	**F1-score**	**support**
DR	0.90	0.97	0.93	588
DME	0.98	0.92	0.95	824
accuracy			0.94	1412
macro avg	0.94	0.95	0.94	1412
weighted avg	0.94	0.94	0.94	1412

Table 6.6: Classification report of FiLM fusion (DR vs DME)

According to the classification report, after implementing three fusion strategies, the total support is 1,412, comprising 588 DR classes and 824 DME classes. The precision, recall, and F1-score for DR classification are 0.86, 0.96, 0.91 for gated

fusion, 0.96, 0.98, 0.97 for MSA, and 0.90, 0.97, 0.93 for FiLM fusion. For DME classification, the precision, recall and F1 score are 0.97, 0.89, 0.93 for gated fusion, 0.98, 0.97, 0.98 for MSA, and 0.98, 0.92, 0.95 for FiLM fusion. The overall accuracy is 0.93, 0.97 and 0.94, respectively, for the 3 fusion strategies. In short, using MSA fusion to fuse the OCT images yields the best performance among the three OCT-biO architectures, with high precision, recall, F1-score, overall accuracy, as well as macro average and weighted average.

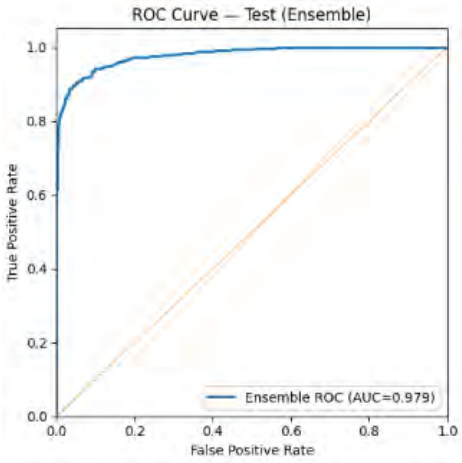

Figure 6.12: ROC curve of gated fusion (DR vs DME)

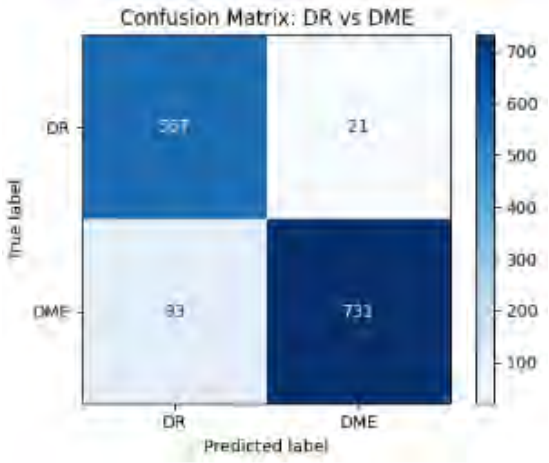

Figure 6.13: Confusion matrix of gated fusion (DR vs DME)

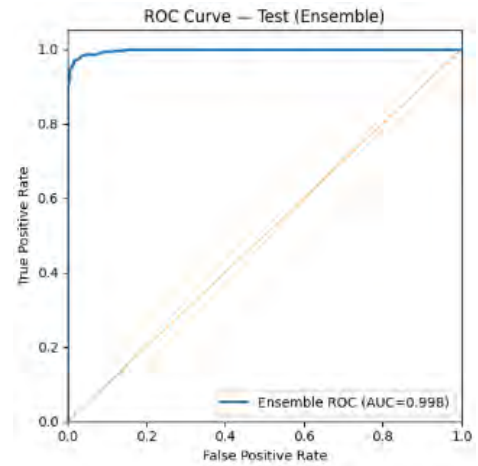

Figure 6.14: ROC curve of MSA fusion (DR vs DME)

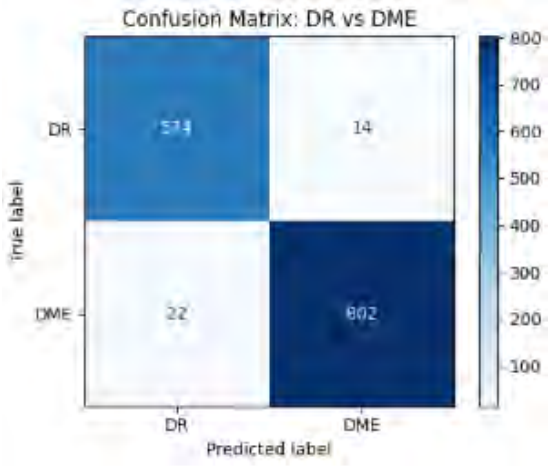

Figure 6.15: Confusion matrix of MSA fusion (DR vs DME)

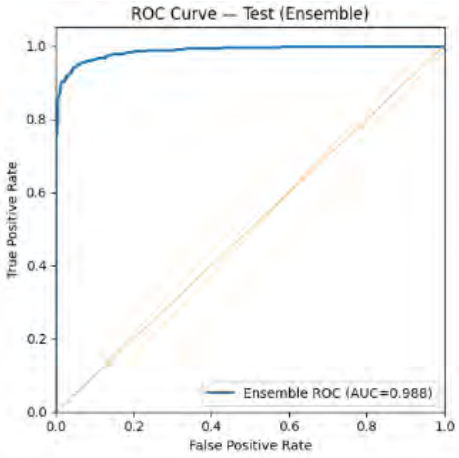

Figure 6.16: ROC curve of FilM fusion (DR vs DME)

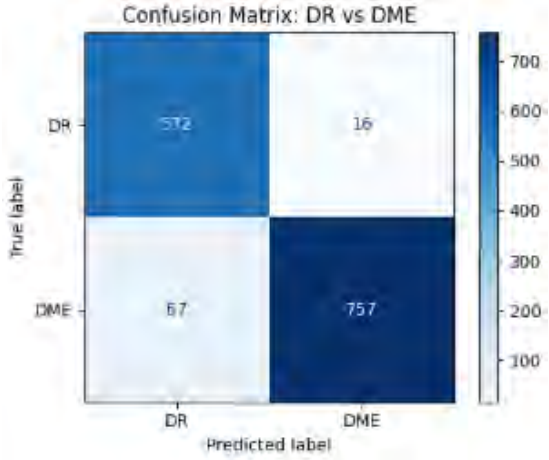

Figure 6.17: Confusion matrix of FilM fusion (DR vs DME)

From the ROC curves and confusion matrix, we can see that applying MSA fusion outperforms the other two fusion mechanisms. The AUC for gated fusion is 0.979. MSA achieves an AUC of 0.998, while FiLM fusion yields an AUC of 0.988. The second ROC curve is closer to the top left corner. However, the other two fusion mechanisms also give a decent output. In the confusion matrices, we can see that there were a total of 1412 data points with both DR and DME classes. On the OCTbiO with gated fusion architecture, 567 data points were correctly classified as positive DR, and 731 data points were correctly guessed as positive DME. But the model misclassified 21 cases as DR, but in reality, they were cases of DME.

Additionally, there were 93 cases where the model incorrectly predicted DME, but they were actually cases of DR. The false positive DR and DME cases for OCTbiO with an MSA fusion architecture are significantly lower. There were 14 cases where the model failed to predict a DME case and 22 cases were incorrectly predicted as DR class. But the model gave an outstanding performance in predicting true DR or true DME cases. A total of 574 cases were correctly predicted as DR and 802 cases were correctly predicted as DME. Finally, the confusion matrix of OCTbiO with FiLM fusion architecture shows that 572 and 757 cases were correctly predicted as DR and DME, respectively, with 16 and 67 cases misclassified. However, despite achieving outstanding performance from all three architectures that utilize OCT images and biomarkers, the real-life applicability for automated diagnosis remains low because the biomarkers themselves are labels collected by annotators.

6.2.5 Comparative Analysis of all Architectures

All Architecture Metrics Comparison

Before building our architecture, we worked with only one modality too and applied some pretrained models (ResNet50, EfficientNetB0, Inception V3) on fundus and OCT images separately. Then we used MLP on the biomarkers. After this, we developed two distinct architectures, each with its own versions based on the fusion mechanisms employed. In total, there are 5 models. Then, we built custom CNN models and added variations with different fusion techniques in various ways. Below

are comparative metrics (Parameter Count, FLOPs considering end-to-end pipeline, which includes pre-processing, and Inference Runtime) given in tabular format for batch size 1 and image size 224x224:

Encoders/Architectures	**Parameter Count**	**FLOPs**	**Inference Runtime**
Fundus Encoder	678,668	27.30 G FLOPs	5.299 ms
OCT Encoder (1 OCT Slice)	279,368	13.90 G FLOPs	2.804 ms
Dual Vision V1	1,746,774	86.80 G FLOPs	19.289 ms
Dual Vision V2	2,568,668	93.43 G FLOPs	19.843 ms
OCTbiO Gated Fused	364,430	27.81 G FLOPs	3.070 ms
OCTbiO MSA Fused	728,330	27.81 G FLOPs	3.143 ms
OCTbiO FiLM Fused	348,170	27.84 G FLOPs	3.265 ms

Table 6.7: Comparison of efficiency metrics

All Encoder/Architecture Accuracy Comparison

The accuracy, F1-score with AUC comparison table is given below:

Modality	**Model/ Architecture**	**Accuracy**	**F1-score**	**AUC**
Fundus	ResNet50	76.89%	0.794	0.840
	EfficientNetB0	79.22%	0.806	0.856
	Inception V3	82.89%	0.855	0.911
	Fundus Encoder	55.87%	0.6236	0.5415
Biomarker	MLP	84.48%	0.859	0.926
OCT	ResNet50	94.24%	0.943	0.969
	EfficientNetB0	94.56%	0.942	0.963
	Inception V3	95.32%	0.950	0.965
	OCT Encoder	94.64%	0.948	0.984
Fundus + OCT	Dual Vision V1	95.52%	0.961	0.963
	Dual Vision V2	97.39%	0.975	0.974
OCT + Biomarker	OCTbiO Gated	91.93%	0.928	0.979
	OCTbiO FiLM	94.12%	0.948	0.988
	OCTbiO MSA	97.45%	0.978	0.998

Table 6.8: Comparison of evaluation metrics

Implementing pretrained models on fundus and OCT images separately was a part of our ablation study. For fundus images, ResNet50, EfficientNetB0, and Incep-

tion V3 achieved accuracies of 76.89%, 79.22%, and 82.89%, respectively. Applying these models to OCT images only had a better performance of 94.24%, 94.56% and 95.32% accuracy.

Additionally, we developed custom fundus and OCT encoder models, which we also utilized in our multimodal architecture. The fundus encoder gave an accuracy of 55.87%, F1-score of 0.6236, and AUC of 0.5415. On the other hand, the OCT encoder model achieved an accuracy of 94.64%, an F1-score of 0.948, and an AUC of 0.984. Implementing MLP on OCT Biomarkers showed an accuracy of 84.48%. However, integrating more than one modality on the same dataset outperforms the results of a single modality in accuracy, F1-score, and AUC, which proves our claim that multimodal deep learning improves the overall performance than relying on a single modality.

The ablation study with single-modality experiments highlights the limitations of relying solely on fundus, OCT, or biomarkers. The results of the pre-trained models on OCT slices are better than those on fundus images. Furthermore, the OCT encoder also gives a better performance than the fundus-only encoder. This suggests that OCT slices can capture richer layered structural information and offer greater reliability for diagnosis. Yet, our custom CNN models, which incorporate modalities, demonstrate that multimodality outperforms the single modality experiments. Also, surpassing the OCT encoder which already has good performance.

Now, as our multimodal system has a great success in this multimodal ocular disease research, we explored different architectures and created different versions of them. OCT slices fused with biomarkers showed an accuracy of 91.93%, 94.1%, and 97.45%, with an F1-score of 0.928, 0.948, and 0.978, respectively. They also achieved AUC of 0.979, 0.988, and 0.998, respectively. Our numbers show that our architectures do show high performance. On the other hand, the OCT biomarkers are directly annotated from the OCT slices by the experts. Despite good results, we also need this approach for real-life end-to-end diagnosis. This application has limited applicability compared to Dual Vision architectures. However, trying and experimenting with these architectures helped us compare the effectiveness of the fusion strategies.

Finally, our Dual Vision V1 architecture has 95.52% accuracy, a 0.961 F1-score, and a 0.963 AUC. These values demonstrate the superiority of multimodal approaches that utilize both fundus images and OCT slices. The Dual Vision V2 architecture has 97.39% accuracy, a 0.975 F1-score, and a 0.974 AUC. The step towards multimodality significantly improved diagnostic performance, confirming the hypothesis that combining fundus and OCT enhances the model's ability. This success also highlights the effectiveness of its advanced bi-directional gated fusion strategy, which ensures balanced feature contributions across multiple OCT slices. This special gated fusion strategy reduces the biases of earlier slices and captures both positional relevance from backward and forward traversals. Also, the MSA enhanced the integration with fundus features. This strategy helped the model focus on the most critical features through communicating and keeping their spatial order out of influence. Therefore, Dual Vision V2 architecture provides the most practical, reliable, and robust solution for DR vs DME classification.

Qualitative Analysis with existing papers

The currently existing benchmark datasets and research papers generally focus on binary classification of the presence of DR in imaging data. Chen et al. [29] experimented with the NASNet-Large pretrained deep learning model on the widely recognized EyePACS dataset, validating its performance across multiple independent datasets and demonstrating generalization. Abdussamed et al. [31] reported near-perfect results on Kaggle DR and MESSIDOR datasets. Although these studies proposed capable methods of detecting the presence of DR with less complex and reproducible architectures, a near-field study of differentiating vision-threatening macular edema among the DR patients remains less explored. This classification provides specific insights and richer structural information. Our study aims to fill this research gap by analysing subtle fluid accumulation and detecting retinal layer-level distortions. Although the overall complexity of this research's architecture is higher, our study addresses a research gap by contributing novel insights. Our study is important since it supports clinical prioritization and advanced decisions.

Another qualitative component of our study is using a dataset that allows multimodality. Numerous works in this medical field have utilized single-modality data. Bressler et al [74] worked on AI-based autonomous screening for detecting DME with fundus photography only. While Zhu et al [85] used 7146 3D OCT images to detect DME presence on DR patients using a 3D CNN. Both of them performed well with a comparatively high accuracy. However, these approaches extract only one type of feature information and predict the presence of modular edema among DR cases. In our ablation study, we executed a similar approach. However, we found that DR and DME are overlapping diseases. For instance, DME is the vision-threatening instance on the DR spectrum. So, separating these two classes relying on a single modality may not fulfill our research objective. Following the successful contribution of multimodality in the latest deep learning architectures, we tried to solve the classification using a more insightful approach. Our research aims to provide more holistic insights into the problem by combining complementary features from multiple sources. This approach allows the integration of both surface-level and cross-sectional cues, making it a more robust, reliable, and resilient approach.

Moreover, pretrained models show strong performance in these clinical classification tasks. However, for medical-based research, despite having a lower complexity, custom CNN models have been proven to have less computational overhead than the pretrained models. The task-specific custom CNN models offer computational efficiency, lower parameters, and FLOPS [63]. Taking this into account, in the ablation study of this research, pretrained models such as RestNet, EfficientNet, and InceptionV3 were used. With a target of lowering the parameters and making the model lightweight, in our proposed architecture, we worked with a task-specific custom CNN model that is domain-specific, fully tailored. By offering us high flexibility, the models performed well in terms of quantitative metrics.

In short, this research not only considers quantitative parameters but also qualitative standards.

Quantitative Comparison with existing benchmark

The novelty of our research is that our model classifies between DR vs DME in a multimodal approach. Also, we are using multiple and unique fusion mechanisms with introducing explainability based on the fusion mechanisms.

Below are comparative tables of our best architecture with the existing works in related fields.

Study	Experiment Data	Architecture	Accuracy	Specificity	Sensitivity	Test Bed
Prabhushankar Study [39]	OCT	ResNet18	70.15%	0.608	0.794	NVIDIA GeForce GTX TITAN X (12GB)
	Biomarker	MLP	79.87%	0.826	0.771	
	OCT + Biomarker	ResNet18 + MLP with late fusion	82.33%	0.742	0.904	
Our Study	OCT	OCT Encoder	94.64%	0.959	0.934	NVIDIA Tesla P100 (16GB)
	Biomarker	MLP	84.48%	0.896	0.808	
	OCT + Biomarker	OCT Encoder + MLP with Gated fusion	92.57%	0.964	0.887	
	OCT + Biomarker	OCT Encoder + MLP with FiLM fusion	94.57%	0.972	0.918	
	OCT + Biomarker	OCT Encoder + MLP with MSA fusion	97.47%	0.976	0.973	
	OCT + Fundus	Dual Vision V1	95.52%	0.952	0.957	
	OCT + Fundus	Dual Vision V2	97.39%	0.988	0.962	

Table 6.9: Comparison of evaluation metrics with benchmark

This paper experimented with the OLIVES dataset with multiple strategies. The RestNet model is used on the OCT images only, which acquired a balanced accuracy of 70.15%, specificity of 0.608, and sensitivity of 0.794. In comparison, our OCT encoder provided an accuracy of 94.64%, specificity of 0.959 and sensitivity of 0.934, which are significantly higher than the current benchmark. Moreover, in this paper, the MLP achieved a balanced accuracy of 79.87%, a specificity of 0.826, and a sensitivity of 0.771. Our model surpassed this performance again. Our MLP model has an accuracy of 84.48%, specificity of 0.896, and sensitivity of 0.808. Lastly, this study fused these two modalities through late fusion or feature-level concatenation. This strategy enhances its performance by achieving a balanced accuracy of 82.33%, a specificity of 0.742, and a sensitivity of 0.904. On the contrary, our proposed architecture with three different fusion strategies achieved a balanced accuracy of 92.57%, 94.57% and 97.47% for gated fusion, FiLM fusion and MSA fusion, respectively. We significantly surpassed this benchmark in both specificity and sensitivity. Our proposed architectures provide specificities of 0.964, 0.972 and 0.976, sensitivities of 0.887, 0.918 and 0.973 for the three fusion strategies, respectively. The results of this paper, as well as our OCTbiO architectures, reflect the success in adopting the multi-modal approach again. However, Prabhushankar's experiment was conducted on an NVIDIA GeForce GTX TITAN X with 12GB GPU memory. In contrast, we used NVIDIA Tesla P100 with 16GB GPU memory. Since these metrics are validated over separate GPU configurations, the comparison may introduce inherent biases due to memory bandwidth, clock speed and computational capability.

Following the absolute success of our OCT encoder and OCTbiO models over the

benchmark of this dataset for the classification, we compared the Dual Vision V1 and V2 architectures with the existing current best work. Although this work did not work with the fundus images, but our proposed architectures prove that incorporating fundus images with the OCT slices provides the best result for diagnosis with an accuracy of 95.52%, a specificity of 0.952 and a sensitivity of 0.957 for Dual Vision V1 architecture and an accuracy of 97.39%, a specificity of 0.988, and a sensitivity of 0.962 for Dual Vision V2 architecture.

In all aspects, our proposed model demonstrated strong performance, achieving better accuracy and sensitivity than Prabhushankar's approach, making our research the new state-of-the-art for DR vs DME classification on the OLIVES dataset.

Chapter 7

Architecture Explainability

After all numerical results and graphs, we are also showing the visual explainability of our methods. This visual explainability is based on how each architecture is made and which fusion mechanism we are choosing here. In computer vision tasks, the output with visual reasoning is superior to others because it helps the viewer better understand the decision. Here, for image visuality, we are using Grad-CAM (Gradient weighted Class Activation Mapping) explainability. This method is widely used for Explainable AI (XAI) methods [22]. So, using Grad-CAM we can visualize a heatmap or a focusing area in a certain image. So, we have to pick a layer of our CNN models and pass its information to Grad-CAM and it will output a heatmap of the image based on the weights of that certain layer of the CNN model. This method is used for layer interpretability. In medical computer vision areas, It helps the clinicians to interpret the models output performance, checking even if the model is focusing on the right pathologies.

7.1 Dual Vision V1 Architecture Explainability

In the Dual Vision V1 architecture, we have a 4-stage XAI method for whole architecture interpretation. Each stage is based on how the exact part of the architecture outputs the correct pathological patterns or not, and even it helps the viewer to understand or not. Every stage provides us with numerous rich insights that can be leveraged in the future to refine or tune our architectures for enhanced multimodal integration.

7.1.1 Stage1: Fundus Explainability

We have a fundus encoder in our Dual Vision V1 architecture, which outputs a fundus feature vector for each patient. This fundus feature vector represents the local highlighted patterns from the fundus images. So, Stage 1 is fully focused on the fundus features outputted from the fundus encoder. But, since we have a K-fold system with 5 different folds, we have different fold models, which are model_fold1.pt, model_fold2.pt, model_fold3.pt, model_fold4.pt and model_fold5.pt. So, the explainability will be an ensembled explanation because that gives a better interpretation. The last layer of our fundus encoder is Conv2D Block 5, and after that, the final convoluted features enter the next step, so this final layer is the deciding layer for the fundus encoder. In order to capture the gradient and activations we have a

forward hook and a backward hook for Conv2D Block 5 layer. The formulas are given below:

$$\alpha_c = \text{mean over } (H, W) \text{ of gradients}_c$$

where α_c=Channel Importance

$$\text{CAM} = \sum_c \alpha_c \cdot \text{activation}_c$$

where activation$_c$= Channels Activation Map

$$\text{CAM} = \text{ReLU}(\text{CAM})$$

$$\text{CAM}_{\text{norm}} = \frac{\text{CAM} - \min(\text{CAM})}{\max(\text{CAM}) - \min(\text{CAM}) + \epsilon}$$

where CAM$_{norm}$ = $Normalized\ CAM$

After completing all the calculations, we need to overlay the CAM onto the original fundus image, which then converts into a heatmap. In the case of 5 folds, we are averaging the normalized CAM values from all 5 folds for a robust explanation. Representative fundus image examples are shown below:

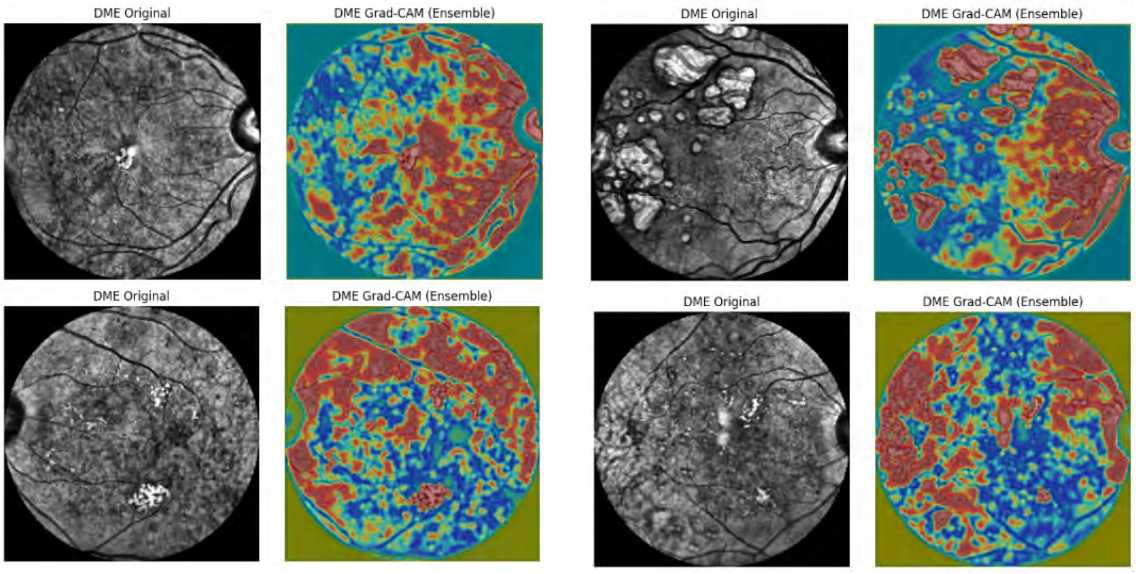

Figure 7.1: DME cases stage 1 visualization

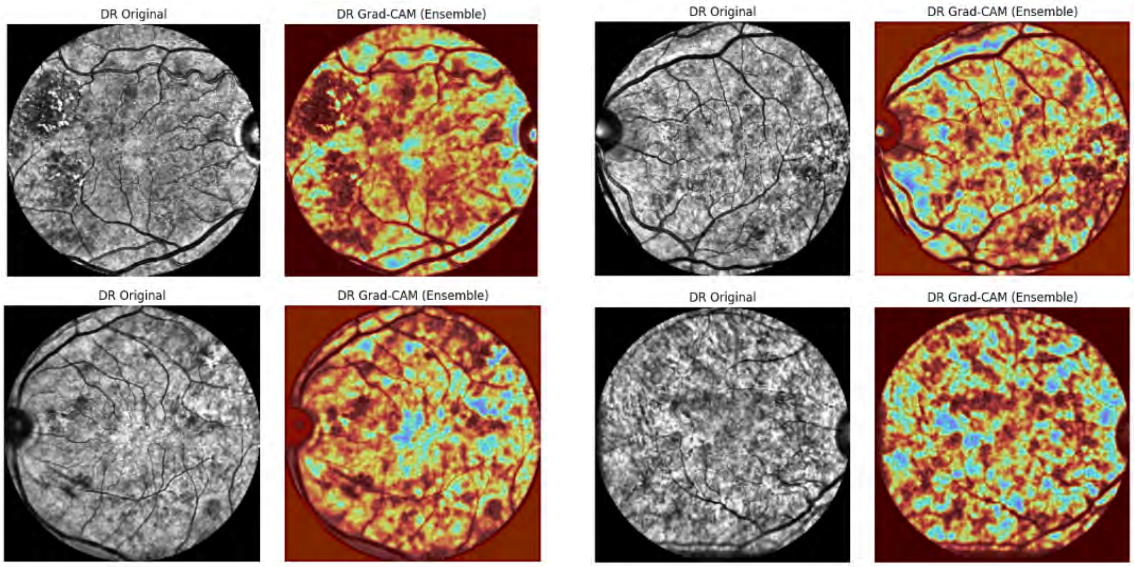

Figure 7.2: DR cases stage 1 visualization

We can see that our fundus encoder is focusing on different patterns for DME and DR cases. That's why the Grad-CAM is producing different color patterns for each class.

7.1.2 Stage2: OCT SCI Explainability

We have an OCT encoder in our Dual Vision V1 architecture, which outputs an OCT feature vector for each patient. This OCT feature vector represents the local highlighted patterns from the OCT slices. Now we have a shared OCT encoder for n OCT slices in the Dual Vision V1 architecture, so if a patient has n OCT slices for each patient, this encoder outputs n OCT slice feature vectors. So, for each OCT slice respectively first we have to perform Grad-CAM each time. To clarify, if we have 5 OCT slices we have to calculate [CAM_1, CAM_2, CAM_3,CAM_4,CAM_5]. But these are CAM values and our main motive is to calculate a value that tells us which slice is contributing how much. We have a numerical score which is SCI (Slice Contribution Index). This score ranges between [0,1] and tells us how much that specific slice contributed to the model's decision for any certain class. So, in a 5 OCT system, we have CAM_{1-5}, we need to calculate the energy of each slice by summing the CAM_i matrix values. After summing, we will get [$energy_1$, $energy_2$, $energy_3$, $energy_4$, $energy_5$]. The given formula is for CAM_i energy and SCI_i calculation.

$$\text{CAM Energy} = \sum_{i=1}^{H} \sum_{j=1}^{W} \text{CAM}[i, j]$$

$$\text{SCI}_i = \frac{\text{energy}_i}{\sum_{j=1}^{5} \text{energy}_j}$$

For each slice, we will get a SCI_i score value representing OCT slice score values [SCI_1,SCI_2,SCI_3,SCI_4,SCI_5]. Now these values are all fraction values between 0 to 1 and we can see that adding all SCI_i values will result in 1 because it is a relative comparison. Below are some SCI values examples:

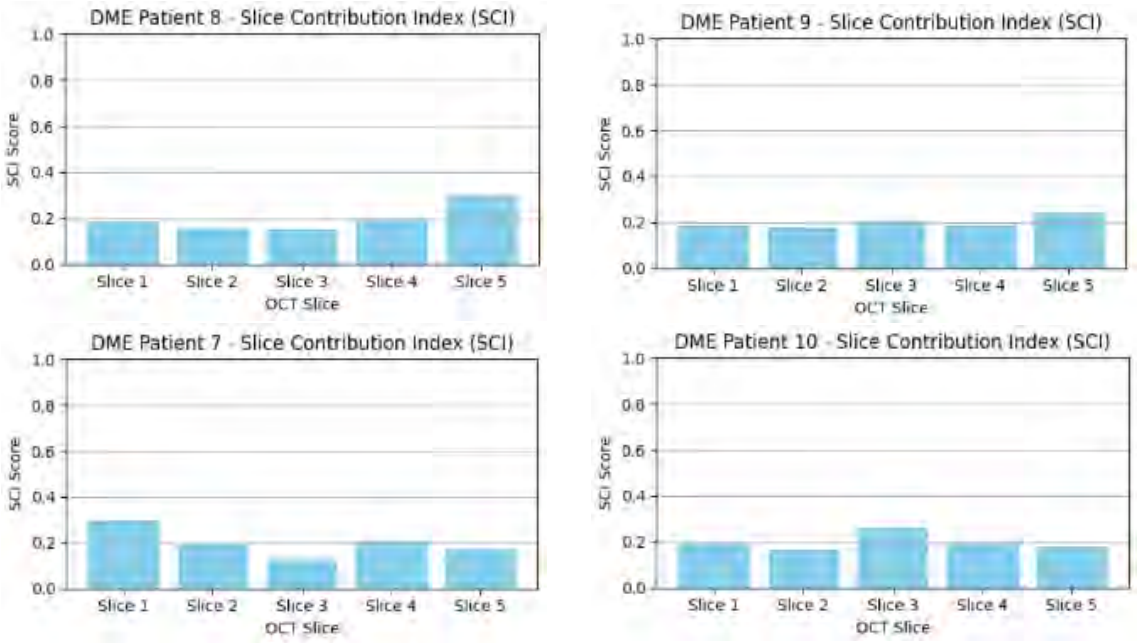

Figure 7.3: DME cases stage 2 SCI visualization

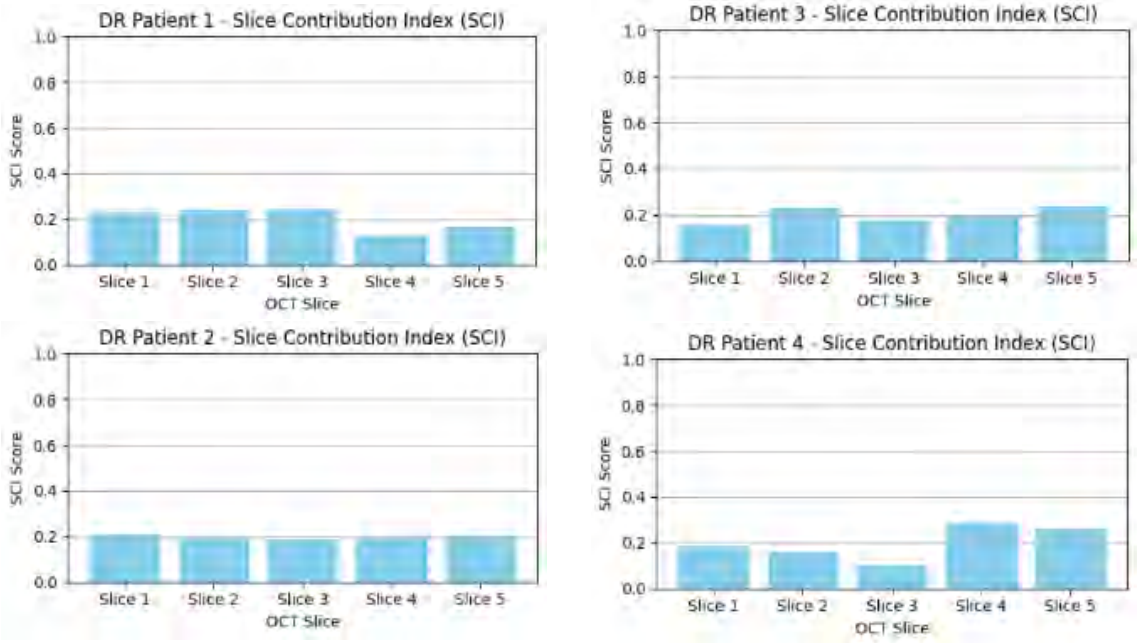

Figure 7.4: DR cases stage 2 SCI visualization

We can see that the SCI_i changes or varies between each class case. Some effective cases are:

- Case1: Only group of later slices dominates
- Case2: Only group of earlier slices dominates
- Case3: Only group of earlier with later slices dominates

- Case4: Only group of Middle slices dominates

This insight can help us in the future to pick on the slices dominance patterns and pick effective slices that matter for our models to be trained on, so that we do not have to use redundant slices that can increase training time and computation power.

7.1.3 Stage3: Attention Fused OCT Explainability

This stage is for after fusion explainability. For Dual Vision V1 architecture, we are performing an MSA fusion between n OCT slices. So, this fusion mechanism outputs a final fused vector among n OCT slices. In other words, this fused OCT feature vector has a mix of all information among n OCT slices. But we have to know among these slices how much they contributed for the final fused vector. The formulas are given below:

$$\mathbf{f}_{\text{fused}} = \text{Fuse}(\mathbf{X}) \in \mathbb{R}^{512}$$

Here, X is the stack of 5 OCT slices feature vectors. So, $\mathbf{f}_{\text{fused}}$ is the MSA fused output.

$$\mathbf{g} = \frac{\partial S_c}{\partial \mathbf{f}_{\text{fused}}}$$

$$\gamma = \text{ReLU}(\mathbf{g} \odot \mathbf{a}) \in \mathbb{R}^{512}$$

Here, g is the gradient of the $\mathbf{f}_{\text{fused}}$ vector a is the fused vector activation. Now we calculate gamma(y) which is the element wise fused vector Grad_CAM.

$$\mathbf{c} = \mathbf{W}_{fc}^{\top} \cdot \gamma \in \mathbb{R}^{128}$$

$$w_i = \mathbf{x}_i^{\top} \cdot \mathbf{c}$$

$$\hat{w}_i = \frac{w_i - \min(w)}{\max(w) - \min(w) + \epsilon}$$

We have $\mathbf{W}_{fc}^{\top}$ which is the FC layer weight matrix after the fusion and we multiply it with gamma(y) to get c which is the slice feature space of projected CAM. Now, using c value, we calculate w_i which is weight of that exact slice and then we normalize it to get $\hat{\mathbf{w}}_\mathbf{i}$. This weight is the value that tells us how much weight that specific slice has for contributing for the final OCT fused vector. Below some examples are shown:

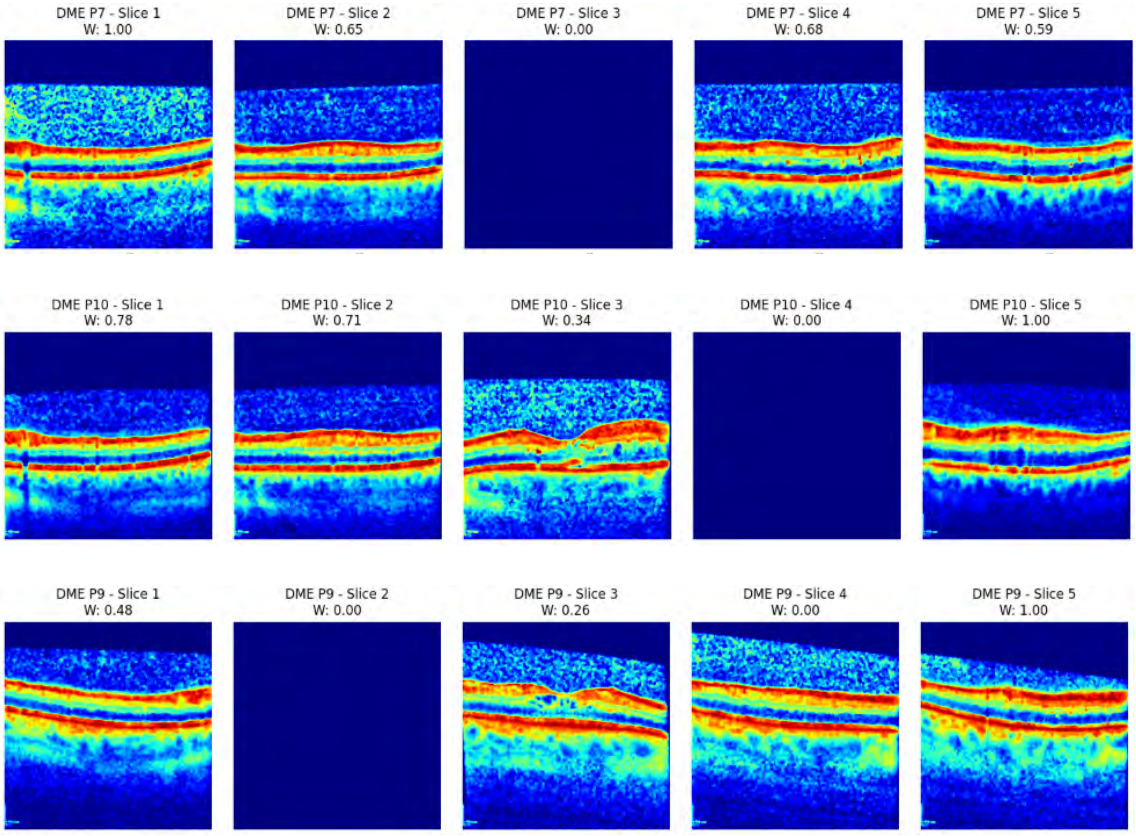

Figure 7.5: DME cases stage 3 slice weight visualization

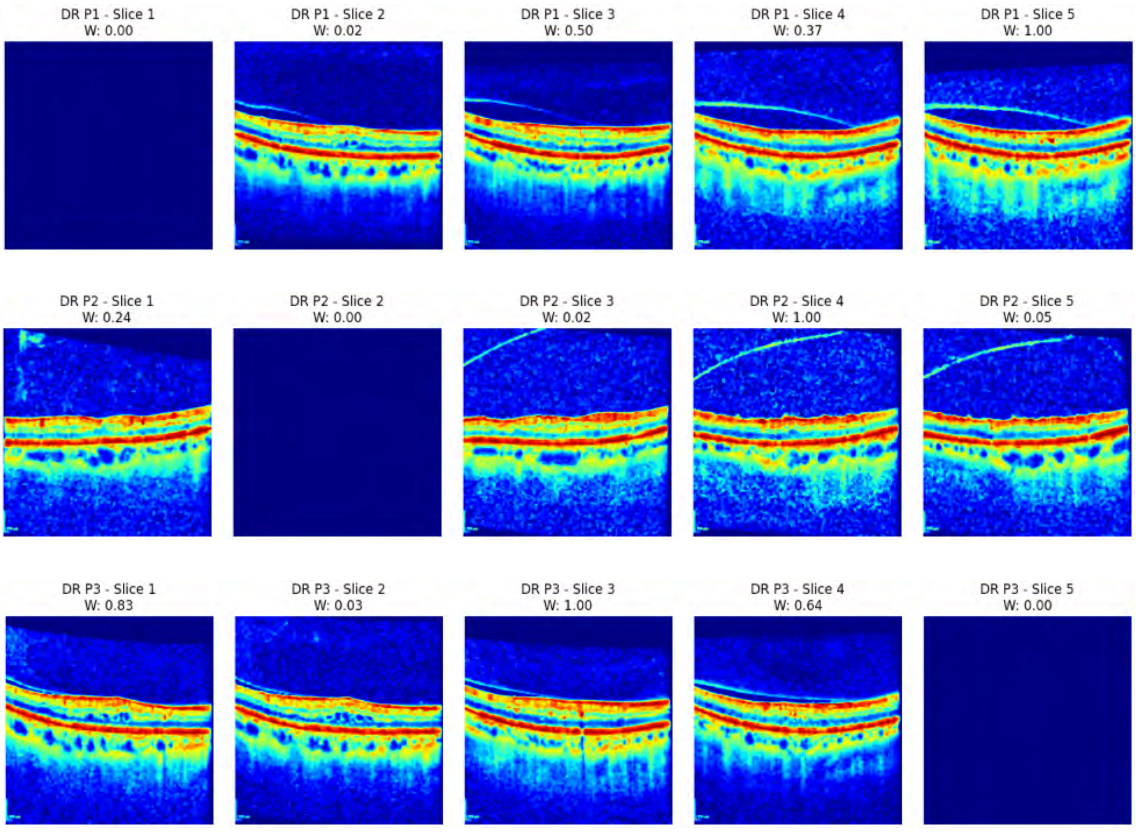

Figure 7.6: DR cases stage 3 slice weight visualization

Looking at the DME and DR cases, we can see that W means the weight value. Here, the image which has no heatmap or blank means that that specific slice has no weight so W = 0. It means that specific slice is not at all contributing for the final fused vector. This tells us that not all slices can contribute to the fusion mechanism. Some contribute mostly and some do not at all. Since MSA fusion in this method is about sharing information among slices or communication between slices, here some for all cases there are some slices which either cannot share their information or do not have any valuable to offer at all.

7.1.4 Stage4: OCT vs Fundus Gated Explainability

Now, after the fundus feature vector and the fused OCT feature vector, we are performing a gated fusion. According to this fusion, we may need to either fully trust a single modality or partially trust both modalities. So, each modality has a contribution value here, and exactly what we are showing is each modality's contribution value in this stage. We have a fundus contribution value λ_f and OCT contribution λ_o value. The formula below is how the fused gated vector is being calculated:

$$\mathbf{g} = \text{Sigmoid}\left(W \cdot [\mathbf{f}_{\text{fundus}}, \mathbf{f}_{\text{oct}}]\right) \in \mathbb{R}^{512}$$

$$\mathbf{f}_{\text{fused}} = \mathbf{g} \odot \mathbf{f}_{\text{fundus}} + (1 - \mathbf{g}) \odot \mathbf{f}_{\text{oct}}$$

Now here $\nabla \mathbf{f}_{\text{fundus}}$ is the fundus feature vector gradient and $\nabla \mathbf{f}_{\text{oct}}$ is the OCT fused fracture vector gradient, we have to calculate each modality energy value by this given formula:

$$E_{\text{fundus}} = \sum |\nabla \mathbf{f}_{\text{fundus}} \cdot \mathbf{f}_{\text{fundus}}|$$

$$E_{\text{oct}} = \sum |\nabla \mathbf{f}_{\text{oct}} \cdot \mathbf{f}_{\text{oct}}|$$

Now, as we have the energy values, we have to calculate λ_f and λ_o by the given formula:

$$\lambda_f = \frac{E_{\text{fundus}}}{E_{\text{fundus}} + E_{\text{oct}}}$$

$$\lambda_o = 1 - \lambda_f$$

This λ_f and λ_o value is a fractional weight between [0,1], meaning that together they sum to 1. If the λ_f is high, then the fundus feature vector has contributed more in predicting that class, and vice versa for λ_o. A diagram for DR and DME cases for gated fusion contribution is given below:

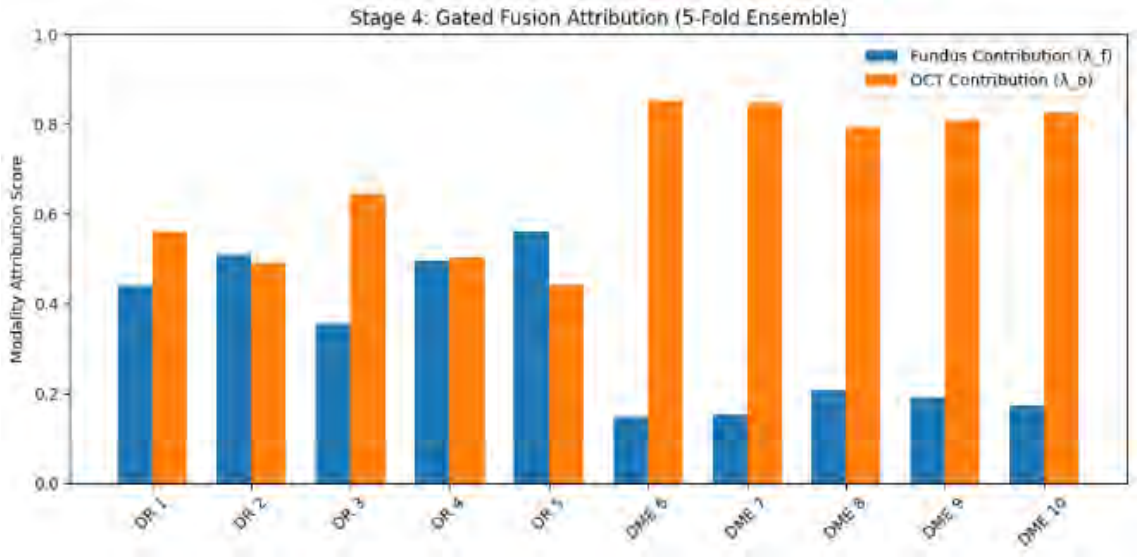

Figure 7.7: DR and DME contribution stage 4 visualization

We can see that for some DR cases fundus modality has better or equal contribution than OCT slices and on the other hand, for almost all DME cases OCT slices are fully contributing and dominant. This tells us that the dataset equality and the fusion mechanism lead us to trust the OCT modality more than the fundus image and for DR cases, both modality has equal contribution. This contribution value is a good insight for the clinicians because there will be cases when a pathological decision needs to be taken after going through three stages will be mixed, where the last stage can be the cherry on top, giving the viewer a multi-visual point of view.

7.2 Dual Vision V2 Architecture Explainability

In Dual Vision V2 architecture, we have a 2 stage XAI method for whole architecture interpretation. Each stage is based on how the exact part of the architecture outputs the correct pathological patterns or not and even it helps the viewer to understand or not. Every stage provides us with many rich insights that can be used in the future to modify or tune our architectures for much better multimodal integration. We are not showing fundus image or OCT slice explanation here because we have shown that in the Dual Vision V2 architecture stage 1,2. Again, we are performing a 5 fold ensemble explanation.

7.2.1 Stage1: Gated Fold Explainability

In Dual Vision V2 architecture, we are using a gated fold fusion mechanism for n OCT slices and this mechanism is different from normal gated fusion. We have a forward pass and backward pass for n OCT slices, meaning it is a bi-directional point of view. So, in this stage, we have a forward and backward fold explainability. In other words, we are going to explain which pass mattered the most for each class and in that particular pass, which slices contributed how much by showing SCI. For each step i, we have to calculate the gate logits:

$$\ell_t = W \cdot [h_{t-1}, x_t, \text{pos}_t] + b + a \cdot \text{pos}_t$$

Then, with the value of gate logit we have to calculate the gate activation and state update:

$$z_t = \sigma(\ell_t), \qquad z_t \in (0,1)^D$$
$$h_t = (1 - z_t) \odot h_{t-1} + z_t \odot x_t$$

Now to measure the SCI. Given are the formulas for SCI calculation normalization:

$$\bar{z}_t = \frac{1}{D} \sum_{j=1}^{D} z_{t,j}$$

$$\Delta_t = \|x_t - h_{t-1}\|_2$$
$$\mathrm{SCI}_t = \bar{z}_t \cdot \Delta_t$$
$$\widehat{\mathrm{SCI}}_t = \frac{\mathrm{SCI}_t}{\sum_{k=1}^{N} \mathrm{SCI}_k}$$

Here, $\bar{z}_t$ is the gate strength and Δ_t is the Euclidean distance. So, doing this calculation for both forward and backward pass we get $\widehat{\mathrm{SCI}}_t^{\mathrm{fwd}}$ and $\widehat{\mathrm{SCI}}_t^{\mathrm{bwd}}$. Now, for forward and backward contribution pie chart, we have their final hidden states as h_N^{fwd} and h_1^{bwd}. Now the formulas below are for the final steps:

$$u = \left[h_N^{\mathrm{fwd}},\ h_1^{\mathrm{bwd}}\right]$$

Starting with the concatenation of the final hidden states, we compute the gate and later, after merging we get the mean of gates across five folds:

$$\lambda_{\mathrm{fwd}} = \mathrm{mean}(g), \qquad \lambda_{\mathrm{bwd}} = 1 - \lambda_{\mathrm{fwd}}$$

Below are some visual examples of DR and DME:

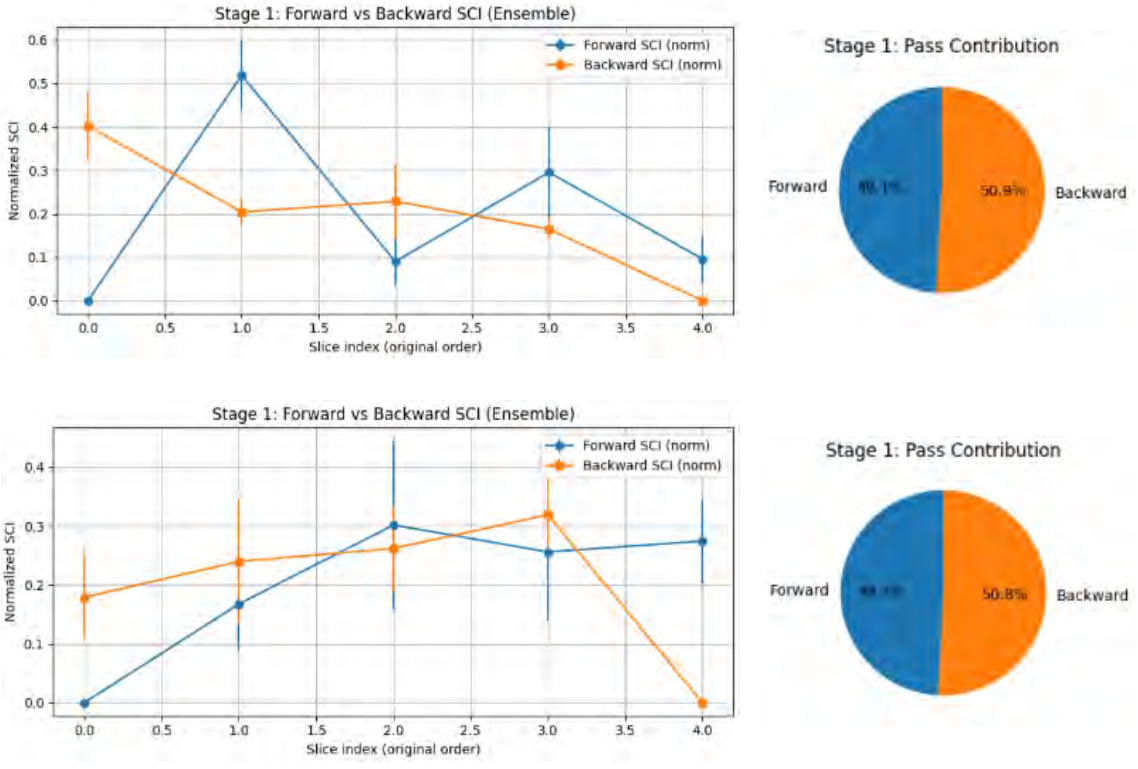

Figure 7.8: DME cases stage 1 forward vs backward pass with SCI visualization

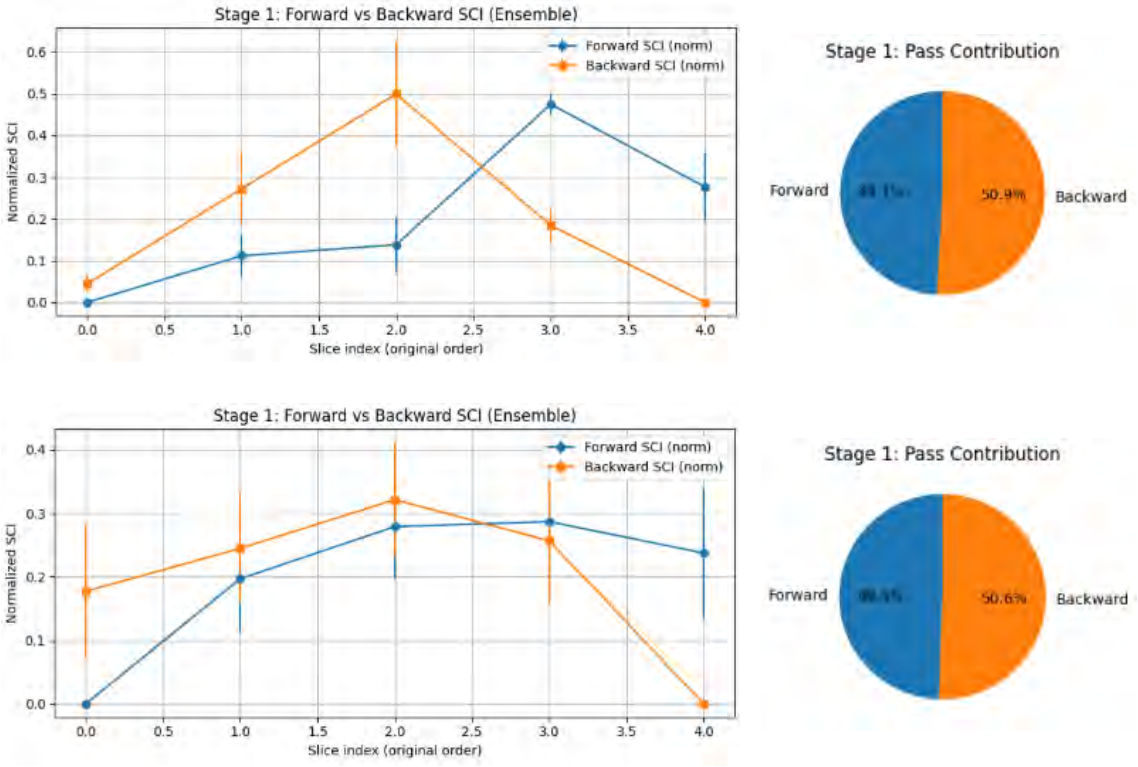

Figure 7.9: DR cases stage 1 forward vs backward pass with SCI visualization

We can see that for the majority cases the middle slices matter the most, meaning they carry much more info. But there could be other cases where earlier slice or later slices matters in a single passing fold. The overlapping bars in forward and backward slices mean that no matter which direction, that particular slice is important. Then the pie chart indicated that for almost all cases the backward pass is a little bit dominating, meaning that the backward sequence of gated fold fusion is giving a little bit more information than the forward fold pass. This tells us when to rely on which point of view scanning of multiple OCTs.

7.2.2 Stage2: OCT VS Fundus MSA Explainability

After the OCT fused vector form gated fold fusion, we are fusing the fundus feature vector with it using an MSA fusion. This fusion is for modality communication and fused contribution. But we need to see which modality has contributed how much here. For that, we have to calculate the modality contribution score for fundus modality (λ_f) and OCT modality (λ_o). For that, the given formulas are how we do it:
We have input tokens $T = [f; o] \in \mathbb{R}^{2\times D}$ then we have to perform an attention calculation using the softmax function based on Q, K and V:

$$\text{Attn}(Q, K, V) = \text{softmax}\left(\frac{QK^\top}{\sqrt{d_k}}\right)V$$

Then, using this attention value, we have to calculate contribution weights and normalize them:

$$c_j = \sum_{i=1}^{2} A_{i,j}, \quad j \in \{0, 1\}$$

$$\lambda_f = \frac{c_0}{c_0 + c_1}, \qquad \lambda_o = \frac{c_1}{c_0 + c_1}$$

As we have the contribution weight, now we can plot them based on the 5 fold ensemble system. A DR and DME attribution chart is given below:

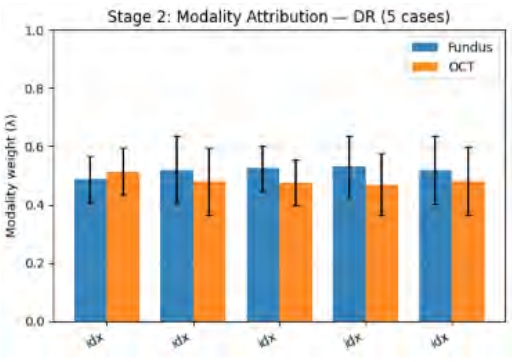

Figure 7.10: DR cases stage 2

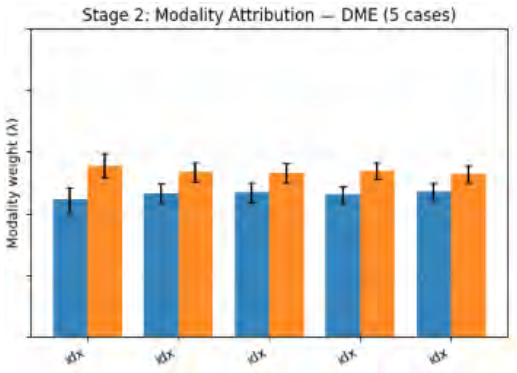

Figure 7.11: DME cases stage 2

Here, we can see that for DR cases, the fundus images contribute less or share less information. Where, for DME cases is vice versa. The Black bars are the variation among 5 folds, which means how much the folds disagree. So, we find a pattern of contribution for each class which is a heart clinical interpretation for the experts to identify the reasons for that certain class predictions.

7.3 OCTbiO Architecture Explainability

In this architecture, we are trying three different types of fusion. Among them, we are going to explain only the FiLM Fusion part since gated and MSA explainability are pretty much the same as the Dual Vision architectures. Talking about FilM, we are using the biomarker tabular features to modulate the OCT encoder layers. This is a type of mid-level fusion where the OCT encoder's performance gets modulated based on the biomarker densely embedded tabular features. So, here we have two stage explainability based on the architecture's workflow.

7.3.1 Stage1: Integrated Gradient Explainability

In stage 1, we have calculated IG (Integrated Gradient) for each biomarker, tabular and clinical features of each patient. This IG helps clinicians understand which biomarker tabular values have contributed to the model prediction. In other words, IG indicates each trait's contribution to the model's progression from a baseline patient to the current patient. Since we have a K-fold system with 5 different folds, we have different fold models, which are model_fold1.pt, model_fold2.pt, model_fold3.pt, model_fold4.pt and model_fold5.pt. So, the IG explainability will be an ensembled explanation because that gives a better interpretation. Below, the IG calculation flow is given:

First, we have to pick a reference input as $\mathbf{x}_0 = 0$ or $\mathbf{x}_0 = \mu$, which is the zero

baseline or mean baseline, respectively. Then we have to gradually slide from baseline to input and IG by these formulas:

$$x^{(\alpha)} = x_0 + \alpha(x - x_0), \quad \alpha \in [0, 1].$$

$$IG_i(x) = (x_i - x_{0,i}) \cdot \int_0^1 \frac{\partial y(x^{(\alpha)})}{\partial x_i} \, d\alpha$$

Here, $(x_i - x_{0,i})\cdot$ is the change in the exact feature and with respect to that exact feature, we have to multiply it by its average gradient. Now, since we have a 5 fold system where each model is trained on that particular fold, we have to ensemble average among 5 folds and normalize for plotting. The formula is given below:

$$IG_i^{(m)}(x), \quad m = 1, \ldots, 5.$$

$$IG_{i,\text{mean}}(x) = \frac{1}{5} \sum_{m=1}^{5} IG_i^{(m)}(x),$$

$$IG_{i,\text{std}}(x) = \text{StdDev}\left(IG_i^{(1)}, \ldots, IG_i^{(5)}\right).$$

$$IG_{i,\text{disp}} = \frac{IG_{i,\text{mean}}}{\sum_j |IG_{j,\text{mean}}|}$$

After all calculations, we will get all IG values related to the Mean and Std (Standard Deviation). Below we have some DR and DME examples for the IG Explanability.

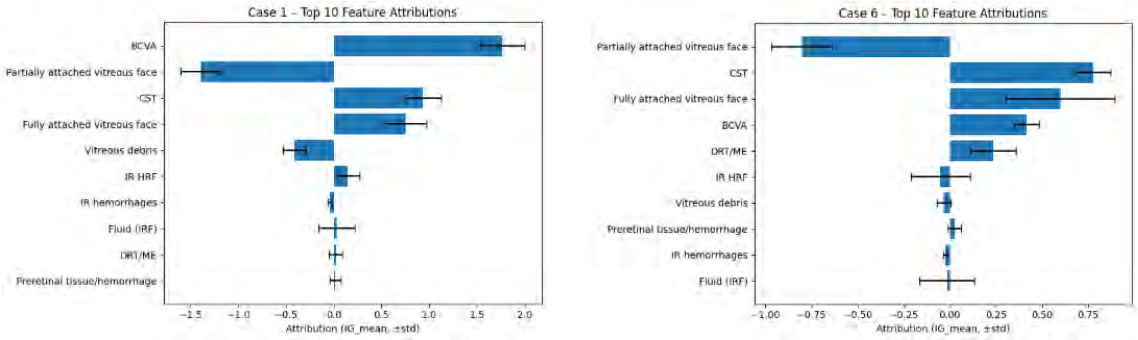

Figure 7.12: DME cases stage 1 IG visualization

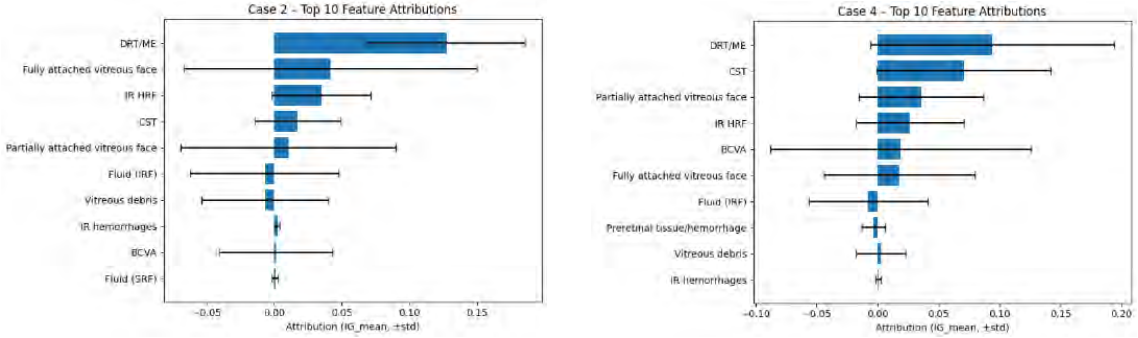

Figure 7.13: DR cases stage 1 IG visualization

Here, by looking at the IG Values, we can see that for both DR and DME cases BCVA, CST, DRT/ME, Fully attached vitreous face, Partially attached vitreous face features have either positive impact or negative impact on the models decision. The x-axis values are the attribution (IG_mean, $\pm std$). Here, blue bars are the IG_mean and the black lines are IG_std. The black bars (IG_std) mean folds agreement or disagreement for that particular feature. This helps the clinicals to understand for which class cases the relevant feature is important or not.

7.3.2 Stage2: Channel Influence Explanibility

Stage 2 focuses on the channel influence explanation. As we know that the FiLM fusion is responsible for channel modulations, we are using it in our Conv2D Block 3 and Conv2D Block 4. For each layer, there are multiple channels and these channels are being modulated with the help of FiLM fusion by the tabular biomarker embedded vectors. First, we have to calculate the channel importance score.

$$\frac{\partial y_c}{\partial A_{ij}^k}$$

$$\alpha_k = \frac{1}{H \times W} \sum_{i,j} \frac{\partial y_c}{\partial A_{ij}^k}$$

First, we have to calculate the channel gradient and average it spatially. Here, $\frac{\partial y_c}{\partial A_{ij}^k}$ is the gradient of that channel and α_k is the channel k's Grad-CAM weight. Then we have to perform a FilM modulation by using this given formula:

$$A^k \mapsto (1 + \gamma_k)A^k + \beta_k$$

- γ_k = multiplicative scaling (boost or suppress the channel)
- β_k = additive shift

$$s_k = \alpha_k \cdot (1 + \gamma_k)$$

Lastly, we get s_k which is the final score for that particular channel k after considering the importance of that channel and how FiLM fusion has boosted it or suppressed it. Some cases are given below for visual explainability:

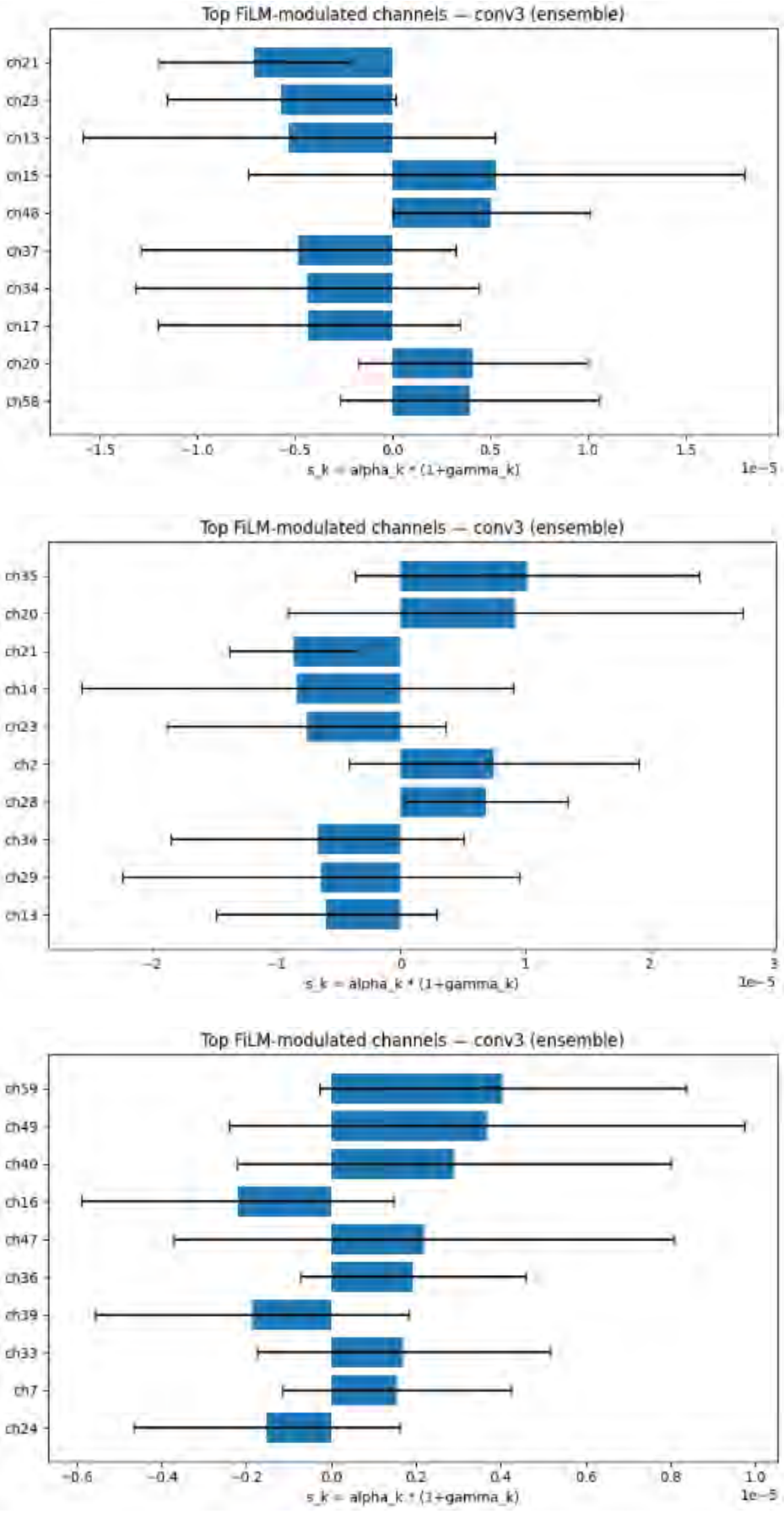

Figure 7.14: Stage 2 Conv2D Block 3 channel modulation visualization

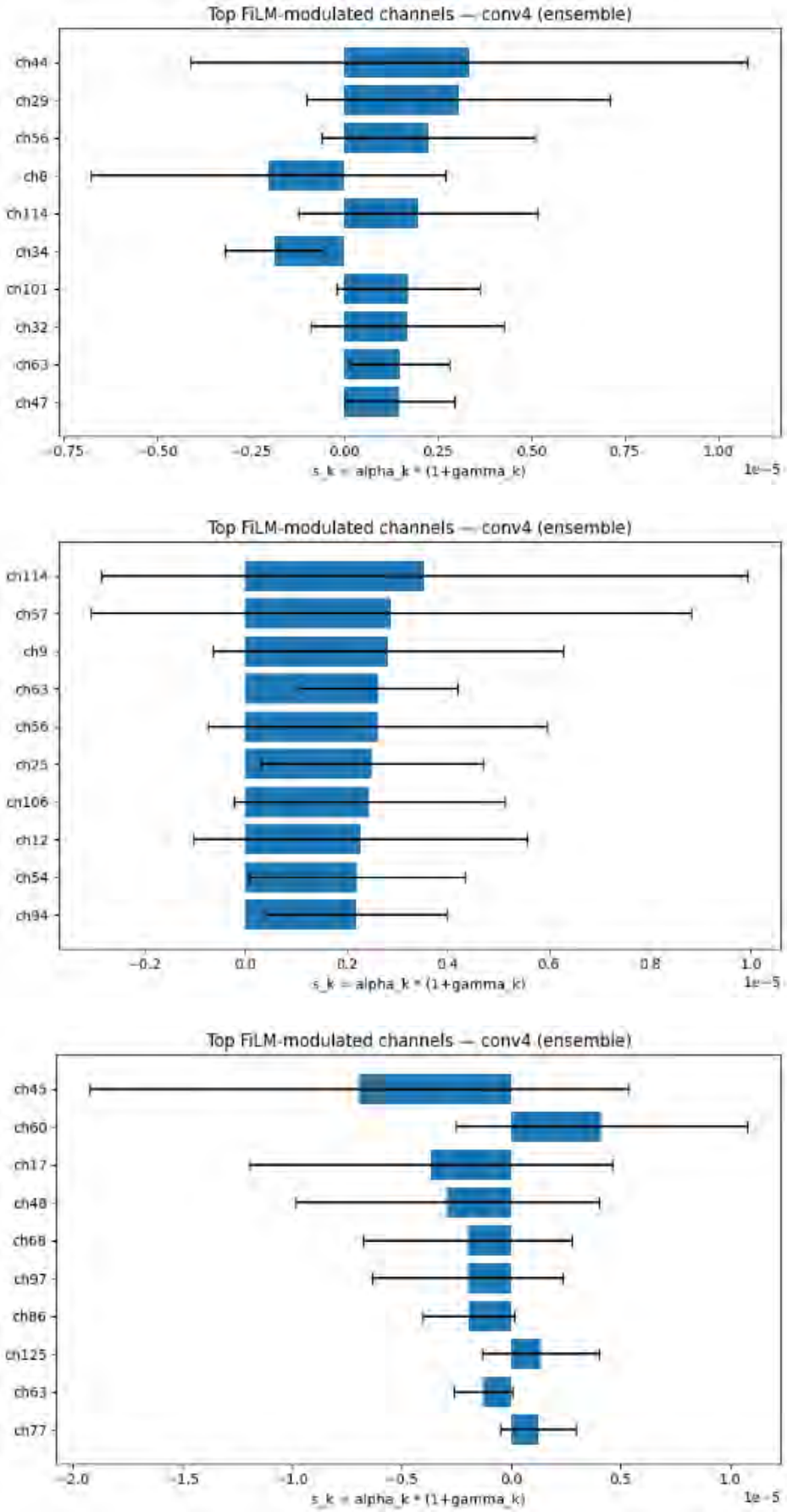

Figure 7.15: Stage 2 Conv2D Block 4 channel modulation visualization

Here, we can see that the channels are either boosted or suppressed by the tabular features in Conv2D Block3 and Conv2D Block4. The number of channels is greater in Conv2D Block4 than in Conv2D Block3. This figure also tells us that certain tabular features can drift, which channels, and it does not matter if the drift is negative, meaning suppressed, or positive, meaning boosted, because for a model to generalize better, both are needed. Here, in the x-axis we have, we have the final score s_k. The left side of 0 on the x-axis indicates that the specific channel is contributing to the prediction of the class, and the right side of 0 means that the specific channel is suppressed.

Chapter 8

Conclusion

8.1 Summary of Findings

This research is based on multimodal DR and DME detection, utilizing image data and biomarker data that interact with several pre-processing techniques, incorporating multiple layers of CNN models, and fusing the results from several pipelines. The previous studies commonly focused only on retinal fundus images or only on OCT slices. These methods are proven insufficient for patients with a variety of conditions and other medical complications. From Table 6.9, it is noticeable that introducing multi-modality with various fusion mechanisms increases this research's test accuracy rate for the classification of DR and DME. Additionally, this method is affordable, reliable, robust, and feasible.

8.2 Contributions to the Field

Proper classification of ocular diseases enables doctors to provide their patients with adequate treatment. At present, there has been minimal research to distinguish between DR and DME, as these two diseases overlap. So we aim to contribute to this medical field with the benchmark accuracy for class identification. The blindness rate will decrease with this approach because our methodology not only classifies the ocular condition but also provides a reason for that exact classification. This research would be helpful for doctors to distinguish between mild and severe DR patients. Since we are using more than one modality, determining which modality is most important in each case provides valuable information. It helps clinicians know when to look into which modality and how to investigate each modality. Our research provides a visual explanation for clinicians to look into the insights of the mixed modalities. Also, our research enhances diagnostic accuracy, makes DR and DME detection more affordable, and establishes a solid foundation for future studies. In short, this study uses a multimodal strategy for classification and provides visual explanations.

8.3 Limitations in Our Research

The primary obstacle in multi-modal medical research is the limited number of datasets. Each sample must contain data for every modality to establish a significant relationship between two or more modalities. If a patient only has an OCT or

fundus image, we can use that single piece of data for classifying their eye condition. Meanwhile, our research covers a variety of modalities, which is why we need both OCT and fundus images for every patient. Unfortunately, such multi-modal ocular data are not widely available. Therefore, it becomes challenging for models to train smoothly on these unorganized datasets.

We discovered the OLIVES dataset during our research, which contains a significant number of images with patient occurrences that overlap across modalities. But the OCT images contain a particular amount of noise and disruption. Thus, randomly choosing OCT slices does not allow the model to practically prove our alternate hypothesis that a larger number of OCT slices leads to better accuracy.

Additionally, we fused the OCT modality with the biomarkers and successfully surpassed the benchmark of the OLIVES dataset. However, the biomarkers are annotated data created by the annotators who used OCT slices. So, fusing these two modalities has a limitation in real-life applicability.

8.4 Recommendations for Future Work

In order to obtain more accurate outcomes, our research utilizes the resources of the well-known public dataset known as OLIVES, which offers real-world data. In the future, if a hospital is comfortable sharing their ocular database with additional modalities publicly, researchers can use this knowledge to improve these research outcomes.

This research's approach is affordable, but we can further improve its usability and affordability. Researchers may explore and treat other eye conditions in the near future using the same methodology. But it is necessary to ensure that the classes are balanced for every single eye condition when we get additional data.

We primarily classify DME and DR in our current research study. But there are a lot of additional conditions related to the eye, including RVO, AMD, Glaucoma, CNV, and CSC. Furthermore, future research can include disease progression analysis or classification of disease severity level with proper labeled dataset.

In the future, this research can contribute to the software industry. Technology-based companies can create telemedicine-based software. This will reduce the need for patients to travel to specialized hospitals, thereby lowering healthcare expenses and indirect charges such as transportation. An additional feature can be added to receive clinical decision feedback and integrate the system with the overall clinical workflow.

The overall outcome of our study opens up many opportunities for other researchers and provides valuable information for those interested in investigating multimodal fusion-based approaches.

Bibliography

[1] S. M. Pizer et al., "Adaptive histogram equalization and its variations," *Computer Vision, Graphics, and Image Processing*, vol. 39, no. 3, pp. 355–368, 1987, ISSN: 0734-189X. DOI: 10.1016/S0734-189X(87)80186-X. eprint: https://doi.org/10.1016/S0734-189X(87)80186-X. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0734189X8780186X.

[2] J. Lowell et al., "Optic nerve head segmentation," *IEEE Transactions on Medical Imaging*, vol. 23, no. 2, pp. 256–264, 2004. DOI: 10.1109/TMI.2003.823261. eprint: https://doi.org/10.1109/TMI.2003.823261. [Online]. Available: https://ieeexplore.ieee.org/document/1263614.

[3] J. Staal, M. D. Abramoff, M. Niemeijer, M. A. Viergever, and B. van Ginneken, "Ridge-based vessel segmentation in color images of the retina," *IEEE Transactions on Medical Imaging*, vol. 23, no. 4, pp. 501–509, 2004. DOI: 10.1109/TMI.2004.825627. eprint: https://doi.org/10.1109/TMI.2004.825627. [Online]. Available: https://ieeexplore.ieee.org/document/1282003.

[4] M. D. Abramoff, M. Niemeijer, and S. R. Russell, "Automated detection of diabetic retinopathy: Barriers to translation into clinical practice," *Expert Review of Medical Devices*, vol. 7, no. 2, pp. 287–296, Mar. 2010. DOI: 10.1586/erd.09.76. eprint: https://pubmed.ncbi.nlm.nih.gov/20214432/. [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/20214432/.

[5] R. Kafieh, H. Rabbani, M. D. Abramoff, and M. Sonka, "Intra-retinal layer segmentation of 3d optical coherence tomography using coarse grained diffusion map," *Medical Image Analysis*, vol. 17, no. 8, pp. 907–928, Dec. 2013, ISSN: 1361-8415. DOI: 10.1016/j.media.2013.05.006. eprint: https://doi.org/10.1016/j.media.2013.05.006. [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/23837966/.

[6] X. Chen, Y. Xu, D. Wong, T.-Y. Wong, and J. Liu, "Glaucoma detection based on deep convolutional neural network," in *Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, Aug. 2015, pp. 715–718. DOI: 10.1109/EMBC.2015.7318462. eprint: https://doi.org/10.1109/EMBC.2015.7318462. [Online]. Available: https://doi.org/10.1109/EMBC.2015.7318462.

[7] M. Esmaeili, A. M. Dehnavi, H. Rabbani, and F. Hajizadeh, "Three-dimensional segmentation of retinal cysts from spectral-domain optical coherence tomography images by the use of three-dimensional curvelet based k-svd," *Journal of Medical Signals & Sensors*, vol. 6, no. 3, pp. 166–171, Jul. 2016. eprint: https://journals.lww.com/jmss/fulltext/2016/06030/Three_dimensional_Segmentation_of_Retinal_Cysts.6.aspx. [Online]. Available: https://journals.

lww.com/jmss/fulltext/2016/06030/Three_dimensional_Segmentation_of_Retinal_Cysts.6.aspx.

[8] H. Fu, Y. Xu, S. Lin, D. W. Kee Wong, and J. Liu, "Deepvessel: Retinal vessel segmentation via deep learning and conditional random field," in *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2016*, S. Ourselin, L. Joskowicz, M. R. Sabuncu, G. Unal, and W. Wells, Eds., ser. Lecture Notes in Computer Science, vol. 9901, Cham: Springer International Publishing, 2016, pp. 132–139, ISBN: 978-3-319-46723-8. DOI: 10.1007/978-3-319-46723-8_16. eprint: https://doi.org/10.1007/978-3-319-46723-8_16. [Online]. Available: https://doi.org/10.1007/978-3-319-46723-8_16.

[9] H. Pratt, F. Coenen, D. M. Broadbent, S. P. Harding, and Y. Zheng, "Convolutional neural networks for diabetic retinopathy," *Procedia Computer Science*, vol. 90, pp. 200–205, 2016, 20th Conference on Medical Image Understanding and Analysis (MIUA 2016), ISSN: 1877-0509. DOI: 10.1016/j.procs.2016.07.014. eprint: https://doi.org/10.1016/j.procs.2016.07.014. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1877050916311929.

[10] S. Rahman, M. M. Rahman, M. Abdullah-Al-Wadud, G. D. Al-Quaderi, and M. Shoyaib, "An adaptive gamma correction for image enhancement," *EURASIP Journal on Image and Video Processing*, vol. 2016, no. 1, p. 35, Oct. 2016, ISSN: 1687-5281. DOI: 10.1186/s13640-016-0138-1. eprint: https://doi.org/10.1186/s13640-016-0138-1. [Online]. Available: https://doi.org/10.1186/s13640-016-0138-1.

[11] C. S. H. Tan, M. C. Y. Chew, L. W. Y. Lim, and S. R. Sadda, "Advances in retinal imaging for diabetic retinopathy and diabetic macular edema," *Indian Journal of Ophthalmology*, vol. 64, no. 1, pp. 76–83, 2016, ISSN: 1998-3689. DOI: 10.4103/0301-4738.178145. [Online]. Available: https://doi.org/10.4103/0301-4738.178145.

[12] S. R. Flaxman et al., "Global causes of blindness and distance vision impairment 1990-2020: A systematic review and meta-analysis," *The Lancet Global Health*, vol. 5, no. 12, e1221–e1234, 2017, Published by Elsevier Ltd. under CC BY 4.0 license., ISSN: 2214-109X. DOI: 10.1016/S2214-109X(17)30393-5. [Online]. Available: https://doi.org/10.1016/S2214-109X(17)30393-5.

[13] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2017. arXiv: 1412.6980 [cs.LG]. [Online]. Available: https://arxiv.org/abs/1412.6980.

[14] A. Vaswani et al., "Attention is all you need," *arXiv preprint arXiv:1706.03762*, 2017. arXiv: 1706.03762 [cs.CL]. [Online]. Available: https://arxiv.org/abs/1706.03762.

[15] S. K. Pandey and V. Sharma, "World diabetes day 2018: Battling the emerging epidemic of diabetic retinopathy," *Indian Journal of Ophthalmology*, vol. 66, no. 11, pp. 1652–1653, Nov. 2018, ISSN: 0301-4738. DOI: 10.4103/ijo.IJO_1681_18. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6213704/.

[16] M. Shakandli, "Deep learning for retinal image analysis," PhD Thesis, University of Leeds, Jan. 2018. [Online]. Available: https://etheses.whiterose.ac.uk/id/eprint/19306/1/M%5C%20Shakandli%5C%20January%5C%202018%5C%20Thesis.pdf.

[17] F. G. Venhuizen et al., "Deep learning approach for the detection and quantification of intraretinal cystoid fluid in multivendor optical coherence tomography," *Biomedical Optics Express*, vol. 9, no. 4, pp. 1545–1569, Apr. 2018. DOI: 10.1364/BOE.9.001545. eprint: https://doi.org/10.1364/BOE.9.001545. [Online]. Available: https://opg.optica.org/boe/abstract.cfm?URI=boe-9-4-1545.

[18] X. Zhang, H. Dong, Z. Hu, W.-S. Lai, F. Wang, and M.-H. Yang, "Gated fusion network for joint image deblurring and super-resolution," *arXiv preprint arXiv:1807.10806*, 2018. arXiv: 1807.10806 [cs.CV]. [Online]. Available: https://arxiv.org/abs/1807.10806.

[19] G. An et al., "Glaucoma diagnosis with machine learning based on optical coherence tomography and color fundus images," *Journal of Healthcare Engineering*, vol. 2019, no. 1, p. 4 061 313, 2019. DOI: 10.1155/2019/4061313. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1155/2019/4061313. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1155/2019/4061313.

[20] J. Hu, L. Shen, S. Albanie, G. Sun, and E. Wu, "Squeeze-and-excitation networks," *arXiv preprint arXiv:1709.01507*, 2019. arXiv: 1709.01507 [cs.CV]. [Online]. Available: https://arxiv.org/abs/1709.01507.

[21] J. Sahlsten et al., "Deep learning fundus image analysis for diabetic retinopathy and macular edema grading," *Scientific Reports*, vol. 9, no. 1, p. 10 750, 2019, ISSN: 2045-2322. DOI: 10.1038/s41598-019-47181-w.

[22] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-cam: Visual explanations from deep networks via gradient-based localization," *International Journal of Computer Vision*, vol. 128, no. 2, pp. 336–359, Oct. 2019, ISSN: 1573-1405. DOI: 10.1007/s11263-019-01228-7. [Online]. Available: https://doi.org/10.1007/s11263-019-01228-7.

[23] J. D. Bodapati et al., "Blended multi-modal deep convnet features for diabetic retinopathy severity prediction," *Electronics*, vol. 9, no. 6, p. 914, May 2020, ISSN: 2079-9292. DOI: 10.3390/electronics9060914. [Online]. Available: http://dx.doi.org/10.3390/electronics9060914.

[24] A. Gaudio, A. Smailagic, and A. Campilho, "Enhancement of retinal fundus images via pixel color amplification," in *Image Analysis and Recognition*. Springer International Publishing, 2020, pp. 299–312, ISBN: 978-3-030-50516-5. DOI: 10.1007/978-3-030-50516-5_26. eprint: https://doi.org/10.1007/978-3-030-50516-5_26. [Online]. Available: https://doi.org/10.1007/978-3-030-50516-5_26.

[25] J. C. M. dos Santos, G. A. Carrijo, C. de Fátima dos Santos Cardoso, J. C. Ferreira, P. M. Sousa, and A. C. Patrocínio, "Fundus image quality enhancement for blood vessel detection via a neural network using clahe and wiener filter," *Research on Biomedical Engineering*, vol. 36, no. 2, pp. 107–119, Jun. 2020, ISSN: 2446-4740. DOI: 10.1007/s42600-020-00046-y. eprint: https://doi.org/10.1007/s42600-020-00046-y. [Online]. Available: https://doi.org/10.1007/s42600-020-00046-y.

[26] W. L. Alyoubi, M. F. Abulkhair, and W. M. Shalash, "Diabetic retinopathy fundus image classification and lesions localization system using deep learning," *Sensors*, vol. 21, no. 11, 2021, ISSN: 1424-8220. DOI: 10.3390/s21113704. [Online]. Available: https://www.mdpi.com/1424-8220/21/11/3704.

[27] M. T. Al-Antary and Y. Arafa, "Multi-scale attention network for diabetic retinopathy classification," *IEEE Access*, vol. 9, pp. 54 190–54 200, 2021. DOI: 10.1109/ACCESS.2021.3070685.

[28] T. Aziz, A. E. Ilesanmi, and C. Charoenlarpnopparut, "Efficient and accurate hemorrhages detection in retinal fundus images using smart window features," *Applied Sciences*, vol. 11, no. 14, p. 6391, 2021, ISSN: 2076-3417. DOI: 10.3390/app11146391. eprint: https://doi.org/10.3390/app11146391. [Online]. Available: https://www.mdpi.com/2076-3417/11/14/6391.

[29] P.-N. Chen, C.-C. Lee, C.-M. Liang, S.-I. Pao, K.-H. Huang, and K.-F. Lin, "General deep learning model for detecting diabetic retinopathy," *BMC Bioinformatics*, vol. 22, no. 5, p. 84, Nov. 2021, ISSN: 1471-2105. DOI: 10.1186/s12859-021-04005-x. [Online]. Available: https://doi.org/10.1186/s12859-021-04005-x.

[30] J.-B. Cordonnier, A. Loukas, and M. Jaggi, "Multi-head attention: Collaborate instead of concatenate," *arXiv preprint arXiv:2006.16362*, 2021. arXiv: 2006.16362 [cs.LG]. [Online]. Available: https://arxiv.org/abs/2006.16362.

[31] A. Erciyas and N. Barışçı, "An effective method for detecting and classifying diabetic retinopathy lesions based on deep learning," *Computational and Mathematical Methods in Medicine*, vol. 2021, p. 9 928 899, May 2021, ISSN: 1748-670X. DOI: 10.1155/2021/9928899. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8184323/.

[32] W. He et al., "Incremental learning for exudate and hemorrhage segmentation on fundus images," *Information Fusion*, vol. 73, pp. 157–164, 2021, ISSN: 1566-2535. DOI: 10.1016/j.inffus.2021.02.017. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1566253521000427.

[33] E. Y.-C. Kang et al., "A multimodal imaging–based deep learning model for detecting treatment-requiring retinal vascular diseases: Model development and validation study," *JMIR Med Inform*, vol. 9, no. 5, e28868, May 2021, ISSN: 2291-9694. DOI: 10.2196/28868. [Online]. Available: https://doi.org/10.2196/28868.

[34] K. Oh, H. M. Kang, D. Leem, H. Lee, K. Y. Seo, and S. Yoon, "Early detection of diabetic retinopathy based on deep learning and ultra-wide-field fundus images," *Scientific Reports*, vol. 11, no. 1, p. 1897, 2021, ISSN: 2045-2322. DOI: 10.1038/s41598-021-81539-3.

[35] S. H. Abbood, H. N. A. Hamed, M. S. M. Rahim, A. Rehman, T. Saba, and S. A. Bahaj, "Hybrid retinal image enhancement algorithm for diabetic retinopathy diagnostic using deep learning model," *IEEE Access*, vol. 10, pp. 73 079–73 086, 2022. DOI: 10.1109/ACCESS.2022.3189374.

[36] S. Ghouali et al., "Artificial intelligence-based teleopthalmology application for diagnosis of diabetics retinopathy," *IEEE Open Journal of Engineering in Medicine and Biology*, vol. 3, pp. 124–133, 2022. DOI: 10.1109/OJEMB.2022.3192780.

[37] Á. S. Hervella, J. Rouco, J. Novo, and M. Ortega, "Retinal microaneurysms detection using adversarial pre-training with unlabeled multimodal images," *Information Fusion*, vol. 79, pp. 146–161, 2022, ISSN: 1566-2535. DOI: 10.1016/j.inffus.2021.10.003. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1566253521002013.

[38] B.-A. Marginean, A. Groza, G. Muntean, and S. D. Nicoara, "Predicting visual acuity in patients treated for amd," *Diagnostics (Basel)*, vol. 12, no. 6, p. 1504, Jun. 2022, ISSN: 2075-4418. DOI: 10.3390/diagnostics12061504. eprint: https://doi.org/10.3390/diagnostics12061504. [Online]. Available: https://doi.org/10.3390/diagnostics12061504.

[39] M. Prabhushankar, K. Kokilepersaud, Y.-y. Logan, S. Corona, G. Alregib, and C. Wykoff, "Olives dataset: Ophthalmic labels for investigating visual eye semantics," *arXiv preprint arXiv:2209.11195*, Sep. 2022. DOI: 10.48550/arXiv.2209.11195. eprint: https://doi.org/10.48550/arXiv.2209.11195. [Online]. Available: https://doi.org/10.48550/arXiv.2209.11195.

[40] G. A. Saleh et al., "The role of medical image modalities and ai in the early detection, diagnosis and grading of retinal diseases: A survey," *Bioengineering*, vol. 9, no. 8, 2022, ISSN: 2306-5354. DOI: 10.3390/bioengineering9080366. [Online]. Available: https://www.mdpi.com/2306-5354/9/8/366.

[41] P. Zang et al., "A diabetic retinopathy classification framework based on deep-learning analysis of oct angiography," *Translational Vision Science Technology*, vol. 11, no. 7, pp. 10–10, Jul. 2022, ISSN: 2164-2591. DOI: 10.1167/tvst.11.7.10. eprint: https://arvojournals.org/arvo/content_public/journal/tvst/938598/i2164-2591-11-7-10_1657608897.74855.pdf. [Online]. Available: https://doi.org/10.1167/tvst.11.7.10.

[42] M. Ahdi, K. Sjamsuri, A. Kunaefi, B. Nugroho, and A. Yusuf, "Convolutional neural network (cnn) efficientnet-b0 model architecture for paddy diseases classification," in *Proceedings of the 2023 International Conference on Information and Communication Technology and Systems (ICTS)*, Oct. 2023, pp. 105–110. DOI: 10.1109/ICTS58770.2023.10330828. [Online]. Available: https://doi.org/10.1109/ICTS58770.2023.10330828.

[43] Y. Attiku et al., "Comparison of diabetic retinopathy severity grading on etdrs 7-field versus ultrawide-field assessment," *Eye*, vol. 37, no. 14, pp. 2946–2949, Oct. 2023, ISSN: 1476-5454. DOI: 10.1038/s41433-023-02445-8. [Online]. Available: https://doi.org/10.1038/s41433-023-02445-8.

[44] M. El Habib Daho et al., "Improved automatic diabetic retinopathy severity classification using deep multimodal fusion of uwf-cfp and octa images," in *Ophthalmic Medical Image Analysis*. Springer Nature Switzerland, 2023, pp. 11–20, ISBN: 9783031440137. DOI: 10.1007/978-3-031-44013-7_2. [Online]. Available: http://dx.doi.org/10.1007/978-3-031-44013-7_2.

[45] T. Hassan, Z. Li, M. U. Akram, I. Hussain, K. Khalaf, and N. Werghi, "Angular contrastive distillation driven self-supervised scanner independent screening and grading of retinopathy," *Information Fusion*, vol. 92, pp. 404–419, 2023, ISSN: 1566-2535. DOI: 10.1016/j.inffus.2022.12.006. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1566253522002548.

[46] B. Ma, K. Wang, and Y. Hu, "Finger vein recognition based on bilinear fusion of multiscale features," *Scientific Reports*, vol. 13, no. 1, p. 249, Jan. 2023, ISSN: 2045-2322. DOI: 10.1038/s41598-023-27524-4. [Online]. Available: https://doi.org/10.1038/s41598-023-27524-4.

[47] S. Manikandan, R. Raman, R. Rajalakshmi, S. Tamilselvi, and R. J. Surya, "Deep learning–based detection of diabetic macular edema using optical coherence tomography and fundus images: A meta-analysis," *Indian Journal of Ophthalmology*, vol. 71, no. 5, 2023, ISSN: 0301-4738. DOI: https://journals.lww.com/ijo/fulltext/2023/05000/deep_learning_based_detection_of_diabetic_macular.19.aspx.

[48] A. Mao, M. Mohri, and Y. Zhong, "Cross-entropy loss functions: Theoretical analysis and applications," *arXiv preprint arXiv:2304.07288*, 2023. arXiv: 2304.07288 [cs.LG]. [Online]. Available: https://arxiv.org/abs/2304.07288.

[49] W. Nazih, A. O. Aseeri, O. Y. Atallah, and S. El-Sappagh, "Vision transformer model for predicting the severity of diabetic retinopathy in fundus photography-based retina images," *IEEE Access*, vol. 11, pp. 117 546–117 561, 2023. DOI: 10.1109/ACCESS.2023.3326528.

[50] I. Qureshi et al., "Medical image segmentation using deep semantic-based methods: A review of techniques, applications and emerging trends," *Information Fusion*, vol. 90, pp. 316–352, 2023, ISSN: 1566-2535. DOI: 10.1016/j.inffus.2022.09.031. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1566253522001695.

[51] P. Bidwai et al., "Multimodal image fusion for the detection of diabetic retinopathy using optimized explainable ai-based light gbm classifier," *Information Fusion*, vol. 111, p. 102 526, 2024, ISSN: 1566-2535. DOI: 10.1016/j.inffus.2024.102526. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S156625352400304X.

[52] I. G. Brennan et al., "Addressing technical failures in a diabetic retinopathy screening program," *Clinical Ophthalmology*, vol. 18, pp. 431–440, 2024, PMID: 38356695. DOI: 10.2147/OPTH.S442414. eprint: https://www.tandfonline.com/doi/pdf/10.2147/OPTH.S442414. [Online]. Available: https://www.tandfonline.com/doi/abs/10.2147/OPTH.S442414.

[53] R. Chen et al., "Retinal age gap as a predictive biomarker for future risk of clinically significant diabetic retinopathy," *Acta Diabetologica*, vol. 61, no. 3, pp. 373–380, 2024, ISSN: 1432-5233. DOI: 10.1007/s00592-023-02199-5. [Online]. Available: https://doi.org/10.1007/s00592-023-02199-5.

[54] Y.-J. Choi, J.-W. Kwon, and D. Jee, "The relationship between blood vitamin a levels and diabetic retinopathy: A population-based study," *Scientific Reports*, vol. 14, no. 1, p. 491, 2024. DOI: 10.1038/s41598-023-49937-x. [Online]. Available: https://doi.org/10.1038/s41598-023-49937-x.

[55] L. Dai et al., "A deep learning system for predicting time to progression of diabetic retinopathy," *Nature Medicine*, vol. 30, no. 2, pp. 584–594, Feb. 2024, ISSN: 1546-170X. DOI: 10.1038/s41591-023-02702-z. eprint: https://doi.org/10.1038/s41591-023-02702-z. [Online]. Available: https://doi.org/10.1038/s41591-023-02702-z.

[56] S. Al-Fahdawi et al., "Fundus-deepnet: Multi-label deep learning classification system for enhanced detection of multiple ocular diseases through data fusion of fundus images," *Information Fusion*, vol. 102, p. 102 059, 2024, ISSN: 1566-2535. DOI: 10.1016/j.inffus.2023.102059. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1566253523003755.

[57] S. Fleifil, L. Azzouz, G. Yu, C. Powell, N. Bommakanti, and Y. M. Paulus, "Quantitative biomarkers of diabetic retinopathy using ultra-widefield fluorescein angiography," *Clinical Ophthalmology*, vol. 18, pp. 1961–1970, 2024. DOI: 10.2147/OPTH.S462223. eprint: https://www.tandfonline.com/doi/pdf/10.2147/OPTH.S462223. [Online]. Available: https://www.tandfonline.com/doi/abs/10.2147/OPTH.S462223.

[58] O. Islam, K. Kumer, S. Akter, and M. Mohiuddin, "Multi-head self-attention mechanisms in vision transformers for retinal image classification," in *Proceedings of the 2024 International Conference on Computing, Power and Advanced Signal Processing (COMPAS)*, Sep. 2024, pp. 1–5. DOI: 10.1109/COMPAS60761.2024.10795956. [Online]. Available: https://doi.org/10.1109/COMPAS60761.2024.10795956.

[59] A. Jabbar et al., "A lesion-based diabetic retinopathy detection through hybrid deep learning model," *IEEE Access*, vol. 12, pp. 40 019–40 036, 2024. DOI: 10.1109/ACCESS.2024.3373467. [Online]. Available: https://doi.org/10.1109/ACCESS.2024.3373467.

[60] Y. Kang, H. Zhang, X. Wang, Y. Yang, and Q. Jia, "Mmdb: Multimodal dual-branch model for multi-functional bioactive peptide prediction," *Analytical Biochemistry*, vol. 690, p. 115 491, 2024, ISSN: 0003-2697. DOI: 10.1016/j.ab.2024.115491. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0003269724000356.

[61] K. Leth-Møller Christensen, D. B. Kristjansen, A. S. Vergmann, T. L. Torp, T. Peto, and J. Grauslund, "Retinal vascular structure independently predicts the initial treatment response in neovascular age-related macular degeneration," *Acta Ophthalmologica*, vol. 102, no. 1, pp. 116–121, 2024. DOI: https://doi.org/10.1111/aos.15709. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/aos.15709. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1111/aos.15709.

[62] Z. Liu et al., "Cross-modal attention network for retinal disease classification based on multi-modal images," *Biomedical Optics Express*, vol. 15, no. 6, pp. 3699–3714, May 2024, ISSN: 2156-7085. DOI: 10.1364/BOE.516764. [Online]. Available: https://doi.org/10.1364/BOE.516764.

[63] C. Raptis, E. Karavasilis, G. Anastasopoulos, and A. Adamopoulos, "Comparative analysis of conventional cnn v's imagenet pretrained resnet in medical image classification," *Information*, vol. 15, no. 12, 2024, ISSN: 2078-2489. DOI: 10.3390/info15120806. [Online]. Available: https://www.mdpi.com/2078-2489/15/12/806.

[64] D. Rivera-De-la-Parra et al., "Association between uric acid and referable diabetic retinopathy in patients with type 2 diabetes," *Scientific Reports*, vol. 14, no. 1, p. 12 968, Jun. 2024, ISSN: 2045-2322. DOI: 10.1038/s41598-024-63340-0. [Online]. Available: https://doi.org/10.1038/s41598-024-63340-0.

[65] T. Shahzad, M. Saleem, M. S. Farooq, S. Abbas, M. A. Khan, and K. Ouahada, "Developing a transparent diagnosis model for diabetic retinopathy using explainable ai," *IEEE Access*, vol. 12, pp. 149 700–149 709, 2024. DOI: 10.1109/ACCESS.2024.3475550.

[66] H. Si et al., "A dual-branch model integrating cnn and swin transformer for efficient apple leaf disease classification," *Agriculture*, vol. 14, no. 1, p. 142, 2024, ISSN: 2077-0472. DOI: 10.3390/agriculture14010142. [Online]. Available: https://www.mdpi.com/2077-0472/14/1/142.

[67] S. Wang et al., "Advances and prospects of multi-modal ophthalmic artificial intelligence based on deep learning: A review," *Eye and Vision*, vol. 11, no. 1, p. 38, 2024. DOI: 10.1186/s40662-024-00405-1. [Online]. Available: https://doi.org/10.1186/s40662-024-00405-1.

[68] K. D. K. Wardhani, S. Kasim, A. Erianda, and R. Hassan, "Deep learning-based method in multimodal data for diabetic retinopathy detection," *International Journal on Advanced Science, Engineering and Information Technology*, vol. 14, no. 5, pp. 1602–1608, Oct. 2024. DOI: 10.18517/ijaseit.14.5.11677. [Online]. Available: https://ijaseit.insightsociety.org/index.php/ijaseit/article/view/11677.

[69] F. Xiao et al., "Cross-fundus transformer for multi-modal diabetic retinopathy grading with cataract," *arXiv preprint arXiv:2411.00726*, 2024. arXiv: 2411.00726 [eess.IV]. [Online]. Available: https://arxiv.org/abs/2411.00726.

[70] L. Xiong, X. Yuan, Z. Hu, X. Huang, and P. Huang, "Gated fusion adaptive graph neural network for urban road traffic flow prediction," *Neural Processing Letters*, vol. 56, no. 1, p. 9, Feb. 2024, ISSN: 1573-773X. DOI: 10.1007/s11063-024-11479-2. [Online]. Available: https://doi.org/10.1007/s11063-024-11479-2.

[71] Z. Zhao et al., "Image fusion via vision-language model," in *Proceedings of the International Conference on Machine Learning (ICML)*, 2024.

[72] J. Zhu, J. Huang, Y. Sun, W. Xu, and H. Qian, "Emerging role of extracellular vesicles in diabetic retinopathy," *Theranostics*, vol. 14, no. 4, pp. 1631–1646, Feb. 2024, PMC10879872, ISSN: 1838-7640. DOI: 10.7150/thno.92463. [Online]. Available: https://doi.org/10.7150/thno.92463.

[73] S. Zhu, C. Xiong, Q. Zhong, and Y. Yao, "Diabetic retinopathy classification with deep learning via fundus images: A short survey," *IEEE Access*, vol. 12, pp. 20 540–20 558, 2024. DOI: 10.1109/ACCESS.2024.3361944.

[74] I. Bressler et al., "Autonomous screening for diabetic macular edema using deep learning processing of retinal images," *Ophthalmology Science*, vol. 5, no. 4, p. 100 722, Jul. 2025, ISSN: 2666-9145. DOI: 10.1016/j.xops.2025.100722. [Online]. Available: https://doi.org/10.1016/j.xops.2025.100722.

[75] R. Dulal, L. Zheng, and M. A. Kabir, "Mhaff: Multi-head attention feature fusion of cnn and transformer for cattle identification," *arXiv preprint arXiv:2501.05209*, 2025. arXiv: 2501.05209 [cs.CV]. [Online]. Available: https://arxiv.org/abs/2501.05209.

[76] W. Gao et al., "Adf-oct: An advanced assistive diagnosis framework for study-level macular optical coherence tomography," *Information Fusion*, vol. 117, p. 102 877, 2025, ISSN: 1566-2535. DOI: 10.1016/j.inffus.2024.102877. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1566253524006559.

[77] J. Li, T. Chen, X. Wang, Y. Zhong, and X. Xiao, "Adapting the segment anything model for multi-modal retinal anomaly detection and localization," *Information Fusion*, vol. 113, p. 102 631, 2025, ISSN: 1566-2535. DOI: 10.1016/j.inffus.2024.102631. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1566253524004093.

[78] Q. Meng, J. Guo, H. Zhang, Y. Zhou, and X. Zhang, "A dual-branch model combining convolution and vision transformer for crop disease classification," *PLOS ONE*, vol. 20, no. 4, pp. 1–23, Apr. 2025, ISSN: 1932-6203. DOI: 10.1371/journal.pone.0321753. [Online]. Available: https://doi.org/10.1371/journal.pone.0321753.

[79] OpenCV Documentation, *Smoothing images*, https://docs.opencv.org/3.4/dc/dd3/tutorial_gausian_median_blur_bilateral_filter.html, Accessed: 2025-09-11, 2025.

[80] J. Qin, J. Xiong, and Z. Liang, "Cnn–transformer gated fusion network for medical image super-resolution," *Scientific Reports*, vol. 15, no. 1, p. 15 338, May 2025, ISSN: 2045-2322. DOI: 10.1038/s41598-025-00119-x. [Online]. Available: https://doi.org/10.1038/s41598-025-00119-x.

[81] S. Safari Sanjani, J.-B. Boin, and K. Bergen, "Blood vessel segmentation in retinal fundus images," *Stanford University Project Report*, Sep. 2025. eprint: https://jbboin.github.io/doc/vessel_segmentation_report.pdf. [Online]. Available: https://jbboin.github.io/doc/vessel_segmentation_report.pdf.

[82] Y. Shihata, "Gated recursive fusion: A stateful approach to scalable multimodal transformers," *arXiv preprint arXiv:2507.02985*, 2025. arXiv: 2507.02985 [cs.CV]. [Online]. Available: https://arxiv.org/abs/2507.02985.

[83] J. M. Stewart, M. Coassin, and D. M. Schwartz, "Diabetic retinopathy," in *Endotext [Internet]*, K. R. Feingold, S. F. Ahmed, B. Anawalt, et al., Eds., Updated 2025 May 14, South Dartmouth (MA): MDText.com, Inc., May 2025. [Online]. Available: https://www.ncbi.nlm.nih.gov/books/NBK278967/.

[84] X. Zhou et al., "Deep learning multimodal bilinear fusion network based on vibrational spectroscopy for diagnosis of benign and malignant thyroid tumors," *Microchemical Journal*, vol. 209, p. 112 870, 2025, ISSN: 0026-265X. DOI: 10.1016/j.microc.2025.112870. [Online]. Available: https://doi.org/10.1016/j.microc.2025.112870.

[85] H. Zhu et al., "Development and validation of a 3-d deep learning system for diabetic macular oedema classification on optical coherence tomography images," *BMJ Open*, vol. 15, no. 5, e099167, May 2025, ISSN: 2044-6055. DOI: 10.1136/bmjopen-2025-099167. [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/40449950/.