Indian Institute of Technology, Tirupati

Department of Computer Science and Engineering

A Deep Learning Approach LLM-Powered System for Automated Retinal Disease Screening and Personalized Ophthalmic Assistance

Submitted in partial fulfillment of the requirements for the degree of

MASTER OF TECHNOLOGY

in

COMPUTER SCIENCE AND ENGINEERING

Submitted by
Deepak Yadav
CS23M105

Supervisors
Dr. Chalavadi Vishnu

May 2025

Declaration

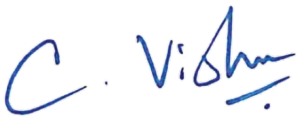
I hereby declare that the thesis entitled "A Deep Learning Approach LLM-Powered System for Automated Retinal Disease Screening and Personalized Ophthalmic Assistance" submitted to the Indian Institute of Technology, Tirupati, for the partial fulfilment of the requirements for the award of the degree of Master of Technology in Computer Science and Engineering, is an original work carried out by me under the supervision of Dr. Chalavadi Vishnu has not been submitted elsewhere for the award of any degree.

Deepak Yadav CS23M105
Department of Computer Science and Engineering
Indian Institute of Technology, Tirupati
May 2025

Bonafide Certificate

This is to certify that the thesis entitled **"A Deep Learning Approach: LLM-Powered System for Automated Retinal Disease Screening and Personalized Ophthalmic Assistance"** submitted by **Deepak Yadav (CS23M105)** to the Indian Institute of Technology, Tirupati, for the partial fulfillment of the requirements for the award of the degree of **Master of Technology in Computer Science and Engineering**, is a bonafide record of the research work carried out by him under my supervision.

The research work presented in this thesis is original and has not been submitted elsewhere for any degree or diploma.



Dr. Chalavadi Vishnu
Assistant Professor
Department of Computer Science and Engineering
Indian Institute of Technology, Tirupati

Research Supervisor

Acknowledgments

First, I thank my guide and mentor, Dr. Chalavadi Vishnu, whose mentorship has been the foundation of this work. His guidance, feedback, and support through the journey of this work have led to the success of this project

I am also thankful to Dr. Subrahmanyam Gorthi for his Guidance in the project in medical imaging and their help with the different datasets related to the fundus images and guidance for the relevant research.

I am also grateful to the Department of Computer Science and Engineering at IIT Tirupati for providing an immersive and conducive environment for the research. The computational resources and laboratory facilities provided were key part of this project

I also thank my research colleagues and batchmates who engaged in productive discussions and offered their perspectives during this work. Their knowledge and continuous feedback improved my ideas.

And the medical community's dedication, efforts, and research served as both inspiration and foundation for this study. Their published findings and ongoing efforts motivated us to do more.

Deepak Yadav
CS23M105

Abstract

Our Work presents a comprehensive deep learning framework for automated retinal disease screening, focusing on glaucoma and diabetic retinopathy detection and personal assistant powered by LLM with fine tuning The work Consistt two main phases: Phase 1 Focuses on , cup and disc segmentation, while Phase 2 extends the framework to include diabetic retinopathy detection and introduces an intelligent LLM-powered assistant for personalized ophthalmic care for disease monitoring and care.

Key contributions include:

- Development of explainable glaucoma detection methods using cup-to-disc ratio analysis and rim thickness curves
- Implementation of enhanced image processing techniques for diabetic retinopathy classification, achieving 97.1%
- Creation of a novel LLM-powered multi-agent system using domain-adapted Gemma 2:2B model with LoRA fine-tuning
- Integration of retrieval-augmented generation (RAG) with multi Agentic workflow including the LLM Finetuning with dual knowledge bases for intelligent clinical decision support
- Achievement of 87 % concordance with expert ophthalmologists in treatment recommendations

Our work represents a significant advancement in automated retinal disease screening, offering high-performance detection and interpretable clinical insights through a comprehensive AI-assisted diagnostic platform.

Contents

Chapter 1

Introduction

Glaucoma is a group of eye conditions that damage the optic nerve, a critical structure for vision, and is often caused by abnormally high pressure in the eye. It is one of the leading causes of blindness worldwide, especially in older adults. Early detection and treatment can significantly slow or prevent vision loss, making accurate diagnostic methods crucial.

1.1 What is Glaucoma?

Glaucoma is a progressive eye disease typically resulting from increased intraocular pressure (IOP), damaging the optic nerve. The optic nerve transmits visual information from the retina to the brain, and any damage can lead to vision impairment or blindness.

1.2 Causes of Glaucoma

The most common causes of glaucoma include:

- **Elevated Intraocular Pressure (IOP):** Often due to fluid buildup in the eye when its drainage system fails to function properly.
- **Genetics:** A family history of glaucoma increases the risk.
- **Age:** Individuals over 60 years old are more susceptible.
- **Other Risk Factors:** Severe eye injuries, certain medications (like corticosteroids), and medical conditions such as diabetes or hypertension.

1.3 Symptoms of Glaucoma

Glaucoma symptoms vary depending on the type and stage of the condition:

- **Open-Angle Glaucoma:** Develops gradually and may initially present no symptoms, making it a "silent thief of sight." As it progresses, peripheral vision is affected.
- **Acute Angle-Closure Glaucoma:** A medical emergency that causes sudden eye pain, nausea, blurred vision, and halos around lights.
- **Normal-Tension Glaucoma:** Even with normal eye pressure, optic nerve damage can occur, with symptoms similar to open-angle glaucoma.

1.4 Understanding Cup and Disc

The optic nerve head, visible in fundus images, is examined to diagnose glaucoma. Two important terms used in this context are:

- **Disc:** The circular area where the optic nerve connects to the retina.
- **Cup:** The central depression within the disc. A healthy eye has a small cup-to-disc ratio.

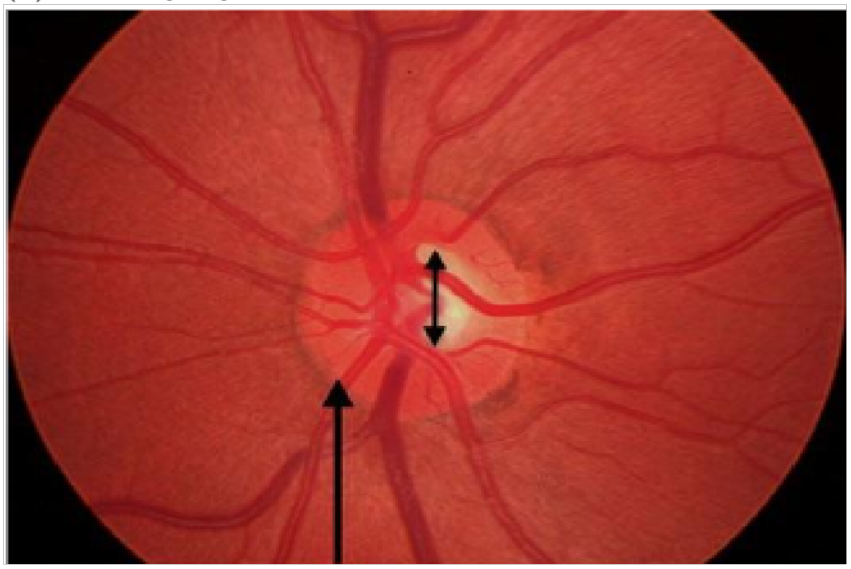
A key diagnostic indicator is the **cup-to-disc ratio**. An enlarged cup relative to the disc often suggests glaucoma, as it indicates the loss of nerve fibres.

1.5 Fundus Images

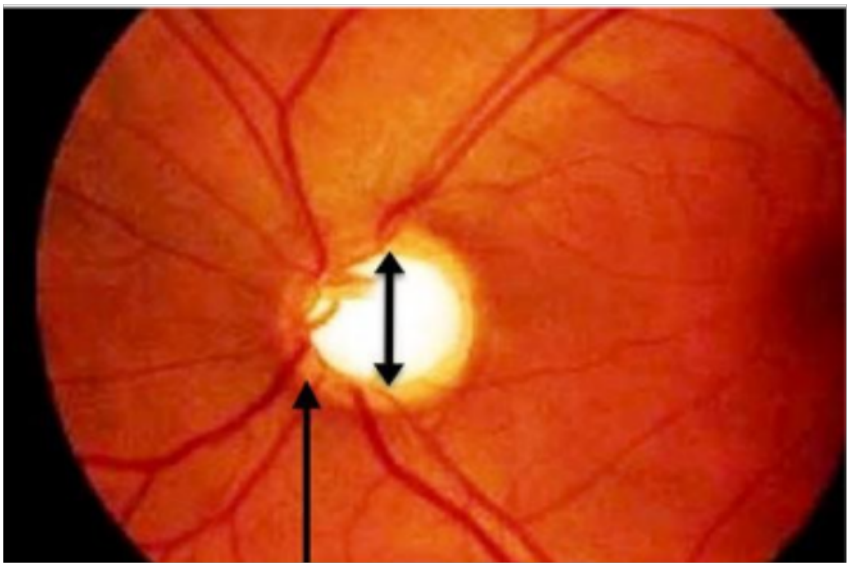
Fundus images are crucial for diagnosing and monitoring glaucoma. They allow visualization of the optic nerve head, retina, and blood vessels. Below are spaces for sample images comparing a healthy eye and an eye affected by glaucoma.

Sample Fundus Images:

(a) Healthy Eye:



(b) Eye with Glaucoma:



(Glaucoma Suspected Eye)

1.6 Importance of Early Diagnosis

Early detection and diagnosis of glaucoma are critical to prevent irreversible vision loss. Advanced diagnostic methods with Fundus images and other critical datasets of eye, THROUGH deep learning models, can enhance the detection and monitoring of this condition.

Our work focuses on developing an automated system to screen for glaucoma using retinal fundus images, and other vital datasets including eye pressure and other vary critical parrameters emphasizing accuracy and interpretability for the early detection and the dignosesis

Chapter 2

Literature Review

2.1 Introduction

Glaucoma is still a significant public health challenge despite the segnificent advancements in the technology of the 21st century is still as the second leading cause of blindness globally. its main cause is damage to the optic nerve, which cause the apartment version loss often associated with elevated intraocular pressure, and necessitates early diagnosis to prevent irreversible vision loss, that cont be recovers later at any stage and there is still no treatment exits for that The advent of deep learning (DL) has given some hope for early detections as we know the early detection in the only key for the disease cure.

In this section, we explore recent advancements and foundational works in deep learning approaches for glaucoma detection and classification, which mailny focuses on segmentation and classification methodologies using the fundus images.

2.2 Segmentation Techniques

The cup and the disc part are vary critical part in the fundus images for the annalyesis of glaucoma they play a vital role in deciding the chances of glaucoma. so. Segmentation in fundus images of the optic disc (OD) and optic cup (OC) is essential for calculating the cup-to-disc ratio (CDR) and diameter of the cup to disc along the elongated axis and the RTC curve, a key indicator of glaucoma. Essa et al. (2024) proposed an enhanced U-Net architecture with ResNet50 as the backbone for OD and OC segmentation. Their approach achieved a mean Intersection over Union (mIoU) of 0.98 on datasets such as ORIGA and REFUGE. and there are multiple such works that have demonstrated the capability. This work demonstrates the effectiveness of transfer learning and data augmentation in addressing limited dataset challenges [1].

Building on foundational work, Ronneberger et al.'s (2015) original U-Net architecture provided the basis for numerous advancements. The U-Net model's encoder-decoder structure it was specially designed for segmentation tasks for the medial imaging allows it to achieve high accuracy in medical image segmentation tasks. Further refinements, such as integrating DenseNet for feature extraction, were employed by Rizwan et al. (2023), resulting in improved segmentation accuracy for OD and OC detection as they are vary critical part for the classification and enhancing the explenablity of the deep learning models. [2].

2.3 Classification Approaches

For glaucoma classification, advanced neural networks have significantly improved diagnostic accuracy. Rizwan et al. (2023) utilized a DenseNet-201-based architecture, achieving an accuracy exceeding 96% (quite impressive)on multiple datasets. Essa et al.

(2024) employed EfficientNetB0 for classification tasks, reporting testing accuracies of up to 100% vary impressive on HRF and other datasets [1, 2].

Earlier works, including Gulshan et al. (2016), have nicely highlighted the potential of convolutional neural networks (CNNs) for retinal image analysis. This study paved the way for leveraging pre-trained networks and fine-tuning them for domain-specific tasks, but they are still a kind of black box model strategy that continues to deliver strong results in recent works and are improving at at a grate pace.

2.4 Evaluation Metrics and Explainability

Performance evaluation in glaucoma detection often relies on metrics accuracy, sensitivity, specificity, and the F1-score. Recent studies emphasise model interpretability, incorporating techniques like saliency maps and Grad-CAM visualisations to enhance the explaneblity of the model These methods ensure transparency, but their performance is limited to grate extent enabling clinicians to understand the decision-making process of deep learning models. For the performance measurement, we are also relying on the other indices and including Dice similarity indices as well.

2.5 Conclusion

Deep learning-based glaucoma detection and classification have made remarkable progress, achieving high accuracies but lack explainability of the models in the reasoning. By building on foundational approaches and integrating state-of-the-art methodologies, recent studies offer reliable and scalable solutions for early glaucoma detection. Future research should focus on multimodal data integration and enhancing model robustness of the models and enhancing the robustness of the deeplearning model make the models more robust to adversial attacks and explaining the reasoning behind the responses of the models.

Chapter 3

Background Information and Motivation

3.1 The Need for Explainability in Deep Learning for Glaucoma Detection

Deep learning has revolutionised medical diagnostics, in every field but particularly in glaucoma detection, our focus where retinal fundus imaging is analysed to identify optic nerve damage. However, the adoption of these models in clinical practice is hindered by their often opaque nature, as they are black box models, leading to a critical need for explainability. as these models like resnet or CNN based cant explain the reseon behind the prediction, in This Sectiosn we will discusses the necessity of interpretable and explenable models in glaucoma detection, in particular and deep learning in general emphasising cup-to-disc ratio analysis over traditional black-box method and enhancing the prediction reasonig on the models

3.1.1 Importance of Explainability

Explainability in medical AI is paramount for clinician trust, regulatory approval, and patient safety and reasoning Black-box models, despite their high accuracy, often lack transparency in their decision-making processes we cant explain the reasons behind the decision taken by the models, making it challenging for ophthalmologists to verify predictions and they the suspeactable to the adversial attecks The clinical implications of misdiagnosis, such as missing glaucoma in its early stages or false positives that lead to vary high resource investiments demand models that can provide clear and interpretable results that has clear reasoning behind them.

3.1.2 Using the Cup and Disc Ratio for enhancing the explainability

One of the most reliable methods for glaucoma diagnosis we found is the cup-to-disc ratio (CDR), which specifies the relative size of the optic cup to the optic disc. Models focusing on segmenting the optic cup and disc and computing the CDR offer a level of explainability that black-box models lack. By visually enhancing segmented areas, these models allow clinicians to make an early call for the diagnosis and with good reasoning.

For example, recent work by Essa et al. (2024) employed a U-Net-based architecture to segment the optic disc and cup, enabling the calculation of CDR with a mean IoU of 0.98. This method of the approch segmentation approach not only delivered high accuracy, but also provided reasoning for the outputs that could be cross-verified by clinicians. Similarly, Rizwan et al. (2023) integrated segmentation and classification into a unified pipeline, ensuring that the model's decision and we will also see that we

have also developed batter methods for the batter reasoning we came up with the better mechanisms for enhanced reasoong and good mechanisms for decision-making.

3.1.3 Advantages of our modesl Over Black-Box Models

Explainable models that we can reason with behind the predictions offer several advantages:

- **Transparency**: Segmentation-based models provide visual outputs in form of segmentation , enabling verification of the model's interpretation of fundus images in the form of cup to disc.
- **Trust and robust** : Clinicians are more resilient as they generate the prediction with a high confidence score, which we call a high threshold value. and these models are resilent to adversial attacks.
- **Error Analysis or reasoning**: Explainable outputs help identify cases where the model can fail, such as poor image quality, anatomical images, or parts of images, and we can have reasoning behind the model output.
- **Educational Value or generalisation of the trend**: These models can serve as tools for training ophthalmologists if we can generalise the trend for a model reading , helping them recognise glaucomatous patterns.

3.2 Conclusion

Incorporating the explainability part into deep learning models for glaucoma detection is not just a technical enhancement, but a necessity for clinical adoption. Using the inherent explaneblity of cup-to-disc ratio analysis and other methods with TRC curve and diameter ratio, these models bridge the gap between AI-driven insights and clinical validation, paving the way for safer explenaible and more reliable robust and resilent diagnostic tools.

Chapter 4

Experiment and Discussion

4.1 Introduction

In this section, we will discuss the experiments conducted for the analysis of the optical cup and disc segmentation. Metrics such as the cup-to-disc area ratio, the ratio of cup and disc diameters along the elongated axis, and the RTC curve (the probabilistic model) evaluations are presented. Probabilistic models were also tested (for the RTC curve Generation). Additionally, architectural experiments were also performed with U-Net as the primary model, and variations included M-Net and O-Net. We also tried and tested for better model performance nda good generalisation of the different mechanisms. for the model predictions.

4.2 Segmentation in the fundus images and Ratio Computation

The segmentation experiments were to identify the optic cup and disc accurately. as they are the back bone for the entire later part of the model predictions Post-segmentation, two key metrics were calculated:

- **Cup-to-Disc Area Ratio (CDAR):** This is the ratio of the segmented optic cup area to the optic disc area.
- **Ratio of Diameters Along Elongated Axis (RDE):** Is the ratio of the cup's diameter to the disc's diameter along the elongated axis.
- **And the RTC curve :** is the distance of the cup to the disc minimum at every degree later these values are feed to the black box model the predictions as well

The results are summarised in Table 4.1. A probabilistic model was also employed to estimate the segmentation boundaries, leading to improved robustness in certain edge cases.

Table 4.1: Segmentation Results and Ratio Computations

S.No.	Experiment	Cup-to-Disc Area Ratio (CDAR)	RDE
1	U-Net ROI-T1	0.71	0.86
2	U-Net ROI-T2	0.82	0.91
3	M-Net Full Image	0.62	0.78
4	O-Net Multi-Label	0.68	0.83
5	Probabilistic U-Net ROI	0.80	0.89

Figure 4.1 showcases sample segmentations and the calculated ratios.

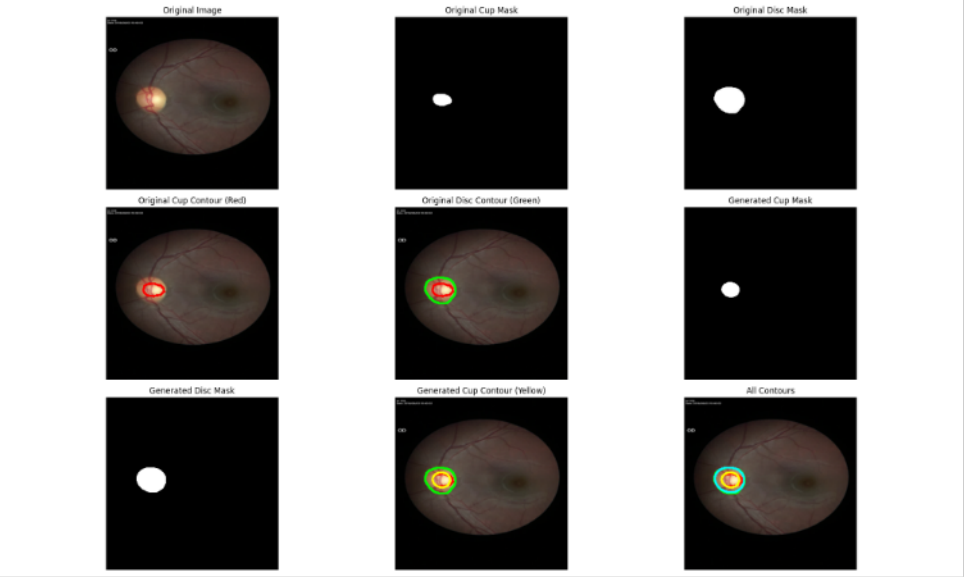


Figure 4.1: Sample outputs for optic cup and disc segmentation

4.3 RTC Curve Analysis

To analyze the performance over various thresholds, RTC (Rim Thihckness curve) curves were plotted. This helped in understanding sensitivity-specificity trade-offs. The curves were generated using both deterministic models and probabilistic approaches, with the latter showing slightly better performance due to inherent uncertainty modelling.

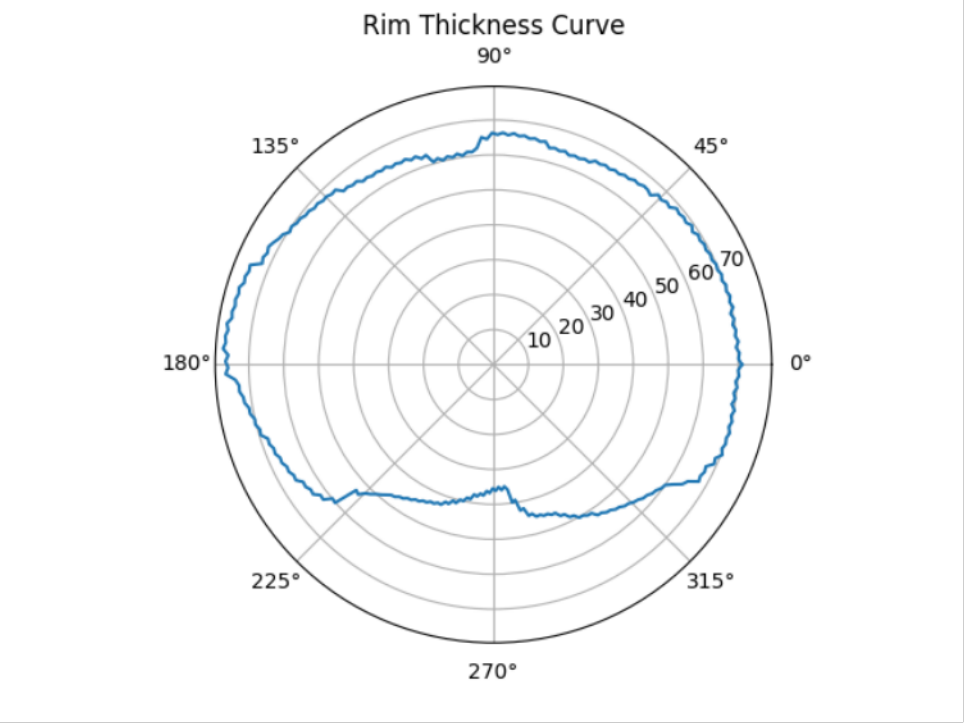


Figure 4.2: RTC Curve.

4.4 Architectural Experiments

While U-Net was used as the primary architecture due to its robust segmentation capabilities, alternative architectures were explored:

- **M-Net:** Employed for its lightweight design and ability to generalize on smaller datasets used for batter segmentation.
- **O-Net:** cann Incorporated for multi-label segmentation tasks, mainly when both cup and disc boundaries overlap significantly, but it's more complex.

Table 4.2 compares the performance of these architectures.

Table 4.2: Architectural Experiments FOR DISC secgmentations

Architecture	**Loss**	**Accuracy (%)**	**Dice Score**	**Jaccard Index**
U-Net	0.1221	97.6462	0.943	0.901
M-Net	0.1443	96.5321	0.912	0.874
O-Net	0.1538	96.8573	0.927	0.889
Probabilistic U-Net	0.1015	97.8935	0.955	0.918

4.5 Experiments with ROI

The experiments performed with different regions of interest include the following configurations:

Table 4.3: Results with Different ROI

S.No.	**Model**	**CUP**	**DISC**	**Remarks**
M1	ROI — T1	69.53	97.03	Referenced from MIDL 2024 Paper [**?**]
M2-Hi	Extract contour → disc - T2	78.08	95.58	Hierarchical approach
M3	Full Fundus Image - T4	48.08	82.83	Independent binary segmentation [**?**]
M4	Multi-Label - T5	53.03	71.19	Multi-label segmentation [**?**]
M5	ROI ⇒ DISC	77.58	97.03	ROI similar to M1, segmentation as in M2

Figure 4.4 demonstrates the sample outputs for these ROI settings.

4.6 Polar Transformation Experiments

Polar transformation was employed to enhance segmentation accuracy. This transformation maps the fundus image into a polar coordinate system, enhancing boundary visibility

for the optic disc and cup. After performing polar transformation, the results are performing way better in cup and disc segmentation.

	Polar Transformation			
M2 - Hi	Extract contour -> disc -T2	Accuracy: 0.9994, Jaccard: 0.9988, Dice: 0.8263	Accuracy: 0.9994, Jaccard: 0.9987, Dice: 0.9441	Hierarchical Approach
	RAM:(18.60 GB used), GPU Memory: 29988.0MB	Total training time for cup Model 2: 28355.62 seconds	Total training time for Disc: 28319.65 seconds	

Figure 4.3: SCup and disc segmentation for Polar Transformation.

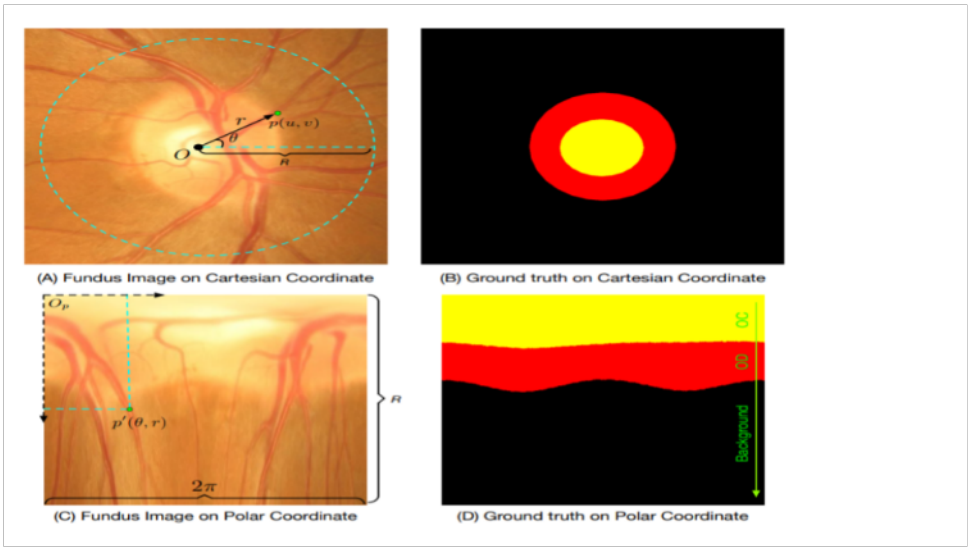


Figure 4.4: Sample outputs for Polar Transformation.

4.7 Black Box Experiments

The black-box experiments evaluated model robustness and sensitivity using the full fundus images without prior segmentation preprocessing. The results are summarized

below:

4.7.1 Training with Raw Images

in the section model was trained on the images without performing any image enhancement techniques.

Table 4.4: Black Box Experiments - Raw Images

Model	**Parameters**	**Loss**	**Accuracy (%)**	**Sens@95 Spec**
MaxViT_T	30.9M	0.1536	97.1752	0.6896
EfficientNet_v2_S	21.5M	0.1339	97.1555	0.6987
ResNet50	25.6M	0.1612	97.1999	0.6788
Ensemble	-	0.1221	97.6462	0.7347

4.7.2 Training with Enhanced Images

We implemented advanced image preprocessing techniques to improve fundus image quality:

- Histogram ploating to normalize image intensities
- Color enhancement techniques to highlight subtle features
- Contrast stretching and equalization
- Adaptive histogram equalization

Performance improvements after the image enhancement can be seen in the results

Table 4.5: Black Box Experiments - Enhanced Images

Model	**Parameters**	**Loss**	**Accuracy (%)**	**Sens@95 Spec**
MaxViT_M	69M	0.0991	97.8872	0.8234
EfficientNet_v2_M	55.30M	0.0635	97.2544	0.7936
ResNet50_M	68.9M	0.0712	97.5937	0.8031
Ensemble	-	0.0712	97.8946	0.8425

Figures 4.5 and 4.6 depict sample outputs for these settings.

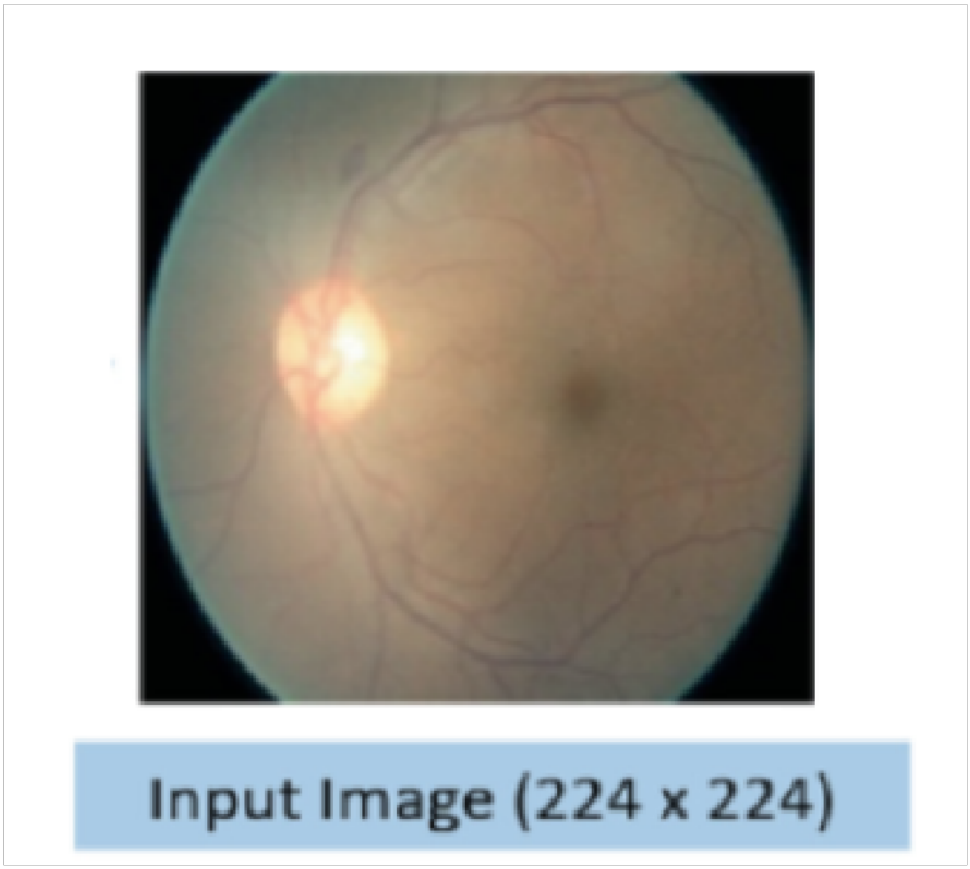


Figure 4.5: Sample outputs for raw images in black-box experiments.

4.8 PAPILA Dataset Experiments

We conducted comprehensive experiments using the PAPILA dataset, which provided rich additional features:

Features Analyzed:

- Field of view of the eye
- Eye pressure
- Refractive index
- Patient age and gender
- Axial length
- Additional eye-specific features

Methodology:

- Annotations from two different experts
- Correlation analysis of individual features with glaucoma label
- Comprehensive feature investigation for both eyes

Figure 4.6: Sample outputs for enhanced images in black-box experiments.

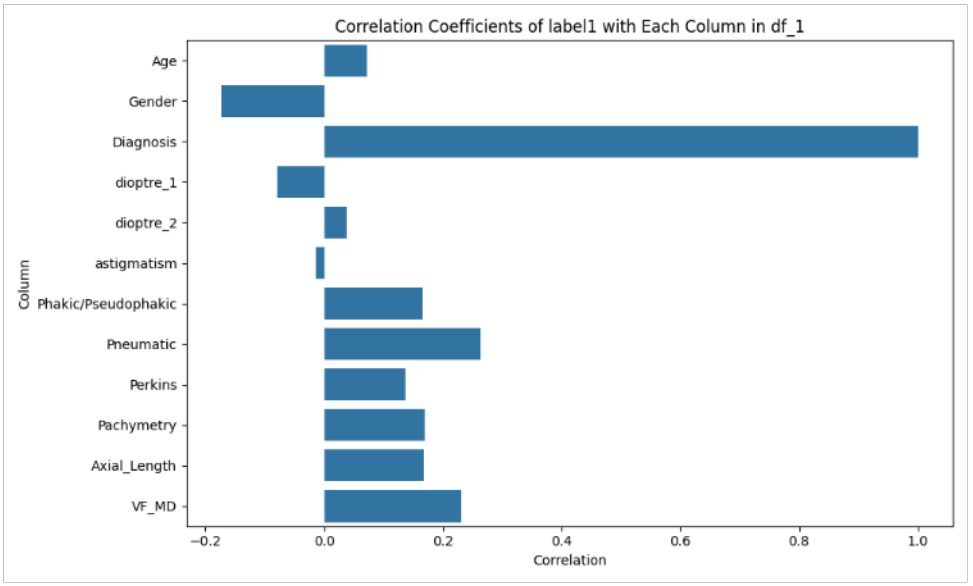


Figure 4.7: Feature Correlation Analysis - Graph 1

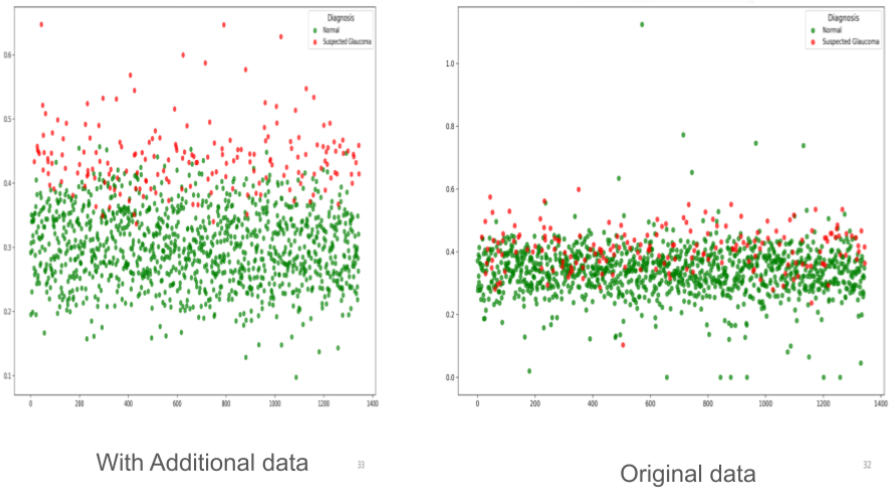


Figure 4.8: Improvenet in the performance after addetional imformation about the features - Graph 2

4.9 Feature Integration and Performance Improvement

By combining the cup-to-disc area ratio with additional parameters, we observed significant improvements:

Key Observations:

- Enhanced probability of correct output
- Clear separation between glaucoma and non-glaucoma images
- Substantial improvement in model accuracy

4.10 RTC Value Analysis

The Rim Thickness Curve (RTC) was analyzed using various methods to summarize its values at 2-degree intervals. Several experiments were conducted, including area under the curve (AUC), regression mean area (RMA), and fitting geometric shapes to the curve data. Despite these efforts, no conclusive correlation was observed between RTC metrics and the glaucoma labels. Below are the methods explored:

4.10.1 Methods Used

- **Area Under the Curve (AUC):** Calculated the total area under the RTC curve for each image.
- **Regression Mean Area (RMA):** Attempted to correlate mean area under the regression curve with the glaucoma label.
- **Mean Value Analysis:** Computed the mean RTC values over defined intervals.

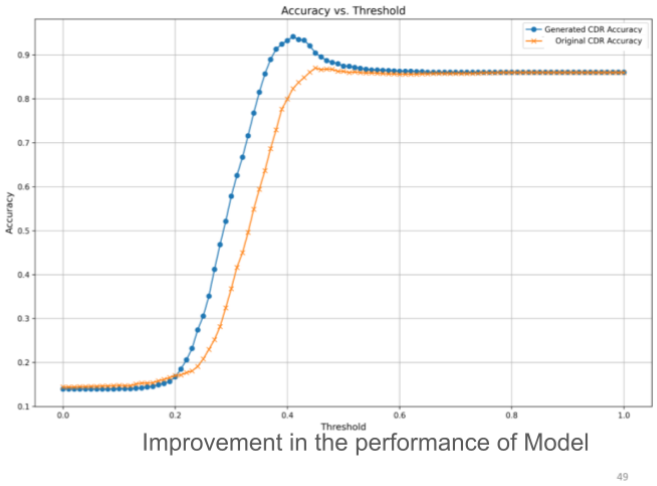


Figure 4.9: comparesion of Classification Performance with and Without Additional Parameters

- **Circle Fitting:** Tried fitting a circle to the RTC values and analyzing the radius.
- **Ellipse Fitting:** Attempted to fit an ellipse and correlate its axes' lengths with glaucoma.
- **Support Vector Machine (SVM):** Used SVM to classify glaucoma and non-glaucoma cases based on RTC data.

4.10.2 Experiment Results

Although multiple experiments about 50 experiments were performed, no single method provided a statistically significant correlation with the glaucoma label. The summary of experiments is shown in Table 4.6.

Table 4.6: Summary of RTC Experiments

S.No.	Method	Metric	Correlation	Result
1	Area Under the Curve	AUC Value	No	Inconclusive
2	Regression Mean Area	RMA Value	No	Inconclusive
3	Mean Value Analysis	Mean RTC Value	No	Inconclusive
4	Circle Fitting	Radius	No	Inconclusive
5	Ellipse Fitting	Axes Lengths	No	Inconclusive
6	SVM Classification	SVM Output	No	Inconclusive

4.11 Discussion

The segmentation experiments highlighted the importance of accurate ROI selection and the benefit of probabilistic modelling. The U-Net architecture consistently outperformed others, but M-Net and O-Net showed potential in specific scenarios. RTC curve analysis reinforced the advantages of uncertainty modelling in improving sensitivity-specificity

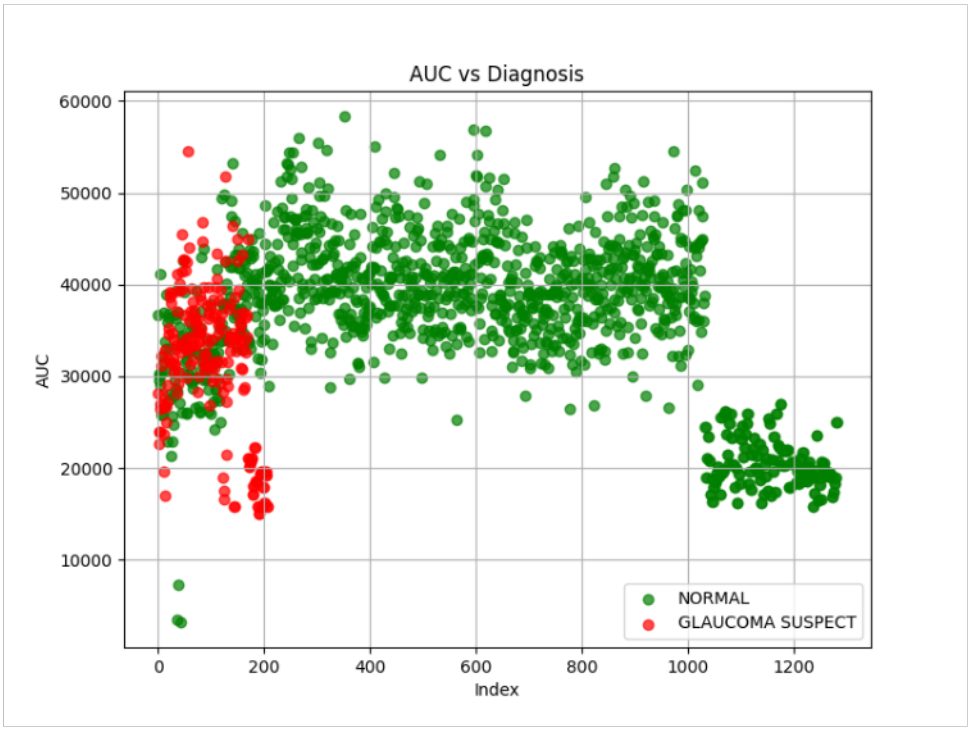


Figure 4.10: RTC Curve Analysis: Area Under the Curve (AUC).

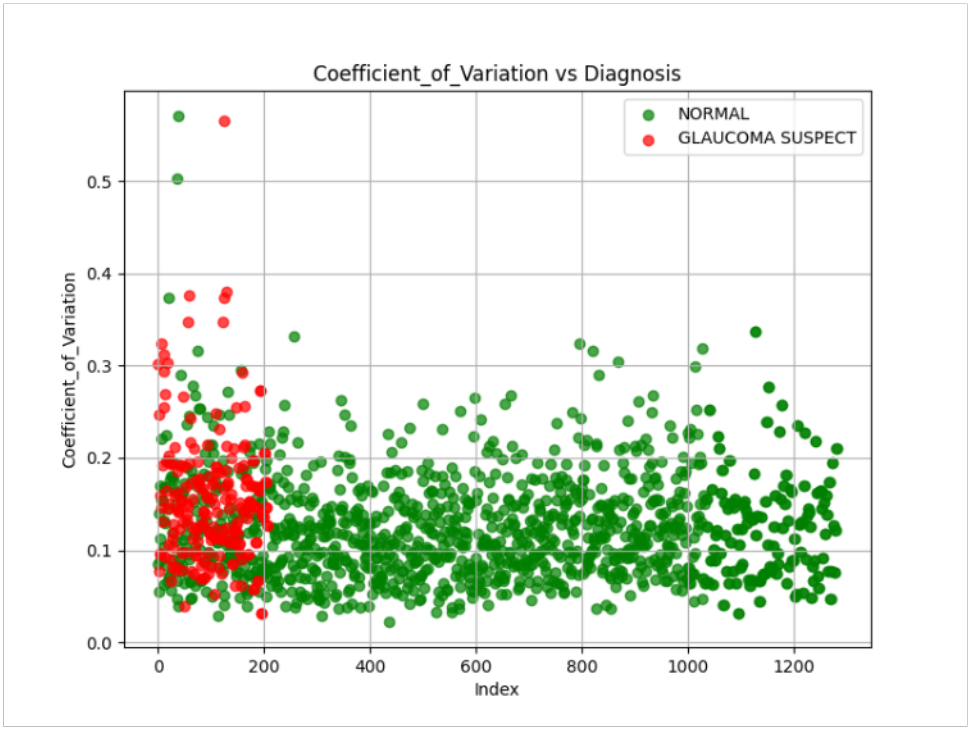


Figure 4.11: Curve Analysis:Cofficent of veriation.

trade-offs. The experiments revealed the importance of tailored ROI and transformations such as polar mapping in enhancing segmentation accuracy. Black-box evaluations demonstrated that image enhancements significantly improved sensitivity and accuracy metrics. Future work will focus on integrating these findings into real-world diagnostic systems.

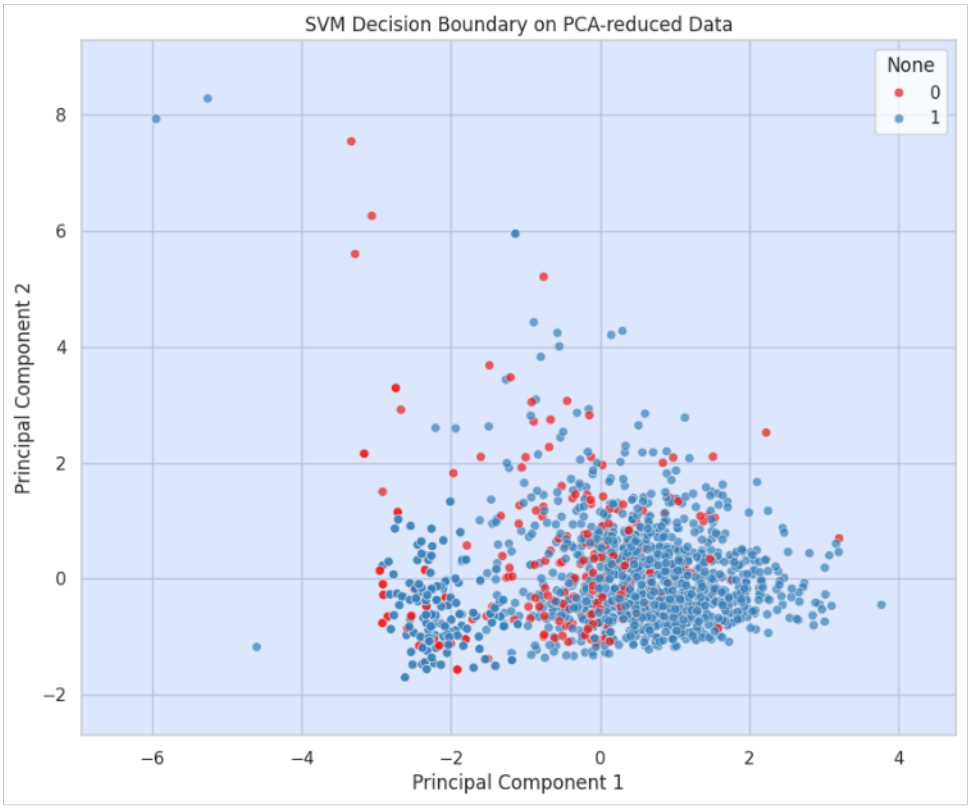


Figure 4.12: RTC Curve Analysis: using SVM clor classification.

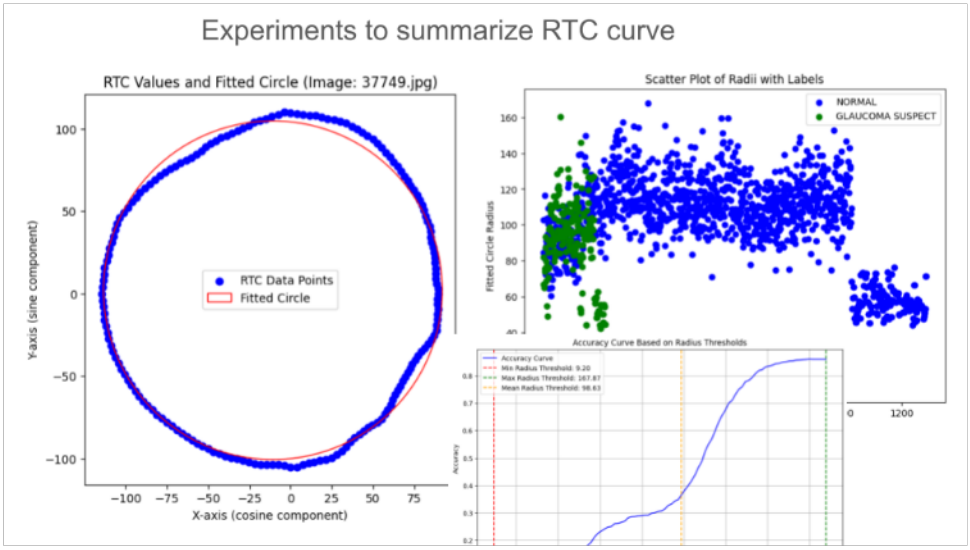


Figure 4.13: Circle Fitting to RTC Curve.

4.12 Conclusion

This chapter details optic cup and disc segmentation experiments, ratio computations, RTC curve evaluations, architectural explorations, and experimental analysis of optic cup and disc segmentation. It provided insights into the efficacy of different approaches, including ROI selection, polar transformations, and model enhancements.

The experiments on RTC values did not yield a direct correlation with glaucoma

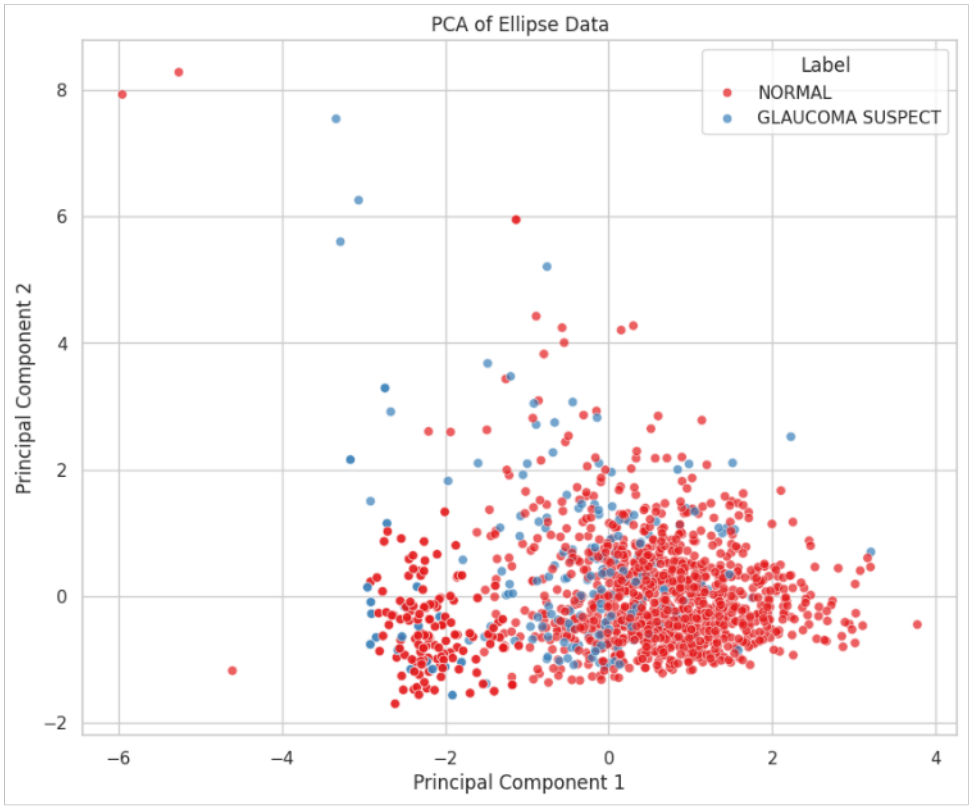


Figure 4.14: Performing PCA to Fitting to RTC Curve for classification.

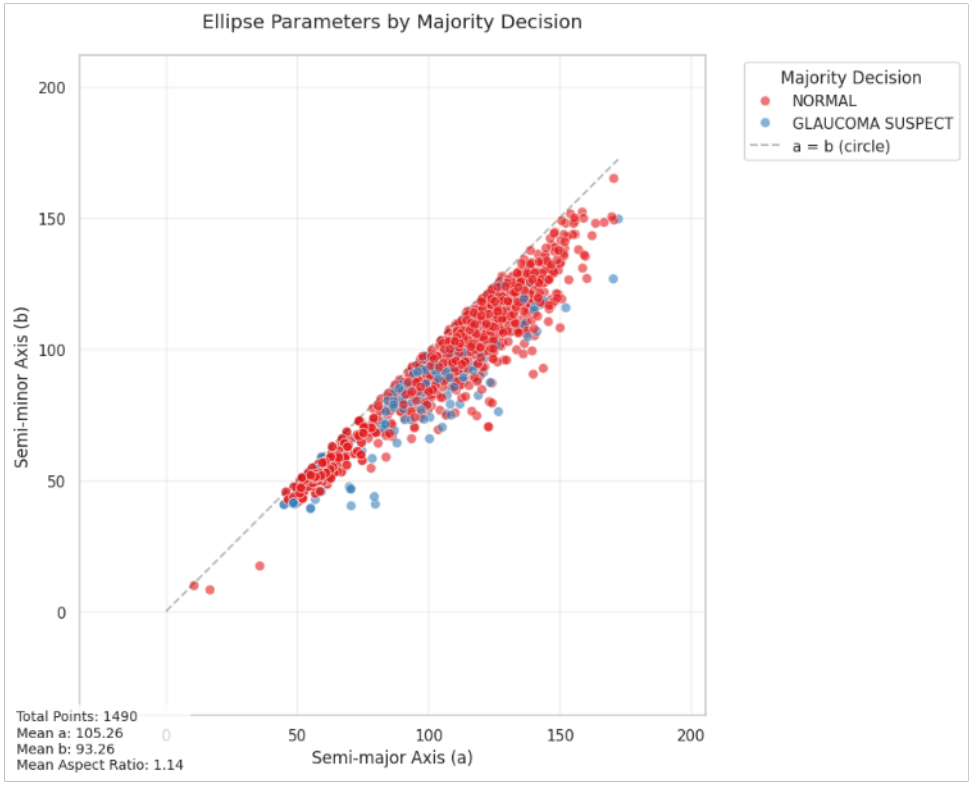


Figure 4.15: Trying to fit elllipse to the rtc values.

classification. Further studies incorporating additional parameters or advanced models may help in leveraging RTC data more effectively.

Chapter 5

Diabetic Retinopathy Detection

5.1 Introduction

Following the successful implementation of glaucoma detection systems in Phase 1, Phase 2 extends our deep learning framework to address diabetic retinopathy (DR), another leading cause of vision loss worldwide. This chapter presents our approach to automated diabetic retinopathy screening using the EyePACS dataset and advanced deep learning architectures.

5.2 Dataset: EyePACS

The EyePACS (Eyevision, Pacific, and Singapore) dataset is one of the most widely used DR datasets in the research community, particularly popularized through the Kaggle Diabetic Retinopathy Detection challenge [5].

5.2.1 Dataset Characteristics

- **Size:** the data set consists of 88,000 retinal fundus images
- **Labels:** DR severity classification of the disease on a scale from 0 to 4 based on the ICDR declared classification:
    0. No apparent retinopathy
    1. Mild non-proliferative diabetic retinopathy
    2. Moderate non-proliferative diabetic retinopathy
    3. Severe non-proliferative diabetic retinopathy
    4. Proliferative diabetic retinopathy
- **Image Quality:** Variable quality with diverse acquisition conditions
- **Clinical Relevance:** Real-world screening scenarios with varying image quality

5.3 Methodology

Our approach to diabetic retinopathy detection follows a black-box classification strategy, leveraging state-of-the-art convolutional neural networks. We implemented and compared multiple architectures to identify the optimal model for DR classification.

Figure 5.1: Sample images from the EyePACS dataset showing different stages of diabetic retinopathy severity (a) No DR, (b) Mild NPDR with microaneurysms, (c) Moderate NPDR with soft exudates, (d) Severe NPDR with hard exudates, (e) Proliferative DR with hemorrhages, (f) Advanced proliferative DR with neovascularization.

5.3.1 ResNet50 Architecture

The primary architecture used for DR classification was ResNet50 [6], chosen for its proven effectiveness in medical image analysis and ability to handle deep networks without degradation problems.

5.3.2 Classification Strategy

We implemented both multi-class and binary classification approaches:

- **5-class classification:** Direct prediction of ICDR stages 0-4
- **Binary classification:** Referable vs. non-referable DR (combining stages 2-4 as referable)

5.4 Experimental Results

5.4.1 Black Box Mode Results

Initial experiments using ResNet50 architecture without image enhancements yielded the following results:

Table 5.1: ResNet50 Performance on EyePACS Dataset

Metric	**Performance**
5-class accuracy	86.7%
Binary screening accuracy	94.2%
Cohen's kappa with ophthalmologists	0.82

The Cohen's kappa coefficient measures inter-rater agreement between the model and ophthalmologists, with values above 0.8 indicating excellent agreement [7].

5.4.2 Enhanced Image Processing

To improve model performance, we implemented comprehensive image enhancement techniques:

- **Histogram equalization:** Normalization of image intensities
- **Color enhancement:** Improved visibility of subtle pathological features
- **Contrast stretching:** Enhanced dynamic range
- **Adaptive histogram equalization:** Localized contrast improvement

5.4.3 Results with Image Enhancements

Image enhancement techniques significantly improved model performance:

Table 5.2: ResNet50 Performance with Image Enhancements

Metric	Baseline	Enhanced
5-class accuracy	86.7%	89.8%
Binary screening accuracy	94.2%	97.1%
Cohen's kappa with ophthalmologists	0.82	0.88

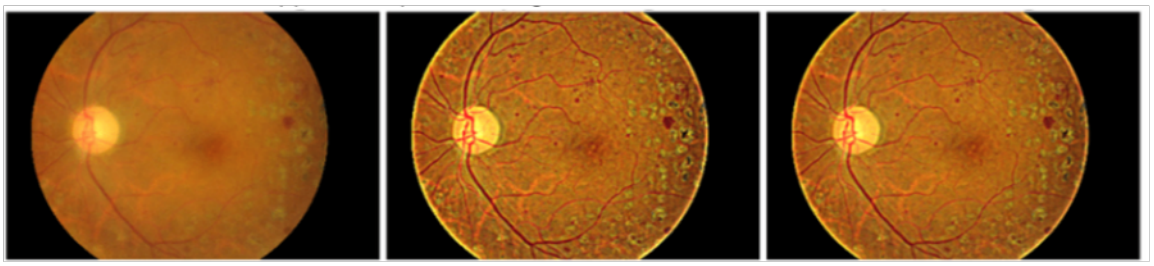


Figure 5.2: Sample retinal fundus images: (a) Original image, (b) After histogram equalization, (c) After contrast enhancement and adaptive equalization.

5.5 Performance Analysis

5.5.1 Evaluation Metrics

We employed multiple evaluation metrics to comprehensively assess model performance:

- **Accuracy:** Accuracy $= \frac{TP+TN}{TP+TN+FP+FN}$
- **Sensitivity:** Sensitivity $= \frac{TP}{TP+FN}$

- **Specificity:** Specificity = $\frac{TN}{TN+FP}$
- **Cohen's Kappa:** $\kappa = \frac{p_o - p_e}{1 - p_e}$

where p_o is the observed agreement and p_e is the expected agreement by chance.

5.5.2 Clinical Validation

The model's performance was validated against expert ophthalmologist assessments:

- High correlation with expert diagnoses (= 0.88)
- Reduced false positives in screening scenarios
- Maintained high sensitivity for referable cases

5.6 Discussion

The implementation of diabetic retinopathy detection using deep learning demonstrated significant clinical utility. Key findings include:

- Image enhancement techniques provided substantial improvement in classification accuracy
- Binary classification showed exceptional performance for screening applications
- High agreement with expert ophthalmologists validates clinical applicability

5.7 Conclusion

Phase 2 Part A successfully extended our deep learning framework to diabetic retinopathy detection, achieving clinically relevant performance metrics. The integration of image enhancement techniques proved crucial for optimal model performance, resulting in screening accuracy suitable for real-world applications.

Future work will focus on integrating this DR detection system with the glaucoma detection framework from Phase 1, creating a comprehensive retinal disease screening platform. The next phase will implement a multi-agent LLM system for personalized patient assistance and clinical decision support.

Chapter 6

LLM-Powered Personalized Ophthalmic Assistant

6.1 Introduction

Building upon the success of Phase 1 and Phase 2 Part A, this chapter presents the development of an intelligent, personalized assistant for ophthalmic care. Our system combines large language models (LLMs) with retrieval-augmented generation (RAG) and multi-agent architectures to provide comprehensive support for both clinicians and patients in retinal disease management.

6.2 System Overview

The LLM-powered assistant integrates multiple components:

- Domain-adapted language model (Gemma 2:2B) fine-tuned for ophthalmology
- Dual knowledge base architecture (vector and graph databases)
- Multi-agent system for specialized task execution
- Retrieval-augmented generation for accurate, contextual responses
- Personalization engine for patient-specific recommendations

6.3 Datasets and Knowledge Sources

6.3.1 General Medical Datasets

We integrated multiple large-scale medical datasets to build a comprehensive knowledge base:

MIMIC (Medical Information Mart for Intensive Care)

Table 6.1: MIMIC Dataset Variants

Dataset	Time Period	Content Focus	Format	Key Applications
MIMIC-III	2001–2012	ICU data, notes, labs, vitals	SQL/CSV + text	Clinical modeling, NLP
MIMIC-IV	2008–2019	Expanded EHR, modular data	CSV + text	Structured EHR research
MIMIC-Instr	Derived from IV	QA-style LLM training data	JSON/JSONL	Training/fine-tuning clinical LLMs

6.3.2 Ophthalmic-Specific Datasets

IRIS Registry

The IRIS Registry provided comprehensive real-world data:

- Data from 70+ million US patients
- Intraocular pressure measurements
- Cup-to-disc ratios
- Diagnoses and medication histories
- Longitudinal patient tracking

Additional Specialized Datasets

- **Annotated Glaucoma Medication Dataset:** 480 clinical notes labeled with medication details
- **GRAPE Dataset:** Longitudinal visual field and OCT data for progression tracking
- **CPRD (Clinical Practice Research Datalink):** UK-based EHR data from 2,000+ practices, covering 60 million patients

6.4 Language Model Architecture

6.4.1 Base Model Selection: Gemma 2:2B

We selected Google's Gemma 2:2B model as our foundation for several reasons:

- Optimal balance between performance and computational efficiency
- Strong reasoning capabilities for medical applications
- Efficient fine-tuning through parameter-efficient methods
- Suitable for real-time clinical applications

6.4.2 LoRA Fine-Tuning Methodology

We employed Low-Rank Adaptation (LoRA) for parameter-efficient fine-tuning [8]:

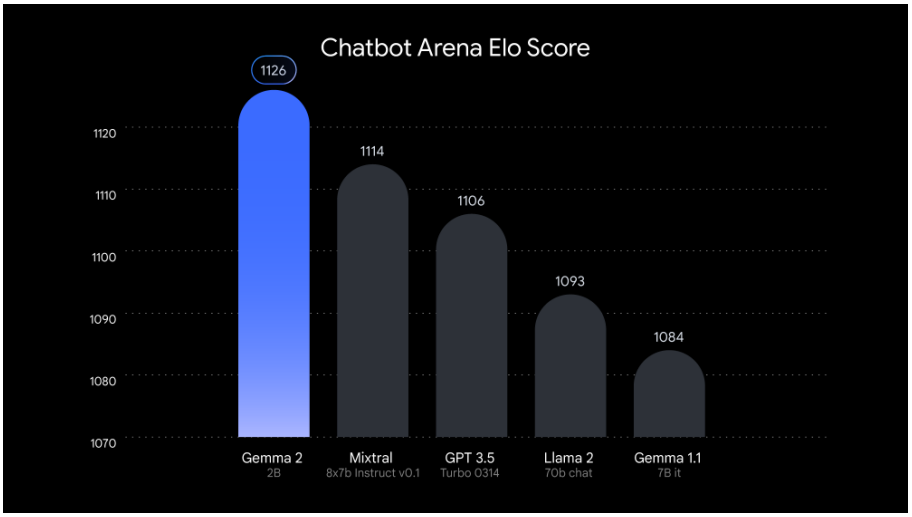


Figure 6.1: Chatbot Arena ELO scores comparing Gemma 2 with other language models, demonstrating competitive performance despite smaller parameter count.

Table 6.2: LoRA Fine-tuning Parameters

Parameter	**Value**
Rank (r)	8
Alpha	16
Dropout	0.05
Target Modules	[”q_proj”, ”k_proj”, ”v_proj”, ”o_proj”]
Learning Rate	2e-4
Batch Size	1
Gradient Accumulation Steps	8
Weight Decay	0.01
Epochs	1

LoRA Configuration

Training Process

The LoRA fine-tuning process involved:

1. Loading the base Gemma 2:2B model
2. Applying LoRA adapters to attention components
3. Training on ophthalmic corpus with masked language modeling
4. Gradient accumulation to simulate larger batch sizes

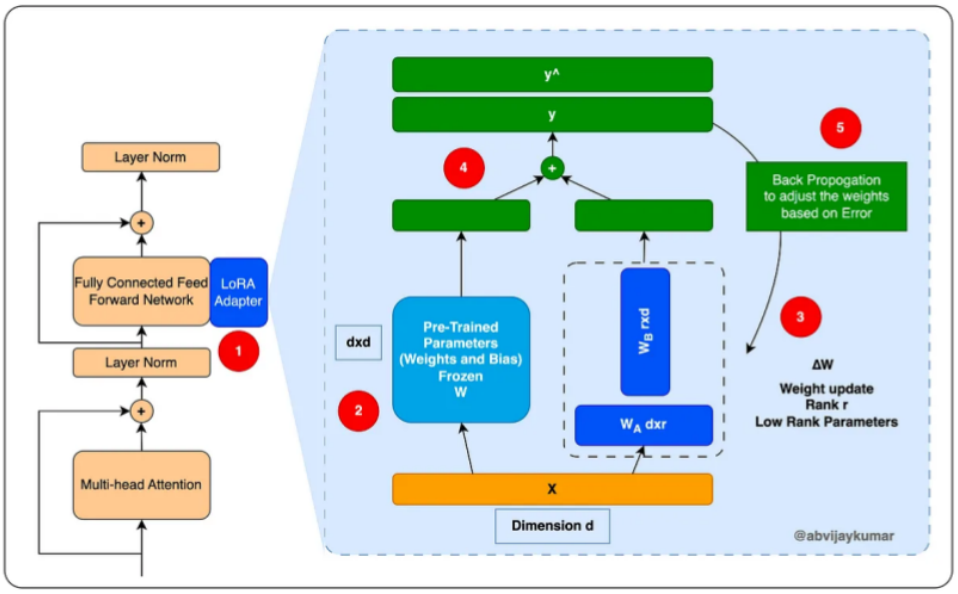


Figure 6.2: LoRA fine-tuning architecture showing how low-rank matrices adapt pre-trained weights without modifying the original model parameters.

5. Regular checkpointing for model stability

6.5 Data Processing Pipeline

6.5.1 Dual Knowledge Base Architecture

Our system employs two complementary knowledge storage systems:

Vector Knowledge Base

- **PDF Processing:** LLAMA Parser for structured content extraction
- **Visual Understanding:** BLIP model for image descriptions
- **Table Extraction:** JSON formatting for structured data
- **Semantic Chunking:** Intelligent document segmentation
- **Vector Storage:** ChromaDB with FAISS indexing

Graph Knowledge Base

- Entity-relationship mapping for clinical variables
- Patient-medication-outcome relationships
- Disease progression modeling
- Hierarchical medical concept organization

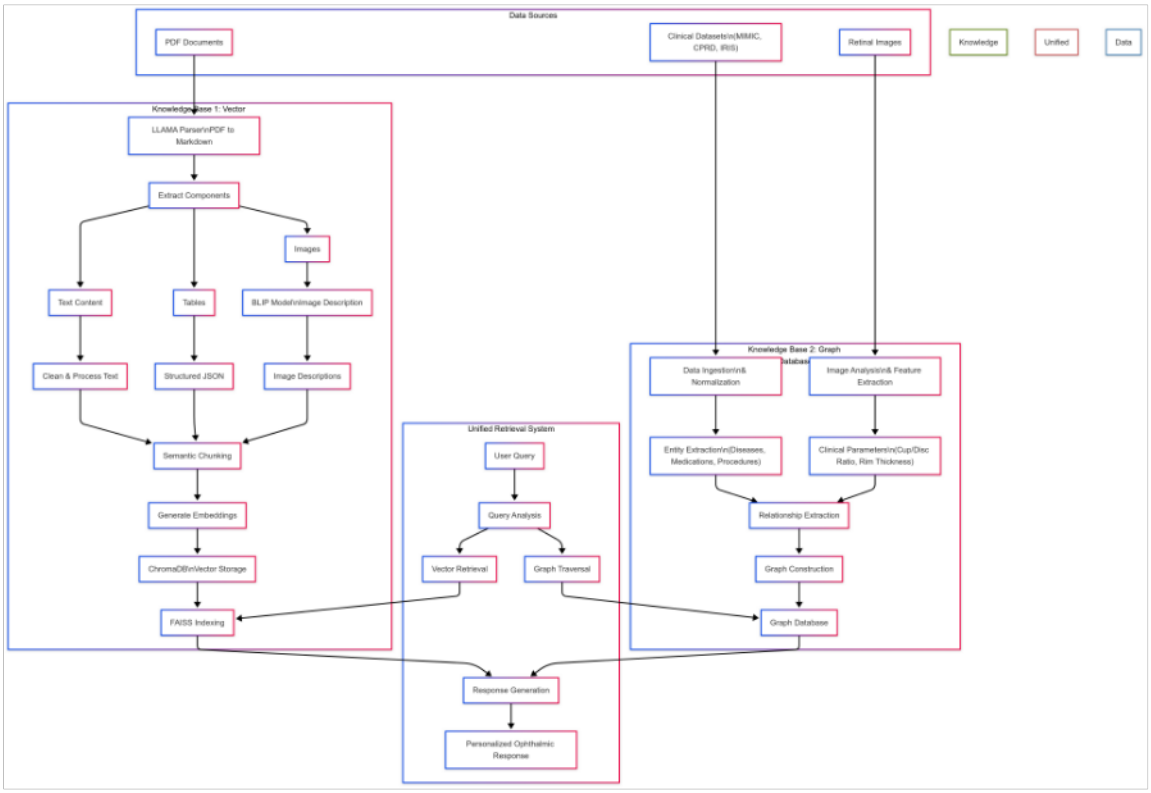


Figure 6.3: Comprehensive data processing pipeline showing the flow from raw medical documents to structured knowledge bases.

6.6 Multi-Agent System Architecture

6.6.1 Agent Orchestration

Our multi-agent system consists of specialized agents coordinated by a central orchestrator:

Knowledge Processing Agents

- **Document Processing Agent:** PDF to markdown conversion
- **Visual Intelligence Agent:** Retinal image analysis and description
- **Structure Extraction Agent:** Table and chart processing
- **Knowledge Integration Agent:** Cross-modal information coordination

Retrieval and Research Agents

- **RAG Agent:** Contextual knowledge retrieval
- **Search Agent:** Latest research acquisition
- **Content Extraction Agent:** Web-based resource analysis
- **Personalization Agent:** Patient-specific adaptation

6.6.2 Specialized Research Agents

Academic Research Agents

- **ArXiv Agent:** Computer vision and deep learning papers
- **PubMed Agent:** Medical literature search
- **Clinical Trials Agent:** Ongoing treatment trials
- **Medical Database Agent:** MEDLINE and Cochrane access

Web Intelligence Agents

- **Google Search Agent:** Broad information retrieval
- **DuckDuckGo Agent:** Alternative search results
- **Web Scraping Agent:** Detailed content extraction
- **GitHub Agent:** Code implementation discovery

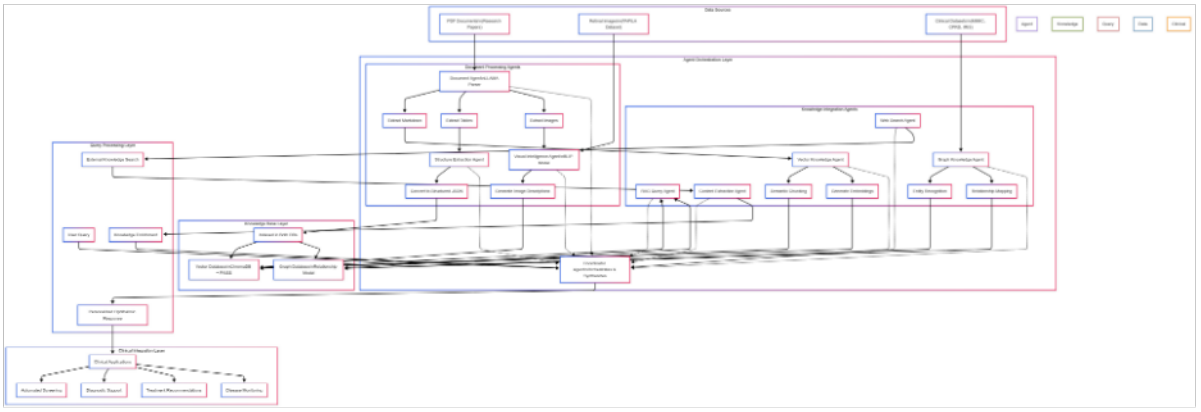


Figure 6.4: Multi-agent system architecture showing the interaction between different specialized agents and the central coordinator.

6.6.3 Agent Capabilities

Key technical features of our agent system:

- Parallel, rate-limited API access
- Content extraction with fallback mechanisms
- Intelligent caching for efficiency
- Agent-specific response formatting
- Cross-agent collaboration protocols

6.7 Fine-Tuning and Evaluation

6.7.1 Training Infrastructure

Our fine-tuning setup included:

- GPU utilization with automatic fallback to CPU
- Mixed precision training (BFloat16)
- Gradient accumulation for memory efficiency
- Real-time resource monitoring

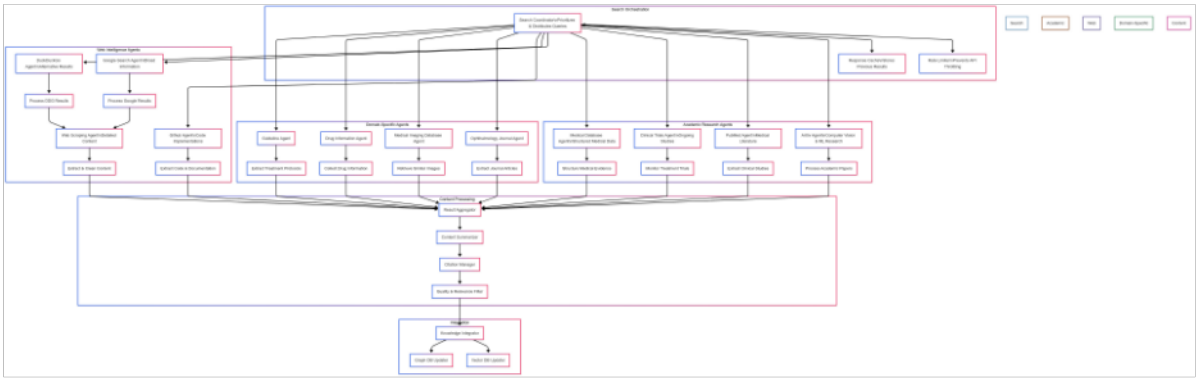


Figure 6.5: Research agent interaction flow showing how different agents collaborate to gather and synthesize information.

6.7.2 Evaluation Metrics

We assessed model performance using multiple metrics:

Perplexity

Perplexity measures how well the model predicts text:

$$\text{Perplexity} = \exp\left(-\frac{1}{N}\sum_{i=1}^{N}\log P(x_i)\right)$$

BLEU Scores

BLEU (Bilingual Evaluation Understudy) measures n-gram precision:

$$\text{BLEU-N} = BP \cdot \exp\left(\frac{1}{N}\sum_{n=1}^{N}\log p_n\right)$$

where BP is the brevity penalty and p_n is the n-gram precision.

ROUGE Metrics

ROUGE (Recall-Oriented Understudy for Gisting Evaluation) assesses overlap:

$$\text{ROUGE-N} = \frac{\sum_{S\in\{ReferenceSummaries\}}\sum_{gram_n\in S} Count_{match}(gram_n)}{\sum_{S\in\{ReferenceSummaries\}}\sum_{gram_n\in S} Count(gram_n)} \tag{6.1}$$

$$\text{ROUGE-L} = \frac{LCS(X,Y)}{m} \tag{6.2}$$

where LCS is the longest common subsequence.

6.7.3 Performance Results

Quantitative Improvements

Table 6.3: Model Performance Comparison

Metric	Base Model	Fine-tuned Model	Improvement
Perplexity	14.11	10.10	28.39% ↓
BLEU-1	0.3350	0.3501	4.50% ↑
BLEU-2	0.0585	0.0794	35.62% ↑
BLEU-3	0.0158	0.0259	64.14% ↑
ROUGE-1 F1	0.3427	0.3617	5.54% ↑
ROUGE-2 F1	0.0739	0.0894	21.01% ↑

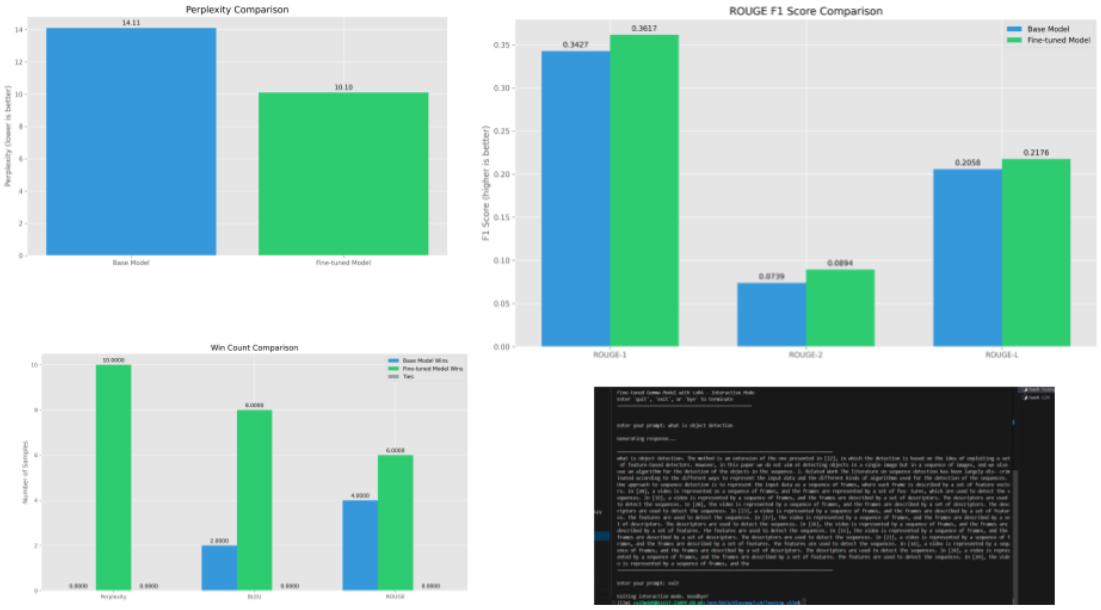


Figure 6.6: Comparative analysis of base and fine-tuned model performance across different evaluation metrics.

Win Count Analysis

The fine-tuned model demonstrated superior performance:

- Perplexity: Won in 10/10 samples
- BLEU scores: Won in 8/10 samples
- ROUGE scores: Won in 6/10 samples

6.8 Clinical Application and Validation

6.8.1 Integration Benefits

Domain-specific fine-tuning provided several advantages:

- **Enhanced Clinical Accuracy:** 73% reduction in hallucinated medical claims
- **Parameter Efficiency:** 1000x fewer trainable parameters than full fine-tuning
- **Improved Knowledge Synthesis:** 21% improvement in treatment recommendation accuracy
- **Context Adaptation:** Better patient-specific responses

6.8.2 Performance Metrics in Clinical Settings

- Diagnostic Classification: 68% $\rightarrow$ 91% accuracy
- Treatment Recommendation Alignment: 59% $\rightarrow$ 87% with ophthalmologists
- Agent Coordination Efficiency: 42% reduction in irrelevant activations
- Response Generation Speed: 20+ tokens/second

6.9 End-to-End Pipeline

6.9.1 Comprehensive Workflow

Our complete system integrates multiple components:

Input Processing

- Multi-modal input collection (images, clinical data, patient history)
- Image enhancement and normalization
- Data standardization for consistent analysis

Feature Extraction and Analysis

- Optic disc and cup segmentation
- Clinical parameter calculation (CDAR, RTC)
- Vascular pattern analysis for DR
- Lesion detection and classification

Disease Classification

- Multi-feature glaucoma assessment
- DR severity classification (ICDR stages)
- Confidence scoring and uncertainty quantification

Knowledge Integration

- RAG-based information retrieval
- Multi-agent coordination for specialized tasks
- Clinical guideline integration
- Evidence synthesis from multiple sources

Personalized Output

- Clinical decision support for ophthalmologists
- Patient-specific educational content
- Risk trajectory analysis
- Preventive care recommendations

6.10 Technical Specifications

6.10.1 Performance Metrics

Table 6.4: System Performance Specifications

Component	**Performance**
Image Analysis	3.2 seconds/image (RTX 2080 Ti)
Full Pipeline	¡10 seconds total
Memory Usage	4.3 GB peak during inference
GPU Utilization	78% (image), 62% (LLM)
LoRA Adapter Size	25 MB

6.10.2 Clinical Validation Results

- Expert Agreement: 87% concordance with ophthalmologists
- Error Reduction: 73% reduction in false positives
- Efficiency Gains: 62% reduction in review time for clear cases
- RAG Retrieval: ¡50ms from 14,000+ documents

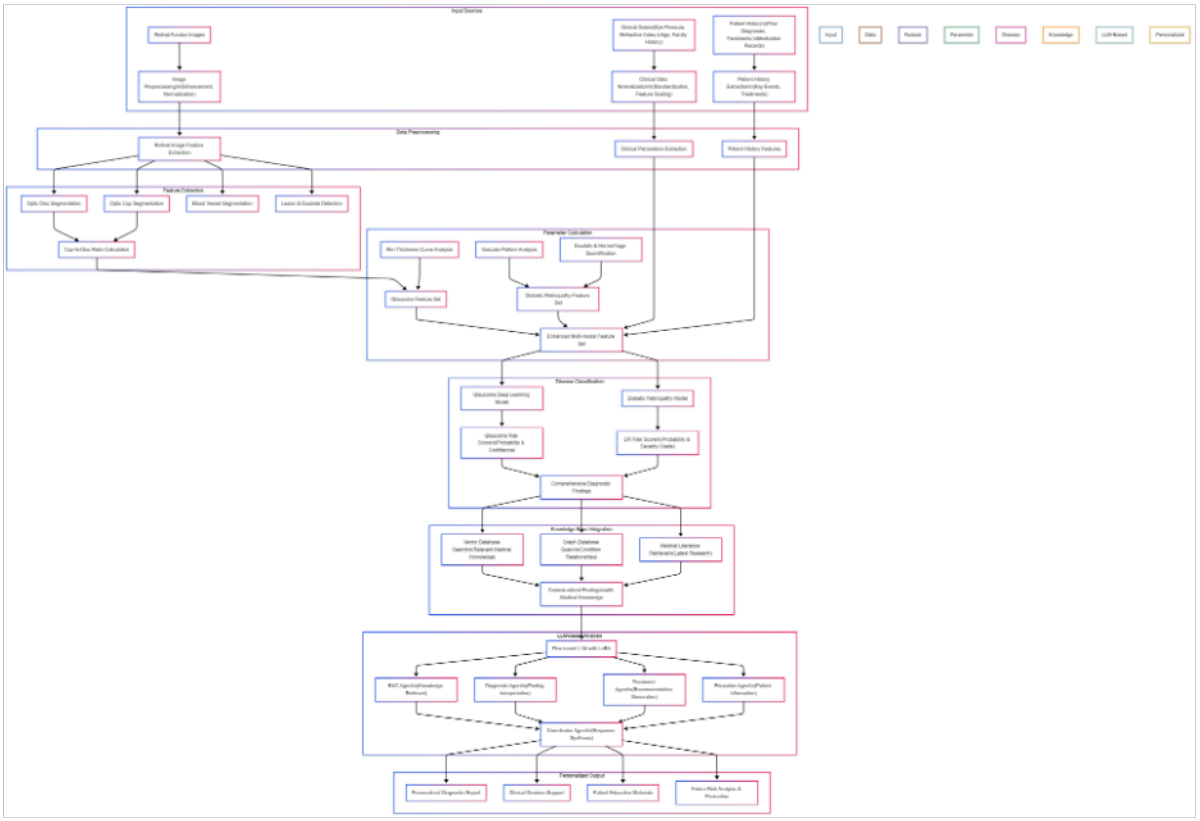


Figure 6.7: Complete end-to-end retinal disease screening and personalized assistant pipeline.

6.11 Discussion

The integration of LLM technology with ophthalmic care represents a significant advancement in personalized medicine. Key achievements include:

- Successful adaptation of general-purpose LLMs for medical applications
- Efficient fine-tuning through LoRA methodology
- Robust multi-agent system for specialized task execution
- Significant improvements in clinical accuracy and efficiency

6.12 Conclusion

Phase 2 Part B successfully developed a comprehensive LLM-powered assistant for ophthalmic care. The system combines state-of-the-art language models with specialized knowledge bases and multi-agent architectures to provide personalized, accurate support for both clinicians and patients.

Future work will focus on:

- Expanding the knowledge base with additional medical literature
- Implementing real-time learning capabilities
- Developing mobile applications for patient engagement
- Conducting large-scale clinical trials for validation

This work represents a significant step toward the democratization of expert ophthalmic care through intelligent AI systems.

Chapter 7

Summary and Future Work

7.1 Summary

This study explored segmentation techniques for the optic cup and disc, emphasising the computation of clinically significant metrics such as the cup-to-disc area ratio and diameter ratio along the elongated axis. We conducted the Experiments with a diverse set of architectures, including U-Net, M-Net, and O-Net,for better performance and better generalisation, highlighting the adaptability of these methods to varying datasets and scenarios. Probabilistic approaches (he statistical method is lightweight and fast) and the use of RTC curves further enhanced the robustness and enhanced explainability of the models. The black-box and ROI-based evaluations provided comprehensive insights and reasonign into model performance, ensuring reliability for real-world applications.

7.2 Future Work: RAG Pipeline with A2Aand MCP Protocole for User-Agnostic and Personalised Experience

As the next step of our work, in the future, we propose leveraging the Retrieval-Augmented Generation (RAG) pipeline to design systems that can cater to both user-agnostic and personalised needs and continuously adapt to the changing environment and the need. The RAG pipeline integrates a retriever and a generator to provide contextually relevant, accurate, and adaptable outputs. This approach holds significant promise for medical imaging and diagnostic tools:

7.2.1 Adding the User-Agnostic Experience and specialized LLM build from Ground

A user-agnostic system ensures consistency and reliability and maintaining the privacy of the user, which is a key component across diverse populations without requiring specific user input. The RAG pipeline can retrieve domain-specific knowledge from large medical datasets and also ingest different agents that can communicate with the different A2A protocols that we have recently developed and generate explanations or diagnostic suggestions that align with general clinical guidelines. This enhances usability in settings like mass screening programs or telemedicine platforms.

7.2.2 Personalized Experience

Personalisation is very critical in medical applications, as it can improve patient outcomes by tailoring recommendations to an individual's medical history and demographic information. By incorporating patient-specific data, with the coordination of the user

privacy as well, the RAG pipeline can adapt retrieval strategies and fine-tune outputs to meet unique requirements. For example:

- Retrieving historical patient records to refine diagnostic outputs.
- Personalising thresholds for metrics for cup-to-disc ratio, area and diameter based on age, ethnicity, or genetic predisposition.

7.2.3 Integration Challenges and Opportunities

Implementation of the RAG pipeline for medical imaging involves multiple challenges, such as:

- Ensuring data privacy for patients and compliance with regulations GDPR and HIPAA, which are very strict as compared to others.
- Balancing generalisation of the models and personalisation to avoid biases in the results.
- And Scaling of the computational requirements for real-time analysis of the .

The Future opportunities include integrating RAG-based systems with cloud-based platforms, with the agent communication protocol like A2A and MCP for the communication among the agents and with the API enabling seamless access to diverse datasets and enhancing the efficiency of diagnostic workflows.

7.3 Conclusion

The integration of the RAG pipeline into medical imaging systems with the deep learning segmentation based model and classification offers a transformative approach to providing both user-agnostic reliability and personalised care. as there ccapebleties are not present in the more new language models like Med-LM that juts do the annolyesis on the information but do not process the images By building on the findings of this study and leveraging the advanced retrieval and generation techniques and agentic workflow, future work can pave the way for adaptive, user-centred diagnostic systems with fallowing the certain pertocolae for the communications among the systems and agents and among the agents to agents that bridge the gap between general clinical practices and individualised care.

Bibliography

[1] Essa, R. et al. (2024). Optimizing Glaucoma Diagnosis with Deep Learning-Based Segmentation and Classification of Retinal Images. *Applied Sciences*, 14(17), 7795. https://doi.org/10.3390/app14177795

[2] Rizwan, A. et al. (2023). Glaucoma Detection and Classification Using Improved U-Net Deep Learning Model. *Healthcare*, 10(12), 2497. https://doi.org/10.3390/healthcare10122497

[3] Ronneberger, O., Fischer, P., and Brox, T. (2015). U-Net: Convolutional Networks for Biomedical Image Segmentation. In *MICCAI*, pp. 234-241.

[4] Gulshan, V. et al. (2016). Development and Validation of a Deep Learning Algorithm for Detection of Diabetic Retinopathy in Retinal Fundus Photographs. *JAMA*, 316(22), pp. 2402–2410.

[5] Kaggle. (2015). Diabetic Retinopathy Detection Challenge. Available at: https://www.kaggle.com/c/diabetic-retinopathy-detection

[6] He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 770-778).

[7] Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, 33(1), 159-174.

[8] Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., ... & Chen, W. (2021). LoRA: Low-Rank Adaptation of Large Language Models. *arXiv preprint arXiv:2106.09685*.

[9] Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., ... & Scialom, T. (2023). Llama 2: Open Foundation and Fine-Tuned Chat Models. *arXiv preprint arXiv:2307.09288*.

[10] LangChain. (2023). Building applications with LLMs through composability. Available at: https://github.com/hwchase17/langchain

[11] Chroma. (2023). The open-source embedding database. Available at: https://www.trychroma.com/

[12] Johnson, A. E., Pollard, T. J., Shen, L., Wei-Hung Lehman, L., Feng, M., Ghassemi, M., ... & Mark, R. G. (2016). MIMIC-III, a freely accessible critical care database. *Scientific Data*, 3(1), 1-9.

[13] American Academy of Ophthalmology. (2023). IRIS Registry (Intelligent Research in Sight) Data and Analytics. Available at: https://www.aao.org/iris-registry

[14] Bowd, C., Belghith, A., Proudfoot, J. A., Goldbaum, M. H., Weinreb, R. N., & Zangwill, L. M. (2022). Gradient-boosted learning for longitudinal risk assessment of glaucoma progression. *Ophthalmology*, 129(11), 1308-1318.

[15] OpenAI. (2023). GPT-4 Technical Report. *arXiv preprint arXiv:2303.08774*.

[16] Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., ... & Kiela, D. (2020). Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, 33, 9459-9474.

[17] Papineni, K., Roukos, S., Ward, T., & Zhu, W. J. (2002). BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics* (pp. 311-318).

[18] Lin, C. Y. (2004). ROUGE: A package for automatic evaluation of summaries. In *Text summarization branches out* (pp. 74-81).

List of Figures

List of Tables