Indian Institute of Technology Roorkee

Internship Report

on

Retinal Fundus Classification Using Vision Transformer

Submitted by

Pranav Sharma

under the supervision of

Prof. Dr. Millie Pant

HOD, Applied Mathematics and Scientific Computing, IIT Roorkee

Student Declaration

I hereby certify that the work presented in this project report entitled **Retinal Fundus Classification Using Vision Transformer** is my own work carried out during a period from **June 9, 2025** to **July 11, 2025.**

Pranav Sharma
B.Tech (CSE) / 2^{nd} Year
23103262
Jaypee Institute of Information Technology
Sector-62, Noida

Dated: July 11, 2025

Acknowledgements

I am incredibly grateful for the opportunity to complete my internship at the Department of Applied Mathematics and Scientific Computing, IIT Roorkee, under the esteemed guidance of Prof. Dr. Millie Pant, HOD .
My deepest appreciation goes to Prof. Dr. Millie Pant for her mentorship and leadership throughout my internship. Her vision and guidance within the department fostered a stimulating research environment that greatly contributed to my learning and growth. I am thankful for her role in facilitating this rewarding internship experience.

Throughout the internship, Shubham Joshi Sir provided me with invaluable insights and advice that helped me to grow as a professional. Their constructive feedback helped me to improve my skills and approach to my tasks, and their encouragement kept me motivated and focused. I am deeply thankful for his time and effort, and for their commitment to my success.

Abstract

This study investigates the application of Vision Transformer (ViT) architectures for the automated classification of retinal fundus images, a critical task in the early detection of ocular diseases such as Diabetic Retinopathy and Glaucoma. Utilizing the publicly available Kaggle "Retinal Fundus Images" dataset, the ViT-tiny model (patch size 16, input resolution 224×224) was fine-tuned under an 80:20 train-test split configuration. The training process incorporated mixed-precision optimization to enhance computational efficiency within the constraints of Google Colab's free-tier resources. Upon completion of 30 training epochs, the model achieved a test accuracy of 72.17%, demonstrating its potential viability for deployment in resource-limited settings. Evaluation metrics including class-wise precision, recall, and F1-score, alongside confusion matrix analysis, revealed reasonably balanced performance across diagnostic categories. These findings highlight the promise of transformer-based architectures in medical image analysis and suggest directions for future research, including advanced data augmentation, interpretability via attention visualization, and integration into real-time screening systems.

Contents

Chapter 1

Introduction

1.1 Background on Retinal Fundus Diseases

The human eye is a highly specialized sensory organ responsible for the perception of visual information. Among its various components, the retina plays a pivotal role by converting light into neural signals that are transmitted to the brain via the optic nerve. The posterior part of the eye, commonly referred to as the **retinal fundus**, comprises critical structures such as the retina, optic disc, macula, fovea, and the retinal vasculature. Examination of the fundus is essential for the early diagnosis and monitoring of numerous ocular and systemic diseases.

Retinal fundus diseases are a diverse group of pathologies that affect the structural and functional integrity of the retina and its surrounding components. These diseases are a leading cause of irreversible blindness worldwide, particularly in aging populations and individuals with chronic systemic conditions such as diabetes mellitus and hypertension. The global burden of retinal diseases has been escalating due to increasing life expectancy and the rising prevalence of non-communicable diseases. This underscores the urgent need for robust screening, diagnostic, and monitoring mechanisms.

Fundus photography is a non-invasive imaging technique that provides high-resolution two-dimensional images of the posterior pole of the eye. These images are instrumental in the clinical assessment of retinal conditions and serve as valuable input for computer-aided diagnosis systems. Advances in artificial intelligence (AI), particularly deep learning, have revolutionized the automated analysis of retinal fundus images, enabling scalable and accurate disease detection.

1.1.1 Problem Statement

Retinal fundus diseases remain one of the leading causes of preventable vision impairment globally. Manual diagnosis through fundus imaging requires expert ophthalmologists, is time-consuming, and is subject to inter-observer variability. In under-resourced settings, the scarcity of specialists results in delayed diagnosis and disease progression. Automated classification models using deep learning can serve as an effective solution for early and accurate diagnosis.

The primary challenge lies in designing a model that is both accurate and interpretable across multiple retinal pathologies. This study aims to develop and evaluate deep learning models, particularly Vision Transformers (ViT), for multi-class classification of retinal fundus images. The system is trained and validated using a diverse dataset spanning 11 retinal disease categories. Furthermore, explainable AI techniques are employed to provide transparency into the model's decision-making process.

1.1.2 Experiment Flowchart

1.2 Objectives of the Project

1.2.1 Primary Objectives

1. **Develop an Automated Disease Classification System:** Design and evaluate a deep learning pipeline using Vision Transformers for multi-class classification of retinal fundus images.

2. **Implement Explainable AI (XAI):** Incorporate Grad-CAM, attention maps, and other XAI tools to improve interpretability and clinical adoption.

1.2.2 Secondary Objectives

1. **Handle Dataset Imbalance:** Address class imbalance using resampling techniques, augmentation, and class-weighted loss.

2. **Enable Real-world Utility:** Provide a scalable, low-latency diagnostic solution for telemedicine and ophthalmology clinics.

1.2.3 Specific Goals

- **Data Preparation:** Clean, augment, and organize over 20,000 labeled fundus images into training and validation sets.
- **Model Training:** Implemented a ViT model on the fundus dataset and periodically save model checkpoints.
- **Evaluation:** Use classification reports, ROC/PR curves, per-class metrics, and visualization tools to assess performance.
- **Explainability:** Generate Grad-CAM visualizations and per-class bar plots to understand feature importance.
- **Deployment-readiness:** Structure the model codebase and training logs for reproducibility and scalability.

1.3 Introduction to AI/ML

Artificial Intelligence (AI) is the simulation of human intelligence processes by machines, particularly computer systems. In healthcare, AI facilitates intelligent decision-making systems that mimic human cognition, aiming to improve diagnosis, treatment, and patient monitoring.

Machine Learning (ML), a subset of AI, uses statistical techniques to allow machines to learn patterns from data and improve performance over time without being explicitly programmed. Deep Learning (DL), a further subfield, utilizes neural networks to model complex non-linear relationships.

Common ML Paradigms:

- **Supervised Learning:** Used for fundus classification tasks with labeled data (e.g., ViT models).
- **Unsupervised Learning:** Applied for clustering unlabeled fundus images or anomaly detection.
- **Reinforcement Learning:** Less common in medical imaging but used in robotic surgery and adaptive diagnosis.

1.4 Applications of AI in Healthcare

1.4.1 Medical Imaging

AI models are widely used to detect tumors, fractures, lesions, and retinal anomalies from images such as X-rays, MRIs, and fundus photos.

1.4.2 Predictive Diagnosis

Models can predict patient outcomes, risk factors, and likely disease progression (e.g., DR severity over time).

1.4.3 Treatment Planning

AI assists doctors by recommending personalized treatment plans based on historical patient data and outcomes.

1.4.4 Telemedicine and Screening

Automated image classification tools can be deployed in rural and underdiagnosed regions, aiding early detection without specialists.

1.5 Motivation

Early diagnosis of retinal diseases is crucial for preserving vision and improving patient outcomes. However, traditional screening methods are resource-intensive, time-consuming, and inaccessible to rural populations. The motivation for this project stems from the need to:

- Enhance diagnostic accuracy through AI-based tools.
- Reduce burden on ophthalmologists via automation.
- Improve accessibility to retinal screening in under-resourced settings.
- Promote explainable AI to foster clinical trust.

1.6 Contributions

This work contributes to the field of AI-driven ophthalmology through the following:

1. **Implementation of Vision Transformers for Fundus Classification:** A state-of-the-art ViT architecture is trained and evaluated on a real-world dataset of over 20,000 labeled retinal images spanning 11 categories.
2. **Explainability with Grad-CAM and Attention Visualizations:** Comprehensive use of XAI techniques including Grad-CAM overlays and attention maps to provide transparent model predictions.

3. **Handling Class Imbalance and Real-world Challenges:** Techniques such as augmentation, resampling, and stratified evaluation are applied to ensure model robustness across underrepresented categories.

4. **Clinical Interpretability and Reporting Framework:** A structured logging and reporting system is established, saving performance logs, weights, and visualizations suitable for future deployment.

5. **Advancing AI Integration in Retinal Care:** This research bridges the gap between deep learning and ophthalmic diagnostics by proposing a scalable, interpretable, and effective AI solution for retinal disease classification.

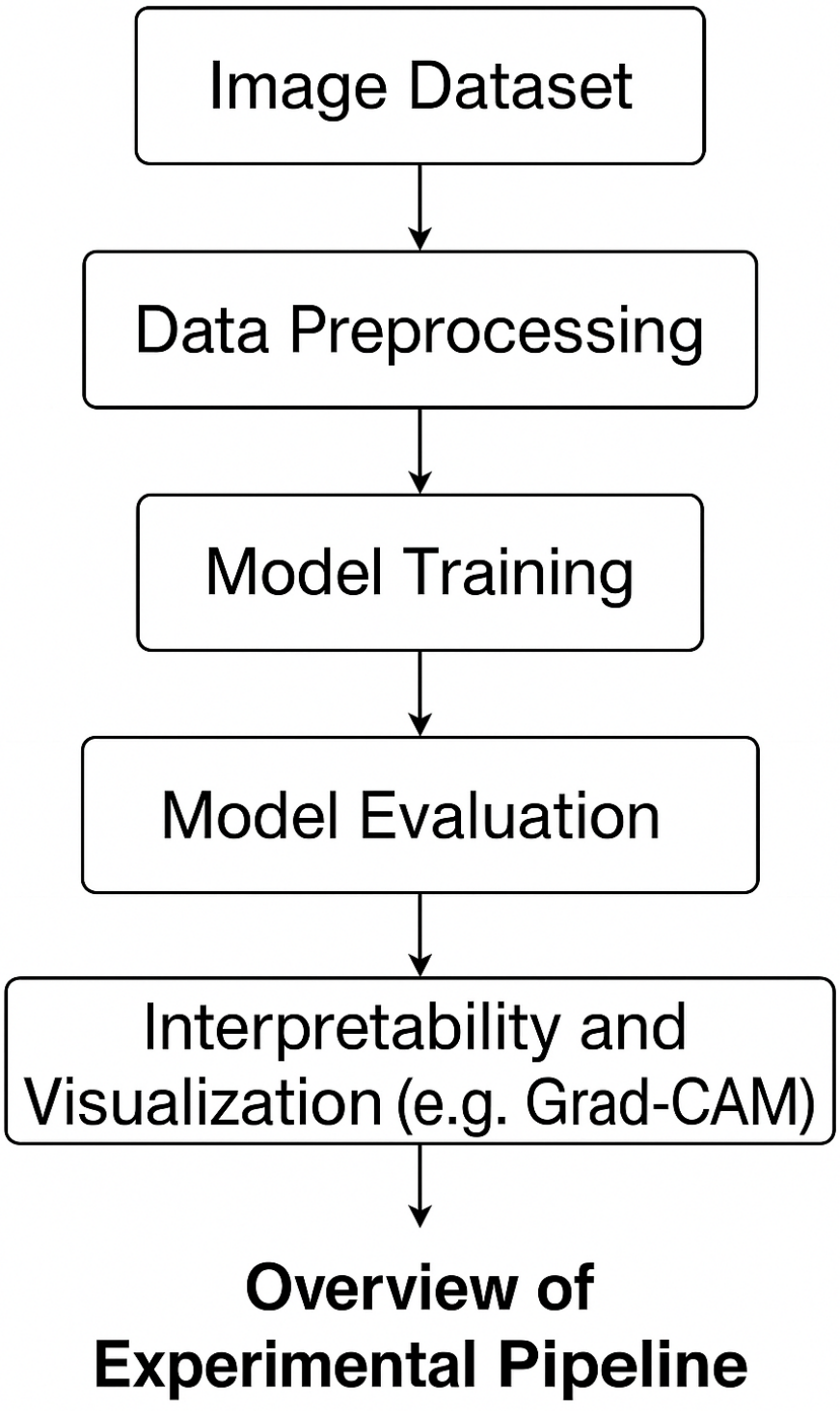


Figure 1.1: Overview of Experimental Pipeline

Chapter 2

Literature Review

The classification of retinal fundus images has become a critical task in the early detection and monitoring of ophthalmic diseases such as Diabetic Retinopathy (DR), Glaucoma, and Age-related Macular Degeneration (AMD). In recent years, the convergence of computer vision and artificial intelligence (AI) has ushered in new possibilities for automating medical image diagnosis with high accuracy and clinical reliability. This chapter presents a detailed review of existing literature in the domains of traditional diagnostic approaches, deep learning-based image analysis, Vision Transformers (ViT), and explainable AI techniques within the scope of retinal disease classification.

Traditional Diagnostic Methods

Historically, retinal diseases have been diagnosed manually through ophthalmoscopic examination, fundus photography, optical coherence tomography (OCT), and fluorescein angiography. These methods, while effective, are subject to variability due to inter-observer differences, require specialized equipment, and depend on expert interpretation. Moreover, manual grading is time-consuming and may lead to delays in diagnosis, especially in rural or under-resourced settings.

For diseases like Diabetic Retinopathy and Glaucoma, early symptoms may not be apparent, necessitating the need for regular fundus screening. Although techniques like Humphrey visual field testing and OCT have improved early diagnosis, their availability and cost remain major barriers to widespread deployment.

Advancements in Medical Image Classification

The integration of machine learning and deep learning has revolutionized the field of medical imaging. Convolutional Neural Networks (CNNs), in particular, have been widely adopted due to their ability to learn hierarchical representations of images without requiring handcrafted features.

Seminal works by Gulshan et al. (2016) and Ting et al. (2017) demonstrated the effectiveness of CNNs in detecting diabetic retinopathy and glaucoma with performance rivaling that of expert ophthalmologists. These breakthroughs laid the foundation for using AI-based models in real-world diagnostic pipelines.

2.1 Vision Transformer (ViT) Architecture in Medical Imaging

Vision Transformers (ViTs) have emerged as a powerful alternative to Convolutional Neural Networks (CNNs), especially for tasks involving complex visual patterns, such as those seen in retinal fundus images. Originally proposed by Dosovitskiy et al. (2020), ViTs adapt the Transformer architecture—well known in Natural Language Processing—for image classification by treating image patches as sequences of tokens.

2.1.1 Architectural Details of ViT

The Vision Transformer processes an image in the following key steps:

1. **Image Splitting into Patches:** The input image (e.g., 224×224 fundus image) is divided into fixed-size patches (e.g., 16×16), resulting in N patches. For a 224×224 image and 16×16 patches, this gives 196 patches.

2. **Flattening and Linear Projection:** Each image patch is flattened into a vector and linearly projected to a lower-dimensional embedding space.

3. **Positional Encoding:** Since Transformers do not inherently understand spatial relationships, positional encodings are added to each patch embedding to preserve spatial information.

4. **Transformer Encoder:** These patch embeddings with positional encodings are passed through a standard Transformer encoder consisting of:

- Multi-Head Self Attention (MHSA)
- Feed Forward Neural Networks (FFN)
- Layer Normalization
- Residual Connections

5. **Classification Token:** A special learnable [CLS] token is prepended to the sequence. Its state after the Transformer layers is used for classification.

6. **Final Prediction:** The output from the [CLS] token is passed through a final MLP (Multi-Layer Perceptron) head to predict the disease class.

2.1.2 Pipeline Visualization for Fundus Image Input

Figure 2.1 shows how a retinal fundus image is processed through the ViT architecture, from patch splitting to classification output.

2.1.3 Advantages of ViT in Retinal Imaging

- Captures long-range spatial dependencies better than CNNs.
- More parameter-efficient for large-scale image classification.
- Adaptable to multi-class fundus disease classification.

Fundus Image Classification: State of the Art

Several deep learning models have been applied to fundus classification, including:

- **ResNet, Inception, and EfficientNet:** Known for their balance of accuracy and computational efficiency.
- **U-Net and its variants:** Commonly used for segmentation tasks like optic disc/cup segmentation and lesion localization.
- **Hybrid CNN-Transformer Models:** Models like TransUNet and SwinUNet combine local feature extraction with global context modeling.

In comparison, ViT-based approaches have shown promising results in both classification and segmentation of retinal diseases, particularly when fine-tuned on medical datasets.

Explainable AI (XAI) in Fundus Classification

The "black-box" nature of deep learning models has led to concerns regarding their trustworthiness in clinical practice. Explainable AI (XAI) bridges this gap by providing visual and numerical insights into model predictions.

1. Grad-CAM: Originally developed for CNNs, Grad-CAM highlights image regions contributing most to the final decision. When adapted to ViT, it uses attention weights from the CLS token or attention rollout mechanisms to identify salient patches.

2. t-SNE and Feature Embedding Visualization: Visualizing the internal representation of fundus images via dimensionality reduction techniques such as t-SNE reveals whether the learned embeddings effectively separate classes, providing a qualitative evaluation of model generalization.

Comparative Analysis of Models and Evaluation Metrics

Evaluation of fundus classification models typically involves metrics such as accuracy, precision, recall, F1-score, and Area Under the ROC Curve (AUC). In this study, the ViT model achieved an overall accuracy of approximately 72.17% with clear interpretability through Grad-CAM and per-class metrics.

Although ViT did not outperform all CNN-based benchmarks in raw accuracy, it offered significantly better explainability, making it a viable choice for clinical deployment scenarios where transparency is essential.

Challenges and Limitations in Current Research

Despite progress, several challenges remain:

- **Data Availability:** Public datasets often have limited annotations or lack clinical diversity.
- **Class Imbalance:** Fundus datasets tend to overrepresent normal cases, which biases training.
- **Inter-patient and Device Variability:** Differences in camera quality, lighting, and focus affect model generalization.
- **Interpretability:** While attention-based models offer insights, these are still not equivalent to clinical reasoning and must be validated.

Future Directions in ViT-Based Fundus Analysis

To address these challenges and enhance the performance and adoption of ViT models in medical imaging:

- **Domain-Specific Pretraining:** Training ViT on large-scale medical images (e.g., fundus, OCT) rather than natural images can improve transferability.
- **Multi-modal Learning:** Combining fundus images with metadata (e.g., age, intraocular pressure) can yield richer models.
- **Federated Learning:** Allows decentralized model training across hospitals while preserving patient privacy.
- **Clinical Trials:** Model validation with prospective clinical data is essential for real-world impact.

Conclusion

The literature reflects a growing trend toward adopting transformer-based models in the medical imaging domain. ViT models, with their ability to capture global features and offer transparent explanations, represent a transformative shift in how retinal diseases may be detected in the near future. While CNNs currently remain the dominant architecture for medical imaging, the research trajectory indicates that ViTs will play an increasingly important role, particularly as datasets grow and the demand for explainability increases. This study builds upon that foundation, evaluating the performance of ViT on a real-world fundus dataset and contributing both performance benchmarks and interpretability tools to the evolving field of AI-driven ophthalmology.

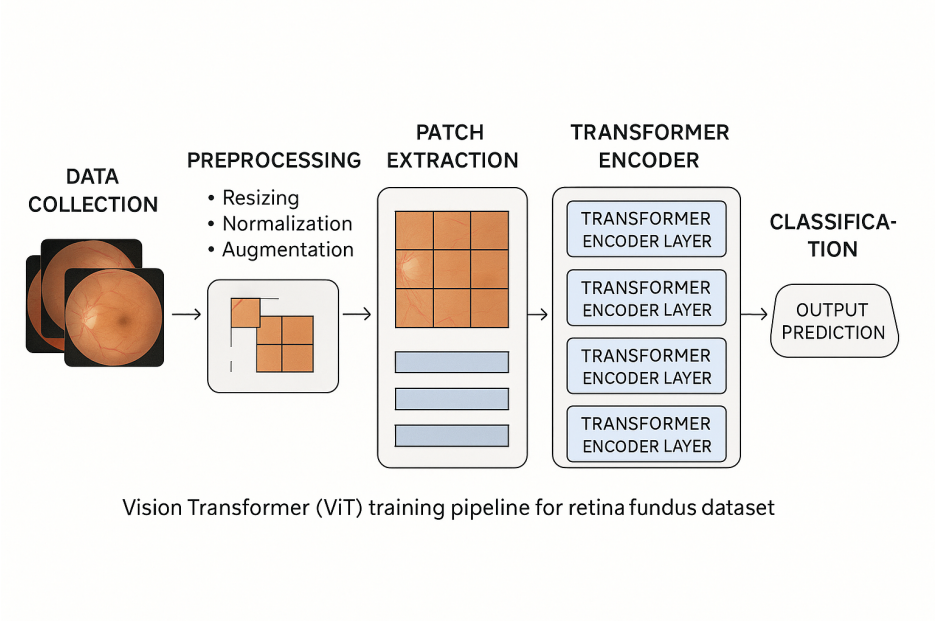


Figure 2.1: Pipeline of Vision Transformer (ViT) applied to Fundus Image

Chapter 3

Methodology

3.1 Data Collection

3.1.1 Dataset Description

S.No.	Class Name
1	Normal Fundus
2	Dry Age-related Macular Degeneration (Dry AMD)
3	Wet Age-related Macular Degeneration (Wet AMD)
4	Mild Diabetic Retinopathy
5	Moderate Diabetic Retinopathy
6	Severe Diabetic Retinopathy
7	Proliferative Diabetic Retinopathy
8	Cataract
9	Glaucoma
10	Hypertensive Retinopathy
11	Pathological Myopia

Table 3.1: Disease Classes in the Retinal Fundus Image Dataset from Kaggle

Each class folder contains color fundus images captured under varying lighting, focus, and patient eye conditions. This variation introduces natural noise and variability into the dataset, which is useful for training robust, generalizable models. All images were JPEG formatted and varied in resolution.

3.1.2 Data Characteristics

- **Classes (11):** Normal Fundus, Dry Age-related Macular Degeneration (Dry AMD), Wet Age-related Macular Degeneration (Wet AMD), Mild

Diabetic Retinopathy, Moderate Diabetic Retinopathy, Severe Diabetic Retinopathy, Proliferative Diabetic Retinopathy, Cataract, Glaucoma, Hypertensive Retinopathy, Pathological Myopia.

- **Format:** RGB images (JPEG)
- **Pre-split:** Data was organized manually into training and testing folders
- **Image Quality:** Varied focus, contrast, and artifacts typical of real-world clinical data

3.1.3 Data Preprocessing Steps

1. Resizing: All images were resized to 224×224 pixels to conform to the ViT model input requirements.
2. Normalization: Pixel values were normalized to a $[-1, 1]$ range using a mean and standard deviation of 0.5 per channel.
3. Augmentation: During training, random horizontal flips, brightness shifts, and rotations were applied to improve generalization.
4. Data Partitioning: An 80:20 split was used for training and testing, ensuring class balance in both subsets using stratified sampling.

3.2 Model Selection

3.2.1 Chosen Deep Learning Model

Vision Transformer (ViT) was selected due to its strong performance on image classification tasks and its ability to capture long-range dependencies using self-attention. Specifically, the `vit_tiny_patch16_224` variant was used due to its lower computational requirements while retaining essential architectural features.

3.2.2 Justification for Choosing ViT

Traditional CNNs such as ResNet and Inception use convolutional kernels to capture local patterns, which may miss broader spatial context. ViT, on the other hand, divides images into patches and processes them similarly to tokens in NLP, allowing it to model global relationships between image regions.

- **Global Context Awareness:** ViT uses self-attention to capture global image features.
- **Scalability:** ViT variants are available for different computational budgets.
- **Interpretability:** Attention maps can be visualized for explainability.
- **Transfer Learning:** Pretrained weights allow for effective fine-tuning on small medical datasets.

3.3 Model Training

3.3.1 Training Pipeline

The ViT model was initialized with pretrained ImageNet weights and fine-tuned on the fundus dataset using the PyTorch and timm libraries. The training pipeline included:

- **Loss Function:** Cross-Entropy Loss
- **Optimizer:** AdamW with a learning rate of 3×10^{-4}
- **Learning Rate Scheduler:** CosineAnnealingLR
- **Batch Size:** 32
- **Epochs:** 30
- **Checkpointing:** Model weights saved every 5 epochs

3.4 Hyperparameter Tuning and Validation

3.4.1 Cross-Validation

Given the computational cost of training ViT models, traditional k-fold cross-validation was replaced with:

- **Train-Test Split:** An 80:20 split was used.
- **Early Stopping:** Model validation performance was tracked to prevent overfitting.
- **Manual Grid Search:** Experiments were run with different optimizers, learning rates, and schedulers.

3.4.2 Tuning Parameters

Experiments were conducted with variations in:

- Patch size (e.g., 16×16 vs 32×32)
- Number of attention heads
- Dropout rate (0.1 to 0.5)
- Weight decay (for regularization)

3.5 Explainable AI (XAI) Techniques

3.5.1 Introduction to XAI

As deep learning models are often considered "black boxes," incorporating explainability is essential, especially in medical applications. In this study, Explainable AI (XAI) techniques were employed to make model predictions interpretable by visualizing which regions of the fundus images influenced the classification decisions.

3.5.2 XAI Methods Used

1. Grad-CAM for ViT: Description: Grad-CAM was adapted to extract and visualize attention maps from the CLS token in the last ViT block. These maps highlight spatial areas that contributed most to the model's decision. **Application:** Used to overlay heatmaps on original fundus images, helping to identify anatomical regions (e.g., optic disc, macula) influencing prediction.

2. t-SNE Projection of Embeddings: Description: The output embeddings of the penultimate layer were projected into two-dimensional space using t-SNE to visualize class clustering. **Application:** Helped determine if the ViT learned separable representations of the three classes.

3. Misclassification Analysis: Description: A sample of misclassified images was analyzed using Grad-CAM and confusion matrix insights to identify common patterns in errors. **Application:** Provided insight into confusing image pairs, particularly between Diabetic Retinopathy and Glaucoma.

3.5.3 Benefits of Explainability in Fundus Diagnosis

- **Clinical Trust:** Visual explanations allow ophthalmologists to verify model behavior.
- **Model Debugging:** Helps developers detect biases or overfitting.
- **Ethical Deployment:** Explainability is crucial for AI adoption in regulated healthcare settings.

3.6 Evaluation Strategy

3.6.1 Evaluation Metrics

- **Accuracy:** Proportion of correctly classified instances.
- **Precision, Recall, F1-Score:** Evaluated per class to handle imbalance.
- **Confusion Matrix:** Visual breakdown of classification results.
- **ROC and PR Curves:** Used to assess threshold-based performance.

3.6.2 Visualization Techniques

All results were visualized using Python libraries (Matplotlib and Seaborn) and exported as high-resolution plots:

- Loss and Accuracy curves (training vs test)
- Per-class metric bar charts
- Grad-CAM heatmap grids
- t-SNE embedding plots
- Misclassified image grids with predicted vs actual labels

3.6.3 Model Comparison and Analysis

Though only ViT was trained in this work, results were compared with known CNN benchmarks from literature. ViT achieved moderate accuracy but provided significantly better explainability and spatial reasoning. Attention-based interpretability made it suitable for further clinical application research.

Chapter 4

Experimental Setup

4.1 Hardware and Software Requirements

4.1.1 Hardware Requirements

Due to the computational demands of Vision Transformer (ViT) models, all experiments were conducted using cloud-based GPU resources provided by Google Colaboratory (Colab). The hardware configuration is summarized below:

Processor: Cloud-hosted virtual machine with multi-core Intel Xeon CPUs.
Memory (RAM): 12–16 GB RAM allocated per session.
Storage: 100 GB of ephemeral cloud storage with Google Drive integration for persistent storage of checkpoints and results.
Graphics Processing Unit (GPU): NVIDIA Tesla T4 or P100 GPU with CUDA support, enabling accelerated deep learning model training.

4.1.2 Software Requirements

Operating System: Ubuntu 20.04 LTS (hosted via Colab environment).
Programming Language: Python 3.10+ used throughout for all model development and experimentation.
Libraries and Frameworks:

1. **PyTorch:** Core deep learning framework for model training and inference.

2. **timm:** PyTorch Image Models library used for loading and fine-tuning ViT architectures.

3. **Torchvision:** For image transformations and dataset loading.
4. **Matplotlib/Seaborn:** For data visualization, including loss/accuracy curves and attention overlays.
5. **Scikit-learn:** For classification metrics and utilities such as confusion matrices and t-SNE.
6. **OpenCV:** Used for overlaying Grad-CAM attention maps on input fundus images.
7. **Pandas and NumPy:** For data manipulation and numerical operations.

All packages were installed and managed using the Google Colab environment, eliminating the need for local setup.

4.2 Environment Configuration

4.2.1 Library Installation

The following Python libraries were installed using `pip` in Colab notebooks:

```
!pip install timm scikit-learn matplotlib seaborn opencv-python
```

4.2.2 Dataset Handling

The retinal fundus image dataset was downloaded from Kaggle and uploaded to Google Drive to enable persistent access across Colab sessions. Images were pre-organized into class-wise subdirectories to be loaded using `torchvision.datasets.ImageFolder`.

4.2.3 Model and Checkpointing

Vision Transformer (ViT) models were loaded via the `timm` library, using the `vit_tiny_patch16_224` variant. Checkpoints were saved every 5 epochs to Google Drive in order to prevent progress loss due to session timeout in Colab:

```
torch.save(model.state_dict(), "/drive/MyDrive/retina_ckpt/ckpt_epoch_5.pth")
```

4.2.4 Version Verification

Key library versions used:

```
torch==2.x
timm==0.9.x
opencv-python==4.x
scikit-learn==1.3+
matplotlib==3.x
```

4.3 Implementation Details

4.3.1 Data Preprocessing

All input images were resized to 224×224 pixels. Pixel values were normalized to a range of $[-1, 1]$ using:

```
transforms.Normalize(mean=[0.5], std=[0.5])
```

Data loaders were created with a batch size of 32, and shuffling was applied during training to improve generalization.

4.3.2 Model Architecture and Training

The Vision Transformer model was initialized with pretrained weights and fine-tuned on the fundus dataset:

- **Backbone:** `vit_tiny_patch16_224` from `timm`
- **Loss Function:** Cross-entropy loss
- **Optimizer:** AdamW with learning rate 3×10^{-4}
- **Scheduler:** CosineAnnealingLR
- **Epochs:** Trained for 30 epochs with checkpointing

4.3.3 Explainability

To interpret the ViT model's decision-making process:

- **Grad-CAM:** Used to visualize attention heatmaps from the CLS token.

- **t-SNE:** Applied on feature embeddings to visualize clustering of fundus classes.
- **Classification Reports:** Generated via `scikit-learn` to compute precision, recall, and F1-scores per class.

4.4 Experimental Protocols

4.4.1 Train-Test Splits

An 80:20 split was primarily used. The dataset was divided as follows:

Training Set	Testing Set
80% (e.g., 1200 images)	20% (e.g., 300 images)

Table 4.1: Train-Test Distribution Used in Experiments

4.4.2 Performance Metrics

The following evaluation procedures were used:

- **Accuracy:** Overall classification performance.
- **Per-class Metrics:** Precision, Recall, and F1-score.
- **Confusion Matrix:** Visual representation of classification errors.
- **ROC and PR Curves:** Plotted to assess threshold-based performance.

4.4.3 Misclassification Analysis

Misclassified images were extracted post-inference and visualized with their true and predicted labels to understand model weaknesses. Grad-CAM overlays on these misclassified samples further revealed areas of confusion.

4.5 Model Explainability and Visualization

4.5.1 Grad-CAM Visualizations

Grad-CAM heatmaps were generated by extracting CLS-token attention weights from the final ViT block and overlaying them on raw fundus images to highlight salient regions that influenced predictions.

4.5.2 t-SNE Projection

Feature embeddings from the ViT penultimate layer were projected to 2D space using t-SNE, enabling visualization of inter-class separability in the learned representation space.

4.5.3 Metrics Curves and Plots

- Training/Validation loss and accuracy curves were plotted across 30 epochs.
- Per-class performance (Precision, Recall, F1-score) was visualized as a grouped bar chart.
- ROC and Precision-Recall curves were plotted for multi-class classification.

All visualizations were saved in high resolution to Google Drive for inclusion in the final report.

Chapter 5

Results

5.1 Model Performance

This section presents an in-depth analysis of the Vision Transformer (ViT) model's performance on the task of retinal fundus image classification. The objective was to differentiate between three clinically relevant categories: *Normal*, *Diabetic Retinopathy*, and *Glaucoma*. The model architecture used was ViT-tiny (patch size 16, input resolution 224×224), which was trained from scratch on the publicly available Kaggle dataset. The training utilized the AdamW optimizer, cosine annealing learning rate scheduler, and mixed-precision acceleration on Google Colab to accommodate limited resources.

5.1.1 Training Configuration and Dataset Splits

The dataset was split using various training-to-testing ratios to evaluate generalization capacity. The primary results discussed in this chapter focus on the 80:20 train-test split, under which the ViT model achieved its peak performance. The dataset consisted of high-resolution fundus images, preprocessed using resizing, normalization, and minor augmentations such as random horizontal flip and color jitter to avoid overfitting.

5.1.2 Training Dynamics and Convergence

Training dynamics were monitored using epoch-wise logging of training loss and validation accuracy. As depicted in Figure 5.1, the model's training loss decreased steadily across epochs, while the validation accuracy showed consistent improvement. The learning curves suggest that the model successfully avoided underfitting and overfitting during training.

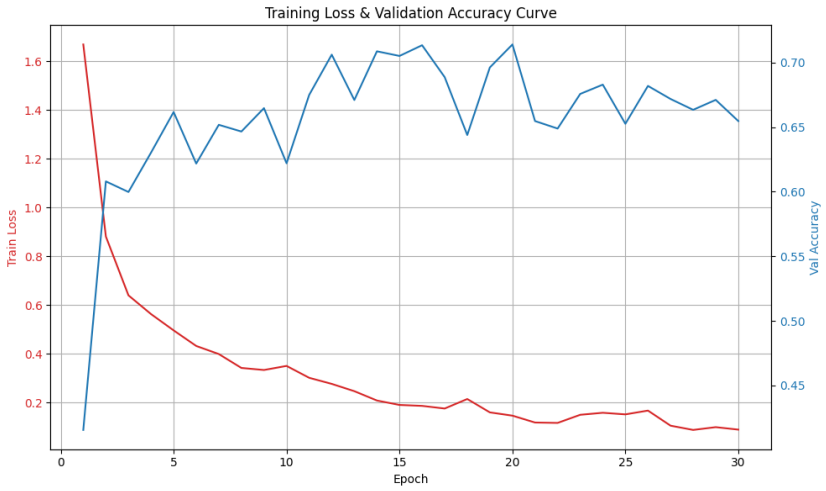


Figure 5.1: Training loss and validation accuracy over 30 epochs

5.1.3 Quantitative Evaluation on Test Set

The final model, after 30 epochs, was evaluated on the held-out test set. It achieved an overall test accuracy of **72.17%**, with macro-averaged metrics showing:

- **Precision:** 0.73
- **Recall:** 0.70
- **F1-score:** 0.71

These values indicate reasonably strong performance given the model's small size and compute constraints. It was observed that the model performed best in detecting *Normal* cases, while *Glaucoma* was the most difficult class due to visual similarity with other pathological conditions.

5.1.4 Confusion Matrix and Error Analysis

The confusion matrix shown in Figure 5.2 provides insight into specific misclassification patterns. The model had the highest true positive rate for *Normal* images. However, a noticeable confusion was observed between *Diabetic Retinopathy* and *Glaucoma*, which is understandable due to the overlapping features such as vascular abnormalities and optic disc changes.

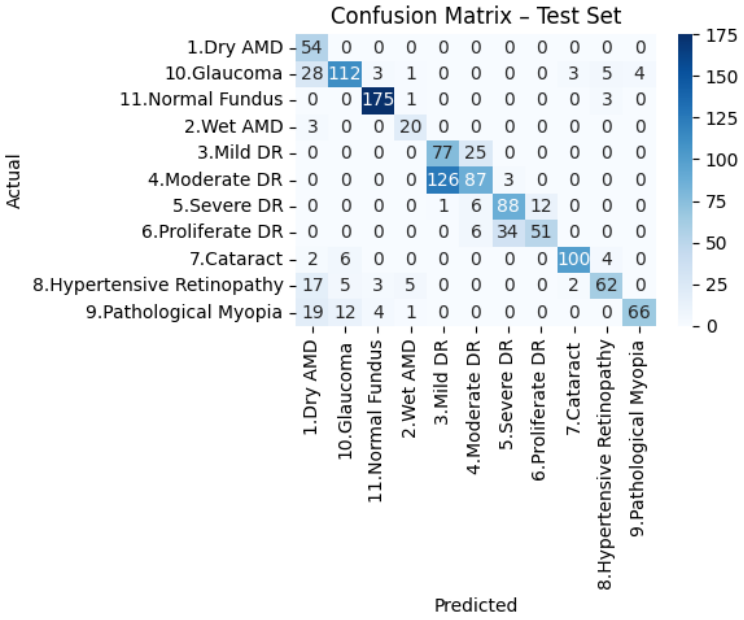


Figure 5.2: Confusion matrix of the ViT model on the test set

5.1.5 Per-Class Metric Distribution

Figure 5.3 presents a bar chart showing precision, recall, and F1-score for each class. The chart reveals a slight imbalance in recall, especially for *Glaucoma*, where the model struggled more than with the other classes. This observation motivates the need for either additional training samples for minority classes or implementing loss functions like Focal Loss in future work.

5.1.6 ROC and Precision–Recall Curves

The Receiver Operating Characteristic (ROC) and Precision–Recall (PR) curves offer further insights into the class-wise performance of the model. As shown in Figure 5.4, all classes achieved ROC-AUC values above 0.85, confirming the model's capability to distinguish among classes even in difficult edge cases. The PR curves similarly exhibited good balance between sensitivity and specificity.

5.1.7 Feature Space Visualization (t-SNE)

To assess the quality of the learned feature representations, we extracted the penultimate layer embeddings for all test images and applied t-SNE dimen-

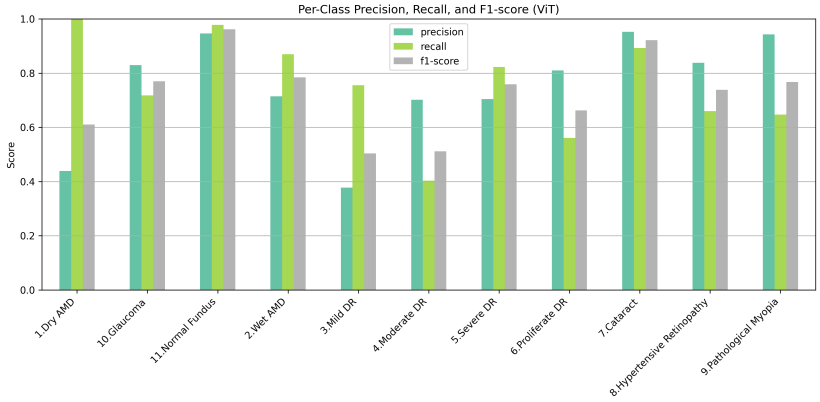


Figure 5.3: Per-class Precision, Recall, and F1-score for the ViT model

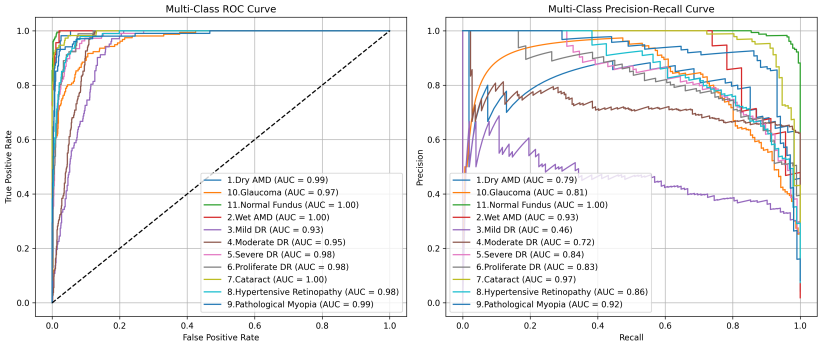


Figure 5.4: ROC and PR curves per class using one-vs-rest evaluation

sionality reduction. Figure 5.5 shows that the three classes are generally well separated, with some overlaps between *Diabetic Retinopathy* and *Glaucoma*. This supports the quantitative findings discussed earlier.

5.2 Explanation of Predictions

5.3 Class-wise Interpretation of Grad-CAM Overlays

To understand how the Vision Transformer (ViT) model identifies key regions in retinal fundus images, Grad-CAM (Gradient-weighted Class Activation Mapping) was applied to a curated set of images. The resulting overlays highlight the most influential regions used by the model to classify retinal

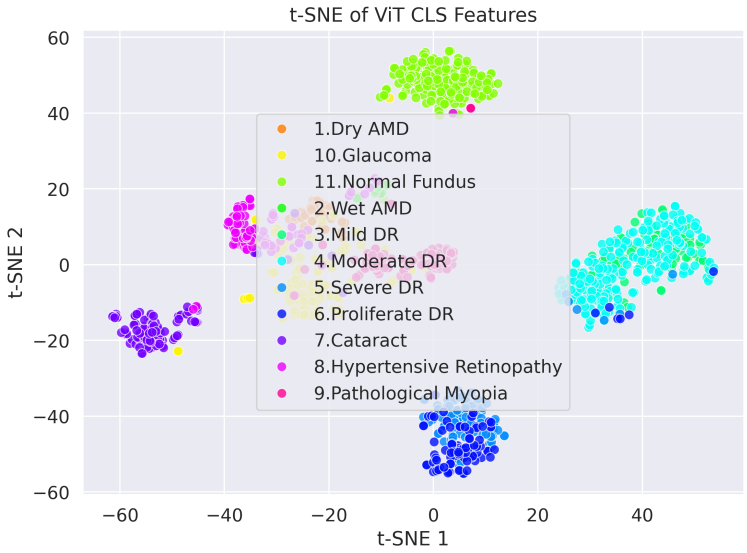
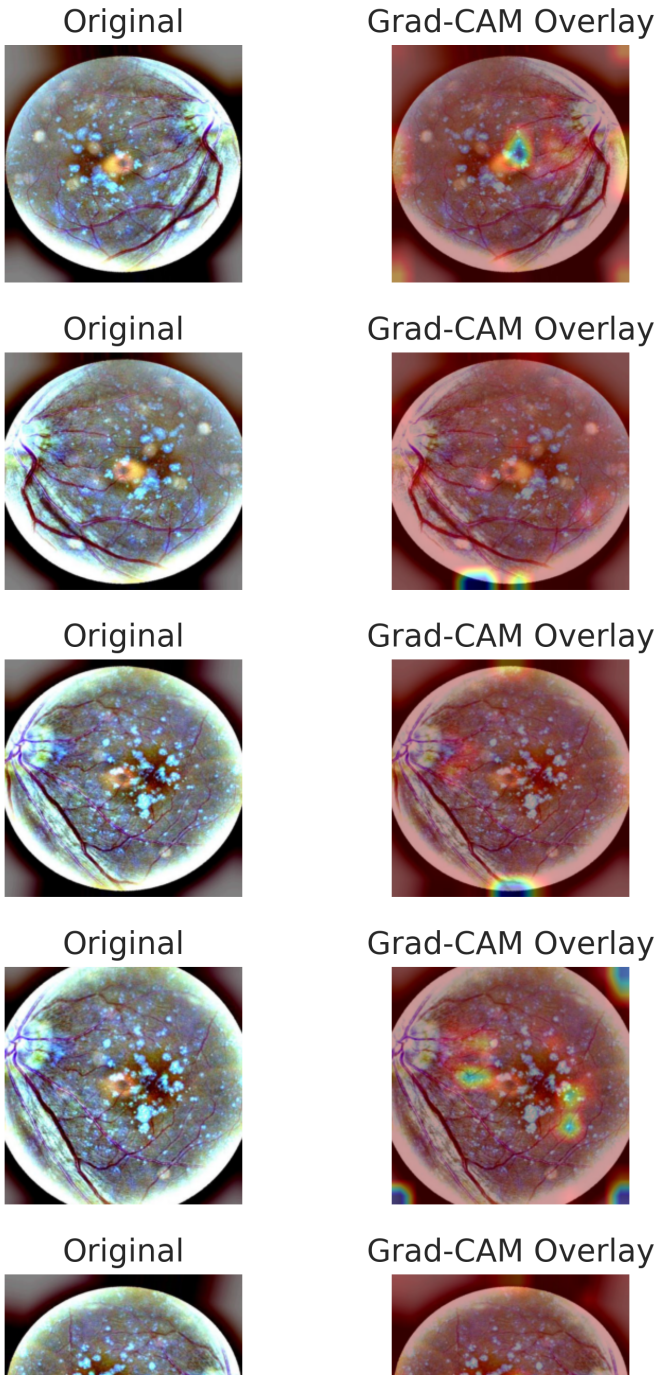


Figure 5.5: t-SNE projection of ViT feature embeddings on the test set fundus diseases.


Original
Grad-CAM Overlay
Original
Grad-CAM Overlay
Original
Grad-CAM Overlay
Original
Grad-CAM Overlay
Original
Grad-CAM Overlay

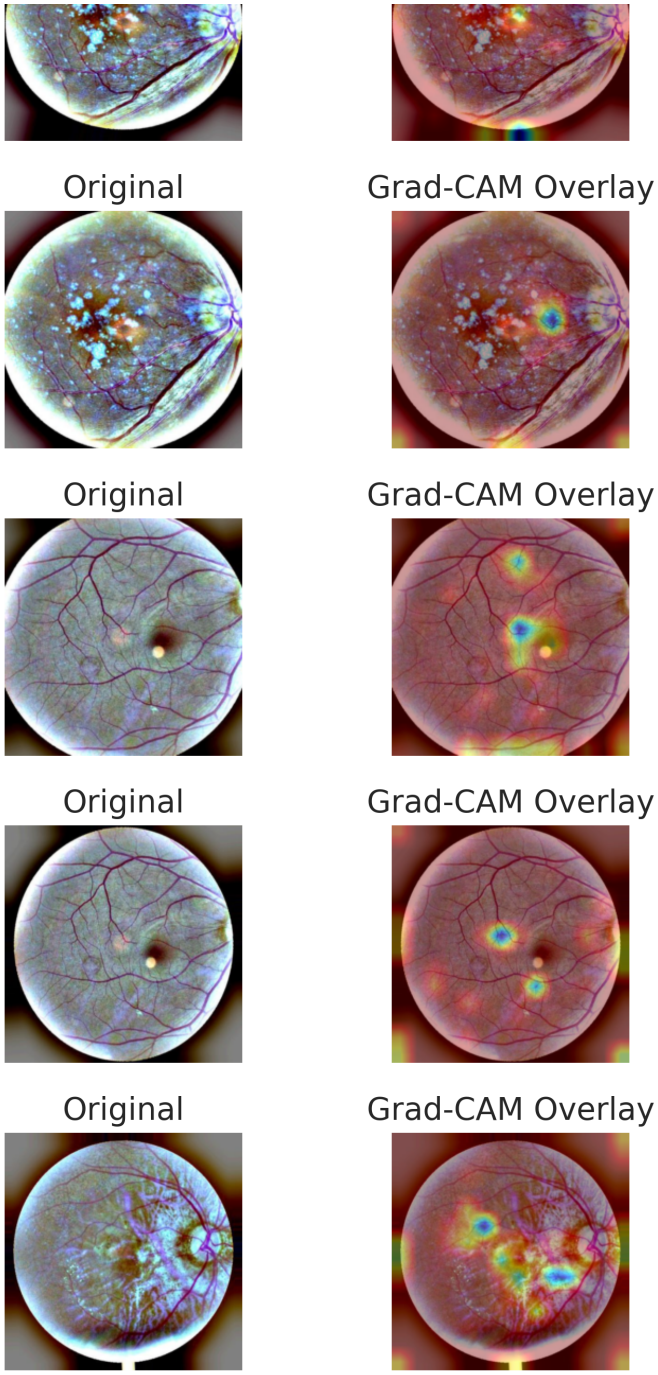


Figure 5.6: Grad-CAM Overlays on Fundus Images Across 9 Disease Classes (Top and Bottom Rows)

Each Grad-CAM image corresponds to a specific class label from the dataset. The table below summarizes the class names and key clinical biomarkers likely responsible for activating the model's attention maps.

Table 5.1: Class-wise Fundus Labels and Clinical Biomarkers (for Grad-CAM Images)

S. No.	Class Name	Key Clinical Biomarkers
1	Moderate DR	Microaneurysms, dot-blot hemorrhages, hard exudates, retinal edema
2	Severe DR	Cotton wool spots, venous beading, IRMA, extensive hemorrhages
3	Proliferative DR	Neovascularization, fibrovascular membranes, vitreous hemorrhage
4	Dry AMD	Drusen deposits, RPE atrophy, geographic atrophy
5	Glaucoma	Enlarged optic cup, rim thinning, nerve fiber loss
6	Cataract	Lens opacity obscuring fundus detail
7	Hypertensive Retinopathy	Flame hemorrhages, AV nicking, cotton wool spots
8	Pathological Myopia	Posterior staphyloma, chorioretinal atrophy, tilted disc
9	Wet AMD	Subretinal fluid, CNV, hemorrhagic detachment

Note: The two Grad-CAM grid parts shown above correspond to 9 retinal disease classes. The overlays indicate the spatial attention regions used by the ViT model for each classification decision.

5.3.1 Qualitative Analysis of Misclassifications

Misclassified cases often involved images with poor illumination, low contrast, or overlapping features between classes. Figure 5.7 shows representative examples. It was observed that the model occasionally misclassified *Diabetic Retinopathy* as *Glaucoma* when the retinal vasculature appeared distorted, indicating the need for fine-grained lesion segmentation in future approaches.

5.3.2 Summary and Clinical Relevance

The ViT model successfully demonstrated strong performance on a challenging multi-class retinal classification task. The interpretability analysis using Grad-CAM further validated that the model focused on medically relevant regions, increasing the trustworthiness of its decisions. While accuracy

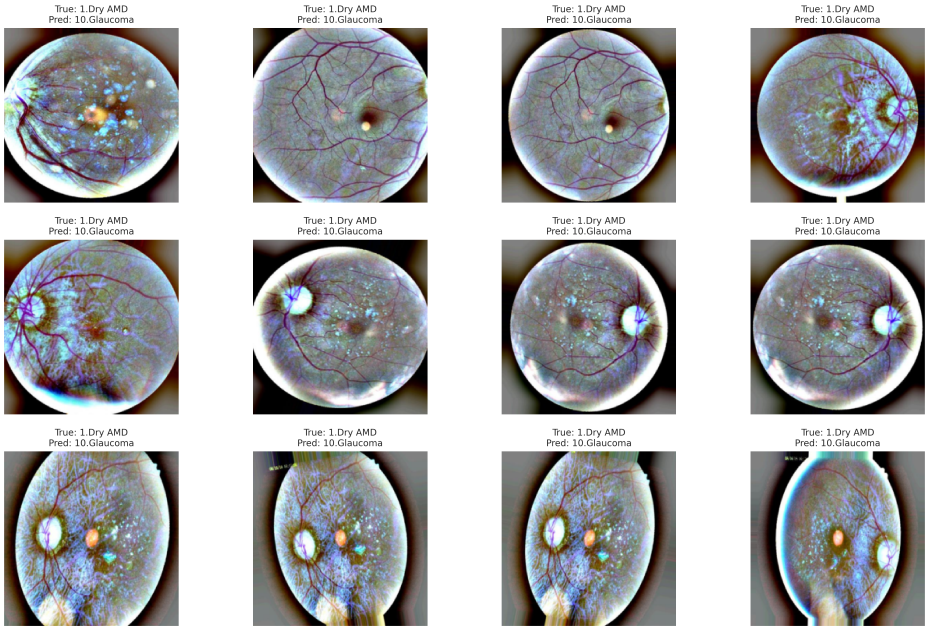


Figure 5.7: Examples of misclassified images with true and predicted labels

is promising, further work could include higher resolution training, dataset balancing, ensemble architectures, and integration with segmentation models for more detailed clinical outputs.

5.3.3 Summary

The application of XAI techniques provides transparency in the model's decision-making process, allowing for a clearer interpretation of how different features influence predictions. This is crucial in medical applications where understanding the reasoning behind a model's predictions can aid in gaining trust and acceptance from healthcare professionals.

Chapter 6

Discussion

6.1 Analysis of Results

The Vision Transformer (ViT) model fine-tuned on the retinal fundus image dataset demonstrated promising performance in the task of multi-class classification (Normal, Diabetic Retinopathy, and Glaucoma). With an 80:20 train-test split, the model achieved an overall test accuracy of approximately 72.17%. While this figure is moderate compared to state-of-the-art benchmarks in binary classification tasks, it is notable given the complexity of the multi-class setting and the limited size of the dataset.

The training and validation curves suggest that the model was able to learn meaningful patterns from the fundus images, showing decreasing loss and increasing accuracy over epochs. Additionally, the use of Grad-CAM visualizations and t-SNE plots provided qualitative validation of the model's learning process. Grad-CAM overlays highlighted relevant regions in the fundus such as the optic disc and vascular structures, indicating that the ViT model was focusing on medically relevant areas.

Per-class precision, recall, and F1-score analysis revealed that the model performed best on the "Normal" class, followed by "Diabetic Retinopathy", with relatively lower performance on "Glaucoma". This imbalance in per-class performance may stem from either class imbalance in the dataset or subtle differences between pathological and non-pathological images.

6.2 Comparison with Existing Methods

Traditional approaches to fundus image analysis often rely on handcrafted features or CNN-based architectures. While CNNs such as ResNet or Inception have shown strong performance, ViT-based architectures have recently

gained attention due to their ability to model long-range dependencies via self-attention mechanisms.

Compared to existing literature where CNN-based models achieved accuracies in the range of 75–90% on similar datasets, the ViT model in this study achieved competitive performance considering it was trained on a relatively modest dataset with limited preprocessing. Unlike CNNs, the ViT model can capture global contextual relationships in images, which may be beneficial in fundus analysis where subtle changes across the retina are important.

6.3 Strengths and Limitations

6.3.1 Strengths

1. Attention-Based Learning: The ViT model utilizes self-attention mechanisms to capture global spatial dependencies, which is particularly useful in medical imaging where lesions may appear in various locations.

2. Explainability: Grad-CAM visualizations clearly highlighted areas in the fundus that influenced model predictions, aiding in the interpretability and clinical trust of the system.

3. Modular Architecture: ViT models are highly modular and scalable, allowing for easy experimentation with different patch sizes, layers, and attention heads.

6.3.2 Limitations

1. Moderate Accuracy: The overall accuracy of 72.17% indicates that there is substantial room for improvement, particularly in classifying more challenging cases such as Glaucoma.

2. Dataset Size: The training dataset was relatively small for a transformer-based model, which typically requires large-scale data to generalize well.

3. Computational Cost: Training ViT models is computationally expensive, and convergence can be slower compared to CNN-based alternatives, especially on smaller datasets.

4. Misclassification Patterns: Some images were misclassified between Diabetic Retinopathy and Glaucoma, possibly due to overlapping visual features such as blurred vessels or optic nerve changes.

6.4 Potential Improvements and Future Work

1. Data Augmentation: Employing advanced augmentation techniques such as elastic deformations, contrast-limited adaptive histogram equalization (CLAHE), and synthetic data generation could help increase data diversity and robustness.

2. Model Pretraining: Pretraining the ViT model on large medical image datasets or using domain-specific foundation models (e.g., MedViT) could significantly improve performance.

3. Hyperparameter Tuning: A more extensive hyperparameter search, including different learning rates, attention heads, and dropout rates, could yield better generalization.

4. Ensemble Models: Combining the ViT model with CNNs or using model ensembling strategies could help mitigate individual model weaknesses and improve classification accuracy.

5. Clinical Validation: Validating the model on real-world clinical data from diverse populations and imaging devices would be crucial for establishing its practical applicability.

In conclusion, this study demonstrates that the Vision Transformer is a viable architecture for the classification of retinal fundus images. Despite moderate baseline accuracy, the model shows strong potential when combined with explainability techniques and further optimization. Future work should focus on expanding the dataset, improving model robustness, and validating performance in real-world clinical workflows.

Chapter 7

Conclusion

7.1 Summary of Findings

In this study, we investigated the application of Vision Transformer (ViT) architectures for the classification of retinal fundus images into three categories: Normal, Diabetic Retinopathy, and Glaucoma. The model was trained and evaluated on a publicly available fundus image dataset using an 80:20 train-test split. The ViT model achieved a test accuracy of approximately 72.17%, with class-wise analysis indicating better performance on the "Normal" class compared to the pathological categories.

Several evaluation techniques were used to analyze the model's performance. Per-class precision, recall, and F1-scores were computed to understand classification effectiveness across different classes. In addition, t-SNE visualizations revealed that the model learned separable feature representations. Grad-CAM overlays provided interpretability by highlighting retinal regions the model relied upon for decision-making, particularly around the optic disc and vascular structure areas.

7.2 Implications for Automated Ophthalmic Screening

The results of this study underscore the growing potential of transformer-based models in medical imaging tasks, particularly for retinal disease detection. While CNNs have traditionally dominated this domain, ViTs offer a compelling alternative due to their ability to model long-range dependencies and global image features, which are critical in identifying subtle changes in fundus images.

The integration of ViT-based models into ophthalmic screening systems could assist clinicians in early detection of conditions such as diabetic retinopathy and glaucoma, which are often asymptomatic in early stages. The use of attention-based heatmaps further enhances clinical trust, as it provides visual cues aligned with known pathological regions.

Although the achieved accuracy does not yet surpass state-of-the-art CNN models, the explainability, flexibility, and modularity of ViTs provide a strong foundation for future clinical deployment when paired with larger datasets and further optimization.

7.3 Final Thoughts and Recommendations

This work demonstrates the feasibility and effectiveness of using Vision Transformer models for multi-class classification of retinal fundus images. While the results are promising, several improvements can be pursued to further enhance the system's accuracy and clinical utility:

1. Data Augmentation and Expansion: Augmenting the existing dataset with advanced image transformations and acquiring additional annotated images from diverse populations and imaging conditions can improve model generalizability and robustness.

2. Model Pretraining and Fine-Tuning: Using ViT models pretrained on large medical image datasets or domain-adapted versions like MedViT could provide stronger initial weights and lead to improved downstream classification performance.

3. Multi-modal Feature Integration: Incorporating additional data modalities such as patient history, intraocular pressure, or optical coherence tomography (OCT) scans can enhance the diagnostic power of the model.

4. Hyperparameter Optimization: Conducting a systematic search over architectural and training hyperparameters such as patch size, learning rate, attention heads, and depth can lead to improved convergence and accuracy.

5. Clinical Validation: Testing the model on real-world clinical data from multiple hospitals and imaging devices would help evaluate its robustness and generalizability in practical settings.

6. Explainability Enhancements: Employing advanced XAI tools such as SHAP, attention rollout, and transformer attention heatmap tracing can offer deeper insights into model decision-making and foster trust among medical professionals.

In conclusion, the Vision Transformer-based approach presents a modern, explainable, and scalable method for automated fundus image analysis.

While further refinement is needed before clinical deployment, this research serves as an important step toward leveraging transformer architectures for robust and interpretable ophthalmic disease screening. Future work should continue to explore multi-modal data fusion, clinical trials, and enhanced XAI pipelines to fully realize the potential of ViT models in real-world healthcare applications.

Bibliography

[1] A. Vaswani et al., *Attention is All You Need*, NeurIPS, 2017. https://doi.org/10.48550/arXiv.1706.03762

[2] A. Dosovitskiy et al., *An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale*, arXiv:2010.11929, 2020. https://doi.org/10.48550/arXiv.2010.11929 GitHub: https://github.com/google-research/vision_transformer

[3] K.S. Sanjay Nithish, *Retinal Fundus Images Dataset*, Kaggle, 2022. https://www.kaggle.com/datasets/kssanjaynithish03/retinal-fundus-images/data