Detection and Exploration of Diabetic Retinopathy Using Advanced Explainable AI (XAI) with Distinctive Features along with Automated Report Generation Utilizing the Deep Learning Method

by

Mugdha Saha
21301645
Maisha Binta Alam
21201596
Md Maruf
21301724
Al-Shahriar Redoy
21301348

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

Department of Computer Science and Engineering
BRAC University
February 2026.

Declaration

It is hereby declared that

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

Student's Full Name & Signature:

Mugdha Saha
21301645

Maisha Binta Alam
21201596

Md Maruf
21301724

Al-Shahriar Redoy
21301348

Approval

The thesis titled "Detection and Exploration of Diabetic Retinopathy Using Advanced Explainable AI (XAI) with Distinctive Features along with Automated Report Generation Utilizing the Deep Learning Method " submitted by

1. Mugdha Saha (21301645)
2. Maisha Binta Alam (21201596)
3. Md Maruf (21301724)
4. Al-Shahriar Redoy (21301348)

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

Examining Committee:

Supervisor:
(Member)

Mr. Md. Sabbir Ahmed

Senior Lecturer
Department of Computer Science and Engineering
Brac University

Co-Supervisor:
(Member)

Mr. Rafeed Rahman

Senior Lecturer
Department of Computer Science and Engineering
Brac University

Thesis Coordinator:
(Member)

Dr. Md. Golam Rabiul Alam

Professor
Department of Computer Science and Engineering
Brac University

Head of Department:
(Chair)

Dr. Sadia Hamid Kazi

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

Acknowledgement

First, we would like to express our gratitude to the Almighty for the continuous blessings which gave us the strength and patience to complete our thesis.

For the past four years of our university's journey, we all have been through thick and thin moving toward graduation. Each step we took made us scared, curious, overwhelmed and excited all at the same time. We have been surrounded by some amazing friends who were our constant support and help throughout this journey of our graduation life. We would like to thank them deeply for being the ultimate teacher, ultimate supporter, ultimate shoulder to cry on and the ultimate people who are always there and will be there for us in the long run.

Now, we would like to acknowledge the utmost guidance, help and the ultimate light of our entire journey of this paper who carried us through it all are our supervisor Mr. Md. Sabbir Ahmed and our Co-supervisor Mr. Rafeed Rahman. Without their impactful insights we would not have been able to finish this. We will be forever grateful to them for their teaching, guidance and overall shield.

Furthermore, we would like to appreciate our other respected faculty members and staff members of our esteemed Brac university for providing us with the necessary academic environment, resources and support during the completion of this work. We are also thankful to the researchers and authors whose studies helped us build the foundation of our research.

Lastly, we would like to take this opportunity to express our whole hearted gratitude to our parents, siblings, and friends who are like our family who are our source of inspiration. Without their love and prayers and their unwavering encouragement, our core would not have been this strong. And we would like to take a moment to reflect on ourselves without whose hardwork, sleepless nights and perseverance this would not be possible.

Abstract

In a world full of technologies and high screening usage, affected eyes need more attention. A serious disease that causes blindness is related to one of the most infamous one named as Diabetic retinopathy (DR). It leads to blindness if not detected early. Unfortunately, the detection process is time consuming and requires prior intensive labour. Our paper mainly focuses on how to propose something useful so that it has a framework which includes solving five class severity DR grading issues along with explainable AI (XAI) with quality-aware deep learning methods examining the fundus images. Merging two datasets one being APTOS 2019 and the other being the kaggle competitions Diabetic retinopathy dataset into a large and curated one which triggers the work of severe class-imbalance and varieties of different quality images. Our paper explores about 5 different state-to-art architectures with some notable results to have recorded for along the entire paper. The architectures listed as ConvNeXt, CoAtNet, Hybrid CoAtNet–ConvNeXt, MaxViT and Vision Mamba which were worked upon some integrated preprocessing-loss, evaluation which stands upon the clinically checkmark point. To further elaborate on the working, this paper works on lesion-level integration to build the clinical trust focusing on the impacted regions more following the predictions of the models. To summarize, our paper proposes a robust architecture which focuses on a practical way to demonstrate building a robust, clinically trustworthy and accurate framework which will make a difference in our living real world.

Keywords: Diabetic Retinopathy (DR), AI-driven system, Deep learning, Explainable AI (XAI), Convolutional Neural Networks (CNNs), Retinal fundus images, Automated reporting module, Quadratic Weighted Kappa (QWK), Grad-CAM, Hybrid Architecture, CoAtNet, ConvNeXt, Vision Mamba, MaxVit, Quality Tier Analysis, Class Imbalance, Black Box AI.

Table of Contents

List of Figures

Chapter 1

Introduction

1.1 Background

Diabetic retinopathy (DR) is a progressive microvascular complication of diabetes mellitus and remains one of the leading causes of preventable blindness worldwide [16]. Prolonged hyperglycaemia damages retinal capillaries and leads to characteristic lesions such as microaneurysms, haemorrhages, hard and soft exudates, and, in advanced stages, neovascularisation [5]. DR severity is typically graded into five stages No DR, Mild, Moderate, Severe non-proliferative DR and Proliferative DR based on the type, extent and distribution of these lesions in colour fundus imaging. Because early stages are often asymptomatic, many patients present only when irreversible retinal damage has already occurred, making early and scalable screening a major clinical and public-health priority [14].

Retinal fundus images has become the standard non-invasive modality for DR screening and longitudinal monitoring. Large public datasets such as APTOS 2019 Blindness Detection and the Diabetic Retinopathy Detection competition on Kaggle have enabled the development of automated DR grading systems by providing tens of thousands of labelled fundus images. However, these datasets are highly imbalanced—severe and proliferative cases constitute only a small fraction of the data—and exhibit substantial variation in image quality due to different cameras, illumination conditions and acquisition protocols. These factors complicate model training and often lead to biased systems that perform well on common grades but fail to reliably detect sight-threatening stages.

Deep learning has transformed medical image analysis and DR detection in particular. Convolutional neural networks (CNNs) and, more recently, transformer-based and hybrid architectures have achieved strong performance on DR grading tasks by learning complex, fine-grained features directly from fundus images [2], [15]. Yet most existing works are built around single-family architectures (CNN-only or transformer-only) and focus primarily on improving overall accuracy or AUC, with comparatively less attention to robustness under extreme class imbalance, explicit modelling of image quality, or systematic interpretability for clinical use. There is also growing interest in state-space models such as Vision Mamba, which extend sequence-modelling ideas to images and offer a different inductive bias compared to standard convolutional or attention-based networks.

At the same time, the "black-box" nature of high-capacity deep models has become a major barrier to adoption in clinical workflows. Ophthalmologists and other clinicians must be able to understand why an automated system predicts a given DR grade, particularly for high-stakes decisions such as referral for laser photocoagulation or intravitreal therapy. Explainable AI (XAI) techniques—such as saliency maps, class-activation maps, attention visualisations and gradient-based attribution—can highlight the retinal regions and lesion patterns that drive a model's decision, thereby bridging the gap between statistical performance and clinical trust [6], [18]. However, in most DR studies, XAI is treated as an auxiliary add-on rather than as an integral design component tied to model choice, data quality and evaluation [12].

In this thesis, a large-scale merged DR dataset of 77,576 fundus images is constructed from APTOS 2019 and the Diabetic Retinopathy Detection Kaggle competition and subjected to systematic quality assessment and class-aware augmentation. On top of this curated dataset, five modern architectures—ConvNeXt, CoAtNet, a Hybrid CoAtNet–ConvNeXt model, MaxViT (Multi-Axis Vision Transformer) and Vision Mamba (a state-space vision model)—are investigated within a unified, quality-aware and explainable DR grading framework. By combining convolutional, hybrid convolution–attention, multi-axis transformer and state-space designs, the study aims to understand how different architectural families behave under real-world constraints of class imbalance and heterogeneous image quality, and how their decisions can be made clinically interpretable through XAI.

1.2 Rationale of the Study or Motivation

Diabetic retinopathy (DR) is becoming more common as the number of diabetes cases worldwide rises. According to worldwide health data, about 90 million individuals are afflicted by DR, with a sizable proportion at risk of developing vision-threatening consequences. Early detection is critical because diabetic retinopathy often progresses silently, with no symptoms up until irreversible damage comes about [16]. It is for these reasons that DR constitutes a significant public health issue indeed, a rising one especially in low- and middle-income countries where ophthalmology service and programed screening are scarce.

Conventional diagnostic techniques involve a manual fundus image examination, which is very time-consuming and inefficient with low accuracy. Manual grading of the retinal images is labor-intensive and suffers from intra- and inter-observer variations leading to non-scalability, making it challenging for doing large scale screenings especially in resource constraints areas. This creates a need for scalable and sophisticated tools to analyze retinal images and aid health professionals.

Deep learning algorithms, especially convolutional neural networks (CNNs), have revolutionized medical imaging and surpassed conventional methods in accuracy and efficiency [15]. These systems may find small patterns such as early microaneurysms or exudates that the eye may miss. Nevertheless, in spite of these high performances achieved by many automated DR tools on benchmark data sets, a few substantial limitations have prevented them from being widely used as daily screening tool in the real world.

Firstly, the datasets for DR are frequently associated with extreme class imbalance (severe or proliferative type of DR tends to be substantially under-represented compared to non-DR and mild cases). This biases the standard training objectives towards the majority categories and lowers its sensitivity for rare but clinically high-risk classes, exactly where early detection is key. Second, variation of image quality due to variations across cameras, illumination, focus and artefacts like blur or vignetting makes fundus images highly variable. Few models explicitly take image quality into account for either training or evaluation, making them less robust when deployed outside of curated training distribution. Third, the interpretability is still a major challenge. In practical situations, one should not only predict accurately, but also understand why. By the use of explainable AI (XAI) with the help of different deep learning models secures clarification with their decision-making [10]. This enhances the openness of the model to be confident by physicians in daily healthcare practice.

Moreover, automatic report generation from detected lesions and categorization outcome can be time efficient, reduce physician fatigue and help to standardize diagnostic reporting. Integration of such a system can substantially enhance clinical workflow and ensure uniform, objective diagnoses, especially for high-volume screening programs [3] [6]. Other studies have also shown that deep learning and automated reporting can decrease the diagnostic timing lag as well as maintain constant decision to output [19]. However, the majority of current DR pipelines do not integrate robustness, imbalance and interpretability in a unique clinically oriented context.

The motivation for the current study is to develop a therapeutically useful AI-enabled system that not only correctly detects DR but also identifies lesion location, explains its reasoning and produces automated diagnostic reports that support day-to-day clinical decision-making. In particular, this dissertation: (i) constructs a curated, quality-aware DR dataset from APTOS 2019 and DR Dataset with explicit consideration of class imbalance; (ii) devises a training strategy that capitalises on the quality sensitized augmentation using the class-aware loss functions to strengthen identification rare severe and proliferative cases; and embeds XAI techniques in the heart of the system so that model predictions are complemented by lesion-level visualisations as well as organised diagnostic narratives. Through the systematic evaluation of five state-of-the-art architectures—ConvNeXt, CoAtNet, Hybrid CoAtNet–ConvNeXt, MaxViT and Vision Mamba on the same data and loss pipeline as well as testing on the same five validation datasets. This thesis would like to show how these high performing models can not only be accurate but also robust, interpretable and viable for scalable DR screening in real-world healthcare settings.

1.3 Problem Statement

Diabetic Retinopathy (DR) is the primary cause of visual loss, usually asymptomatic at early stages, and therefore often not diagnosed until advanced [5]. However, the conventional means of fundus image inspection is timeconsuming with subjective judgement, which is infeasible for mass screening [14]. Despite the remarkable success of deep learning-based DR grading, there still lacks a universally accepted

and clinically-deployable framework that simultaneously tackles three key real-world challenges of extremely imbalanced class distribution, heterogeneous image quality, and interpretability.

Public datasets like APTOS 2019 and DR Dataset have a heavily unbalanced distribution towards non-DR and mild categories, while severe cases along with proliferation manifest much less frequently (benign cases only 5.3% in APTOS 2019). This imbalance is unfair to training models towards majority classes and decreases the sensitivity of sight-threatening DR stages. Furthermore, fundus image quality varies greatly because of differences in cameras and illumination, focus and artefacts but its consideration is mostly neglected during model development which results in low robustness across deployment conditions. Besides, despite their diagnosis capability deep learning methods generally lack intrinsic interpretability, which curtail the path to clinical application even with high accuracy [15]. The "black-box" behavior of such models diminishes clinician trust and hinders clinical validation [2], and commonly systems do not output any sort of lesion-level evidence (i.e., hemorrhages, exudates, microaneurysms) which makes diagnostic feedback less useful. Recent works claim to provide lesion-specific interpretability at the expense of full transparent pipelines [17]. Also, the absence of a report generating mechanism is inefficient; clinicians are required to manually record AI results [18].

The fundamental question to be solved in this thesis is that of a degree system for an AI-based diagnostic framework, providing high diagnostic accuracy along with lesion-level explainability and at the same time generation of reports compatible with real-world clinical workflows [9]. Hence, the above thesis is concerned with the development and rigorous evaluation of an interpretable and quality conscious deep learning model for five-class DR grading, where it utilizes a large merged dataset that performs robustly under severe class imbalance as well as diverse image quality while also presenting with clinically meaningful explanations.

1.4 Objective

The overarching objective of this thesis is to develop and evaluate a quality-aware and explainable deep learning framework for five-class diabetic retinopathy (DR) grading that is robust, clinically interpretable, and suitable for scalable screening. Hybrid deep learning systems combined with explainability mechanisms have demonstrated clinical potential in DR analysis [8].
The objectives of this study are:

- Construct and curate a large-scale DR dataset by merging APTOS 2019 and the Diabetic Retinopathy Detection Kaggle dataset, and to perform systematic image-quality assessment and class imbalance analysis across five DR grades.

- Develop a unified training pipeline and train five modern vision architectures—ConvNeXt, CoAtNet, Hybrid CoAtNet–ConvNeXt, MaxViT, and Vision Mamba—for five-class DR grading.

- Integrate explainable AI (XAI) techniques (e.g., Grad-CAM and attention-based visualisations) to generate lesion-level interpretability outputs that support clinical inspection of model predictions.

- Evaluate model performance using clinically meaningful metrics, including quadratic weighted Cohen's kappa, ROC-AUC, sensitivity, specificity, and confusion matrices, with emphasis on severe and proliferative DR detection.

- Develop a prototype automated report-generation module that combines predicted DR grade and XAI evidence into structured, clinician-readable diagnostic summaries for screening workflows.

- Enhance scalability and deployment readiness through transfer learning and modular design to support practical use across diverse clinical settings.

This study aims to combine advanced AI techniques with real-world clinical needs by offering a DR detection framework that is accurate, understandable, and user-friendly. The proposed method aims to assist ophthalmologists in delivering quick and accurate care, particularly in high-demand or resource-limited settings, by integrating lesion localization, explainability, and automation.

1.5 Methodology in Brief

The paper creates an easily understandable and interactive AI system to identify and analyze Diabetic Retinopathy (DR). It applies deep learning to classify the images, find particular lesions, and then Explainable AI (XAI) to explain itself in a better way, hence being able to write a diagnostic report individually. These principal stages include:

- **Research Design:** This study conducts quantitative experimental research on the construction and application of a deep learning system to diagnose diabetic retinopathy (DR). The experimental setup allows for comparison between different architectures in a controlled setting using the same dataset, preprocessing, training and evaluation procedures.

- **Data collection and Merging:** Labelled retinal fundus images are gathered from two publicly available repositories: APTOS 2019 Blindness Detection and the Diabetic Retinopathy Detection Kaggle dataset. We merge these two datasets in a large scale dataset, with the aim to increase the diversity of samples and spread out images over five DR levels. Following merging, the dataset is analyzed to characterise grade distribution and balance particularly that which under represent severe and proliferative DR like clinically significant retinal signs (microaneurysms haemorrhages or hard exudates) which are indicators of severity as may be obtained for interpretability analysis [16].

- **Image quality and data splits:** The combined dataset is quality-checked to minimize the impact of non-useful images (e.g., very blurred or low contrast fundus photographs). Images are categorized into four quality tiers (A-D) based on blur, brightness, contrast, and usable retinal area metrics, enabling tier-specific preprocessing strategies.

- **Stratified train-validation split:** Following quality-aware balancing to 65,000 images (13,000 per class), the dataset undergoes stratified random sampling into training (52,000 images, 80%) and validation (13,000 images, 20%) sets with a fixed random seed (42) for reproducibility. Each class maintains 10,400 training and 2,600 validation samples, preserving class balance and quality tier distribution. The validation set serves as a held-out test set for final evaluation, remaining completely isolated from the training process to ensure unbiased performance assessment. This split strategy follows standard practices in DR detection literature [11], [15].

- **Preprocessing and Data Augmentation:** All images are processed together in the pre-processing pipeline to homogenise the input format across various acquisition conditions. Tissue and retinal structures are visualized by contrast enhancement using CLAHE with 8×8 tiles for each color channel to observe lesions. Images are resized to the same resolution (224×224) for classification and noise reduction is done to reduce artefacts that can potentially interfere with model learning. Moreover, some common augmentation operations (e.g. rotation, horizontal/ vertical flipping and brightness change) are used to augment the training data for variety and overfitting resistance [5].

- **Quality-Critical Balancing and Class Imbalance Treatment:** And since the majority of DR datasets are imbalanced in favor of non-DR and mild cases, this work proposes a quality-aware balancing approaches which re-enforces under-sampled classes (especially severe and proliferative DR) without increasing low-quality samples. Bias toward majority classes is further mitigated by class-aware loss functions used in training, which elevate sensitivity to rare but clinically high-risk categories.

- **Model Training and Experimental Setup:** We train and evaluate five architectures with a single unified training protocol, ConvNeXt, CoAtNet, Hybrid CoAtNet-Con-vNeXt, MaxViT and Vision Mamba. We always keep the same preprocessing and data splits for fair comparison. Transfer learning is utilized to increase the efficiency of learning, and a staged fine-tuning strategy is employed in which we first freeze the model backbones and then partially unfreeze them for better feature extraction, adapted for retinal imagery. To accelerate training and save memory, we use mixed-precision GPU training. [2] [9].

- **Explainable AI (XAI) Integration:** To meet the clinical need for interpretability, we analyse trained models by means of XAI methodologies generating visual lesion-level explanations. Methods like Grad-CAM, saliency maps and attention-based heatmaps have been used to emphasize retinal regions that are most important in making predictions [18]. These visualizations aid clinical interpretability (enabling ophthalmologists to verify how much the model attends to relevant disease areas, aiding trust and interpretability) [20].

- **Automatic Report Generation Module:** We have developed a prototype automatic reporting module to translate AI outputs into summaries that are user-friendly for any clinician. It also amalgamates the predicted DR grade with the XAI evidence to produce succinct diagnostic text that can be helpful

for screening operations and can ease the burden of manual documentation. This allows for standardized results communication and structured decision support in the clinical setting [11].

- **Evaluation Metrics and Validation Strategy:** Model performance is assessed with clinically relevant metrics, like QWK, ROC-AUC, per-class sensitivity and specificity and confusion matrices. We also show other metrics such as accuracy, precision, recall and F1-score. Assessment focuses on the presence of severe and proliferative DR because these are deemed to represent the highest level of risk for vision loss [16]. For reliability and generalizability, cross-validation and external testing are typically included to evaluate whether the model generalizes across different imaging conditions or dataset distributions [10].

By following this step-by-step plan fig 1.1, we hope that the study will create a system that works well and explains how it thinks. This helps solve the main problem between AI being right and doctors trusting it. Adding details about specific lesions and automatic report writing makes it more useful for eye doctors in their daily work.

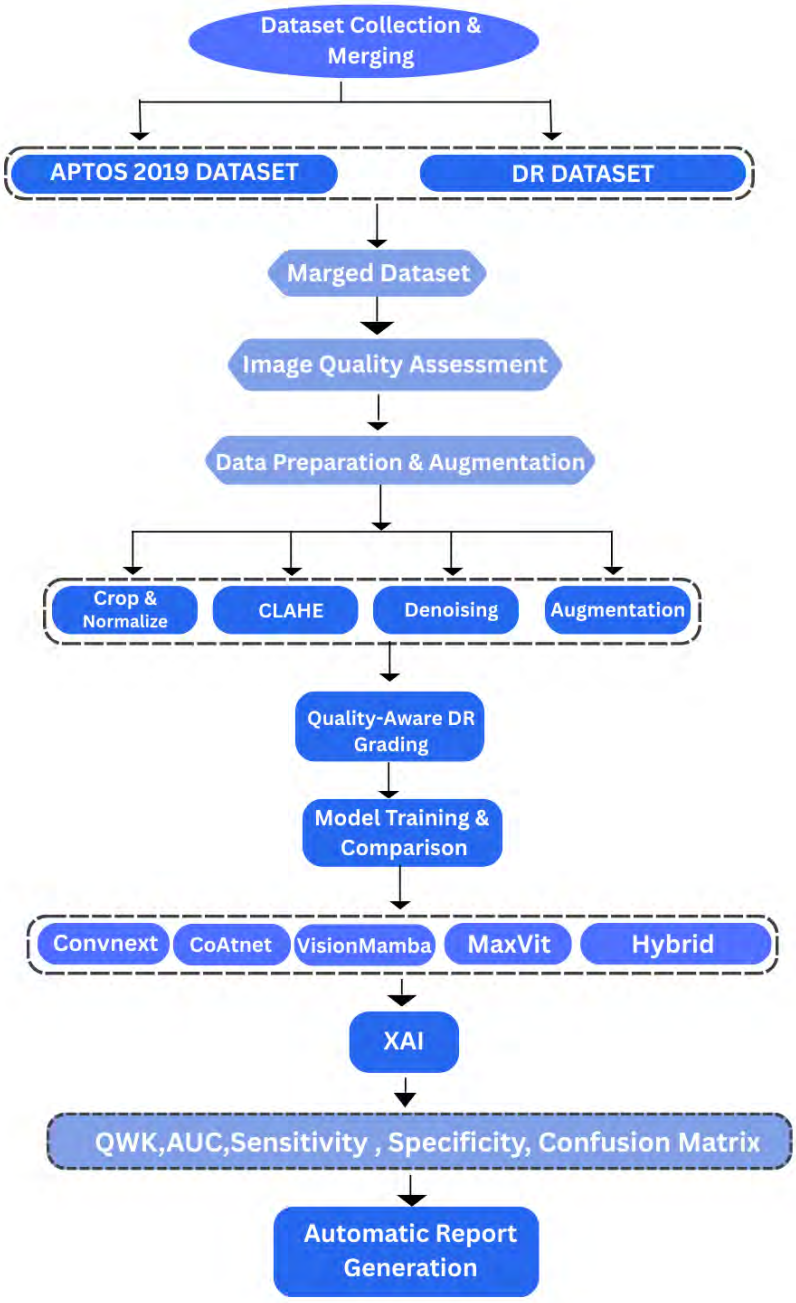

Figure 1.1: Workflow diagram

1.6 Scopes and Challenges

This thesis is focused on the development and validation of an automated deep learning system for five-class diabetic retinopathy (DR) severity grading based on 2D colour fundus photographs. The work utilises the public data, namely APTOS 2019 Blindness Detection (Diabetic Retinopathy) and the combined kaggle's Diabetic Retinopathy Detection (DR Dataset), with uniform distribution on splits to produce a large-scale dataset for training and testing. The research is limited to retrospective image analysis and does not involve prospective clinical trials, real-time applications or interfacing with hospital information systems.

Scope of the Research

- **Fundus images DR grading based on deep learning:** The approach applies deep learning to classify fundus images at different DR severity stages by capturing underlying disease patterns that may not be easily recognized manually [2].
- **Lesion-level feature interpretation for clinical relevance:** The method focuses on clinically relevant lesion evidence, including microaneurysms, haemorrhages and exudates, resulting in a more accurate assessment of the severity of DR [5].
- **Integration of explainable AI (XAI) techniques:** Interpretability is incorporated through visual XAI methods such as saliency maps, class activation mapping (CAM/Grad-CAM), and attention-based visualisations, allowing inspection of the retinal regions influencing predictions [18].
- **Prototype report generation for screening support:** A rule-based report generation module is included to produce concise diagnostic summaries based on predicted DR grades and XAI evidence, supporting structured screening-style outputs [11].The integration of XAI with reporting may improve diagnostic standardisation and clarity in automated reporting formats [13].

Challenges and Limitations

Several challenges are anticipated in developing a clinically reliable DR grading framework:

- **Severe class imbalance:** Public DR datasets are heavily skewed toward non-DR and mild cases, while severe and proliferative DR remain rare. This imbalance can bias model learning toward majority classes, reducing sensitivity for high-risk disease stages and making calibration difficult. [5].
- **Heterogeneous image quality and domain shift:** Fundus images vary widely due to differences in acquisition devices, illumination, focus, and artefacts. Such variability may reduce generalisation performance when models are tested on unseen distributions and can disproportionately affect architectures sensitive to noise.

- **Balancing interpretability and performance:** While XAI techniques give visual justifications, the heatmaps might not match entirely with clinical thoughts. It is still challenging to achieve clinically meaningful interpretability and assessing the reliability of explanations [7].
- **Computational complexity and training restrictions:** Training and optimization of multiple large-capacity models (both transformer-based and state-space-based ones) is computationally expensive and requires a significant amount of attention to optimise the GPU memory management [16].
- **Clinical trust and acceptance constraints:** Although transparency and reporting with the system explainability are granted, clinician-in-the-loop review is necessary to fully verify clinical acceptability of the systems. The notions of trust and adoption have also been found to be paramount for explainable systems in healthcare domain.
- **Ethical and regulatory issues (recognized but not clinically trialed):** Given that DR screening is an important medical application, it's essential to be aware of challenges like fairness, bias, transparency and compliance with medical AI norms. Our focus is not on the detailed verification of regulations, which would be a case of formal regulatory validation.

Chapter 2

Literature Review

This literature review examines current research on Diabetic Retinopathy (DR) detection. It looks at how Artificial Intelligence (AI), deep learning, and Explainable AI (XAI) are used for this job. The section starts with basic concepts. Next, it reviews current methods in detail. Last, it discusses main findings and related work that improve our understanding of DR detection

2.1 Preliminaries

Diabetic retinopathy (DR) is a progressive microvascular complication of diabetes that results in destruction of the retinal capillaries and is induced by long-term hyperglycemia. The resulting ischaemic damage leads to capillary dropout, leakage and retinal ischaemia with clinical manifestations such as microaneurysms, intraretinal haemorrhages, hard and soft exudates, venous beading which finally can result in retinal neovascularisation [16]. At the clinical level, there are generally five categories of severity for Diabetic Retinopathy: no DR or mild Non ProliferativeDR moderate non-proliferative DR Proliferative DR or Proliferative DR, at reference standards of grading that include consideration of number and size of lesions and/or their field allocations to grids. As the early stages are frequently asymptomatic, population-level screening with color fundus images is endorsed to identify referable DR before irreversible sight loss ensues.

Retinal fundus images, a non-invasive imaging modality that results in high-resolution retinal images, is commonly used for the diagnosis of DR in clinical setting. Ophthalmologists or trained graders manually review such images to interpret the existence and severity of DR. However, hand grading is labour intensive and time consuming, and is also affected by inter-observer variation and potential human error, particularly when signs are minimal or used for large scale screening [14]. These limitations motivate the search for automated diagnostic methodologies that enables efficient and identical screening.

Emerging from the recent advances in artificial intelligence (AI), especially deep learning (DL), research in automated DR detection and grading has gained remarkable momentum. Convolutional neural networks (CNNs) are commonly used since they learn a hierarchical feature representations from raw images itself, which minimizes the need for hand-crafted features and can accurately classify DR severity using fine-grained patterns on retinal images [2] [5]. CNN-based systems have

demonstrated high sensitivity and specificity for DR detection and have achieved performance comparable to expert graders in several studies.

Beyond CNNs, newer architectures have expanded the capabilities of automated DR grading. Transformer-based vision models use self-attention mechanisms to capture long-range dependencies across the image, which is useful for modelling global retinal context. Hybrid convolution–attention architectures such as CoAtNet and ConVnext combine local convolutional inductive biases with global attention, offering improved robustness and performance on high-resolution medical images. More recently, sequence state-space models such as Vision Mamba have been adapted for vision tasks, providing an alternative approach for modelling long-range interactions with linear time complexity, and have shown promising results in retinal image analysis.

Despite these developments, several dataset-related factors continue to limit real-world performance. Public DR datasets such as APTOS 2019 Blindness Detection and the Diabetic Retinopathy Detection Kaggle competition dataset provide large collections of labelled fundus images and are widely used benchmarks. However, these datasets exhibit severe class imbalance, where non-DR and mild cases dominate while severe and proliferative cases remain rare. Moreover, the quality of images varies greatly as a result of different cameras, lighting conditions, and focus/control settings. Such properties have implications on model training, evaluation as well as generalisation and thus underline the need for quality-aware and imbalance-tolerant learning schemes.

Interpretability is another important challenge of clinical interpretation. Despite their excellent prediction performance, most of the deep learning models are considered as black-boxes which hardly gives any explanation for why a specific DR grade was predicted. This is an important problem in medicine, where transparency is needed for diagnostic accountability and clinician confidence, as well as patient safety. Explainable AI (XAI) aims to address this issue by providing tools to interpret model decisions. Visual interpretation methods such as saliency map, Grad-CAM and attention heatmaps show the regions in retinal images that contribute most to the model's prediction for physicians to verify whether the model is looking on clinically meaningful evidences [18]. The introduction of XAI might help to improve their confidence in the use of AI-assisted screening and enhance clinical decision-making. This motivation aligns the primary objective of the thesis: a unification of deep learning's diagnostic power with lesion-level explainability to drive early DR detection and enhance screening processes [11].

2.2 Review of Existing Research

Diabetic retinopathy (DR) screening mechanisms have come a long from manual clinical DR image evaluation to AI-based automated DR detection systems. Early work focused on conventional machine learning with hand-engineered feature extraction, while newer studies have emphasized deep learning models that extract the discriminative retinal features directly from fundus images. In addition, XAI has been recognized as an important factor for clinical trust and accountability. We

discuss reviewed studies under 5 branches: classical methods literature, CNNs in DR detection, transfer learning category, XAI in retinal imaging field and the newly proposed hybrid models that integrate accuracy-interpretability balance.

2.2.1 Traditional Methods and Early AI Applications in DR Detection

Traditionally, ophthalmologists visually examined retinal fundus images to detect any pathological indications of DR including microaneurysms, haemorrhages and exudates. Auto body fat segmentations have been proven to be a successful therapeutic solution although the manual grading is time consuming and subjective among observers, especially in mass screening environment. In order to save time from manual inspection, early automated systems utilized conventional machine learning models such as Support Vector Machine (SVM), k-Nearest Neighbours(k-NN) and Decision Trees for the DR detection and grading process[6]. These systems required handcrafted features describing colour, texture, and morphology, making them highly dependent on expert design and often poorly generalisable across datasets. Limited labelled data and variability in image acquisition further constrained their clinical applicability.

2.2.2 Deep Learning in DR Detection

Deep learning, particularly convolutional neural networks (CNNs), introduced a major shift in automated DR analysis by enabling models to learn hierarchical feature representations directly from raw pixels. CNN-based methods have shown strong performance in detecting DR severity stages and identifying clinically important retinal patterns such as microaneurysms, haemorrhages, exudates, and neovascularisation [5]. Several studies demonstrate that deep learning models can achieve expert-level performance on public benchmark datasets, supporting their potential use in screening programmes [1]. Compared to traditional approaches, CNNs generally provide superior accuracy and efficiency, and many works extend classification to lesion localisation by identifying abnormal regions within the fundus image [16] [15].

2.2.3 Explainable AI (XAI) in Medical Image Analysis

Although deep learning models can achieve high predictive accuracy, their black-box nature limits clinical adoption. As DR screening is a high-stakes medical application, interpretability is essential for transparency, accountability, and clinician trust. Explainable AI (XAI) methods aim to visualise and justify model predictions by highlighting image regions that contribute most to the decision. Techniques such as saliency maps, Class Activation Mapping (CAM), Grad-CAM, and attention-based visualisation have been widely used to support clinical validation of AI predictions [10]. In DR, XAI is particularly valuable because clinicians expect lesion-level evidence (e.g., haemorrhages, exudates, microaneurysms) to justify grading decisions. Therefore, integrating XAI into DR detection systems has become increasingly important to support trustworthy deployment [16].

2.2.4 Recent Hybrid and Modern Architectures for DR Grading

Recent research has progressed beyond standard CNNs by adopting modern architectures that improve robustness and representation learning [4], [10]. Hybrid convolution–attention models combine the local feature extraction strength of CNNs with global attention mechanisms to capture long-range dependencies, which is particularly useful for retinal images where lesions may be spatially distributed. Models based on transformer and attention mechanisms have been faring well on complex visual tasks, but most of time hybrid models are more stable for high-resolution inputs [15]. In addition, approaches that focus on interpretability concern more prominently the research field recently (so called interpretability-driven frameworks) and incorporate categorisation with lesion localisation and explainability techniques [5], [7]. Even with the increase in speed, addressing challenges of real-world datasets including highly imbalanced class distribution and varied image quality is still a challenge.

Such constraints motivate the approach followed in this thesis, which evaluates different state-of-the-art designs using a common pipeline, and integrates quality-aware balancing, class-aware optimization and lesion-level XAI to enhance the robustness and clinical relevance of an automatic system.

2.3 Summary of Key Findings

Based on the reviewed literature, several key insights guide this thesis:

- Deep learning models can identify grades of disease severity like microaneurysms, haemorrhages, exudates and neovascularisation. Nevertheless, in class-imbalanced data, usually the performance is deteriorated for minority classes like more severe stages
- Interpretable is a crucial requirement for clinical use. Deep learning models are widely criticized for being black boxes, despite their impressive precision. XAI methods such as Grad-CAM and attention heatmaps provide transparency by highlighting retinal regions that contribute to predictions, thereby enhancing clinical validation and confidence.
- In real applications, however, the imbalance of dataset and fluctuation of image quality are seldom handled. Public DR datasets (e.g., APTOS 2019) are highly imbalanced and also exhibit varying image quality owing to acquisition heterogeneity. These properties, which degrade robustness and constrain the generalisability of learned models in real-world settings—clearly call for a QualityAware learning approach.
- Clinically relevant assessment parameters are required. Moreover, accuracy is not enough for grading in DR, especially the imbalanced case. Metrics like weighted quadratic Cohen's kappa, sensitivity and specificity and confusion matrices are used to give a more clinically meaningful performance estimation particularly for the cases of severe and proliferative DR.

- The requirement of integrative, clinically directed pipelines. There is a growing interest in the literature on consolidating classification and interpretability. Nevertheless, few models have an inclusive pipeline that deals with imbalance, variance in image quality and lesion-level interpretation and even generates structured diagnostic reporting [11]. This motivates the hybrid architecture proposed in this dissertation.

2.4 Research Gap and Motivation

Although deep learning based diabetic retinopathy (DR) detection has achieved impressive success upon public fundus imaging datasets, there are extremely limitations restricting its clinical spread. Most of the existing approaches often emphasize on improving prediction accuracy rather than robust, interpretable and clinically dependable across a wide range of scenarios. A big gap in the class imbalance problem is present especially for popular datasets such as APTOS 2019 and the Diabetic Retinopathy detection dataset. These datasets have a skewed class distribution with majority of No DR and Mild cases, whereas Severe and Proliferative DR samples are scarce. Thus models that are trained using the same objectives will be biased towards majority classes and sensitivity for the most critical clinical stages is decreased [16]. This discrepancy makes it difficult to formulate grading systems for the detection of sight-threatening DR specifically in screening programs.

A second limitation is image-quality variation. Fundus images captured by different cameras with different settings mostly suffer from the blurriness, non-uniform illumination and low contrast or other artifacts. Either recent works ignore these quality differences and test their models on relatively clean benchmark distributions, or performance degradation would be suffered under real-world deployment scenario [14]. Since DR screening in the real world often takes place in low resources settings, it is desired for the system to be robust under varying quality conditions.

In addition, although their diagnostic performance is superior, many deep learning models are inscrutable black-box systems that limit clinical confidence and responsibility. In a healthcare setting, what clinicians need is much more than the predictions, but also evidences along with rationales broadly trusted by medical experts. Explainable AI (XAI) techniques, including Grad-CAM, saliency maps and attention visualisations have also helped mitigate this problem by highlighting image regions important to predictions [18]. However, most DR works consider interpretability as an optional add-on and do not naturally include it in the main workflow. This reduces transparency and obstructs uptake by health decision makers .

Furthermore, the majority of prior work evaluates one or two architectures in isolation, often using different preprocessing, loss functions, and evaluation settings. This makes comparisons inconsistent and prevents a clear understanding of which modern architectures are most suitable for real-world DR grading. Although CNN-based approaches have been widely explored [2] [5], recent advances such as hybrid convolution–attention networks and state-space vision models require systematic evaluation under the same dataset and training conditions.

Chapter 3

Proposed Methodology

3.1 Methodology Overviews

3.1.1 Final Specifications and Requirements

This thesis introduces a quality-aware and interpretable deep learning system for the classification of five categories of diabetic retinopathy (DR) utilising 2D colour fundus images. The system serves as a research-oriented prototype that enables scalable diabetic retinopathy screening by producing (i) predictions of DR severity, (ii) visual explanations at the lesion level, and (iii) structured diagnostic summaries. Unlike DR classifiers that only care about accuracy, the design focuses on being able to handle class imbalance and changing image quality. It also includes elements that make it easier for doctors to understand and trust the system.

The proposed system is trained and tested on publicly available datasets namely APTOS 2019 Blindness Detection and Diabetic Retinopathy Detection Kaggle (DR Datasetand), which are then combined to form a large dataset. The method is based on a five-class DR grading scheme (No DR, Mild, Moderate, Severe Non Proliferative DR and Proliferative DR) that captures the clinically meaningful ordinal structure of DR severity. Since DR screening is a safety-critical medical task, the system is deliberately cast as decision-support tool rather than a replacement for expertise.

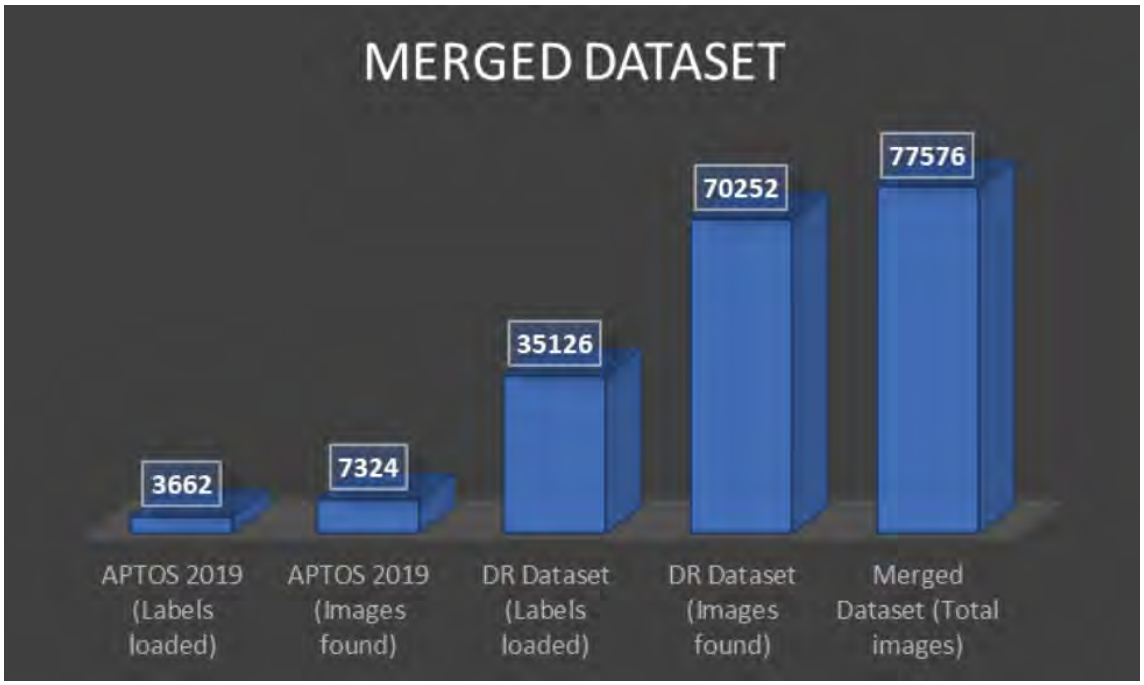

Figure 3.1: Merged Dataset

3.1.2 Functional Requirements

The functional requirements define what the system must accomplish in terms of inputs, outputs, and core capabilities. The system operates on fundus images and produces both predictive and explanatory outputs. To ensure alignment with clinical screening needs, the following functional requirements are specified:

- **Fundus Image Input Handling** The system should be capable to process 2D colour retinal fundus images and computationally tolerant to resolution/image acquisition parameters. To make it compatible to all models and prevent unstable inference, we preprocess the image for standardization.
- **Five-Class DR Severity Grading** At a minimum, the underlying predictive module must return a DR grade based on five severity levels. It is crucial that the system optimizes for accurate identification of clinically significantly categories (Severe and Proliferative DR) with the highest referral urgency.
- **Quality-Aware Learning and Robustness Support** The system should be invariant to the image quality change, including defocused images, low contrast images, variation of illumination condition and camera imperfect. This criterion is important as screening images are commonly taken in diverse scenarios, which are not necessarily part of the benchmark distributions.
- **Imbalance-Aware Training Behaviour** As a result of combined public datasets are still highly imbalanced (majority No DR/Mild vs minority Severe/Proliferative), the system has to include class imbalance management approaches. The goal is to enhance the minority-class sensitivity without directing the performance based on a wrong training assumption.

- **Explainable AI Output Generation** The algorithm should generate a lesion-level visual explanation for each of the prediction using XAI methods, Grad-CAM and attention-based visualization. The clarification should highlight trends in regions of the retina that are relevant to the expected grade, which can help physicians assess and confirm results.

- **Prototype Diagnostic Report Generation** The system needs to generate a structured written report concerning the predicted grade, confidence, and lesion-related evidence from XAI outputs. This standard is applicable to any screening process where a reportable interpretation must be given in an explicit and consistent manner.

3.1.3 Non-Functional Requirements

The performance, dependability, interpretability, and repeatability of the system are defined by the non-functional requirements. When assessing AI systems for healthcare, these factors take on further significance because accuracy alone is not enough.

Performance and Evaluation Requirements

Given the ordinal characteristics of diabetic retinopathy grading, it is essential to utilise therapeutically pertinent criteria for system evaluation rather than absolute accuracy. Quadratic Weighted Cohen's Kappa (QWK) serves as the primary evaluation metric, considering agreement and imposing greater penalties on distant misclassifications compared to those that are closer. In clinical screening, it is less critical to conflate The distinction between Moderate and Severe is problematic, particularly in conflating Severe with No DR; this risk-informed framework raises concerns for QWK.

The evaluation criteria include:

- Primary metric: Quadratic Weighted Cohen's Kappa (QWK).

- Secondary metrics: accuracy, ROC-AUC, per-class precision/recall/F1-score, sensitivity, specificity.

- Diagnostic analysis: confusion matrices with emphasis on severe/proliferative performance.

The experimental findings demonstrate that the proposed unification pipeline achieves clinically acceptable performance across multiple architectures. Among the evaluated models, CoAtNet exhibited the strongest agreement with expert annotations, attaining a quadratic weighted kappa (QWK) of 0.902. The Hybrid CoAtNet-ConvNeXt architecture also demonstrated robust performance with a QWK of 0.8944, followed closely by ConvNeXt alone at 0.8667. These results validate the method's capacity to accurately classify images into five distinct severity grades, indicating its potential utility for clinical deployment.

Reliability and Robustness
A robust diagnostic system must maintain consistent performance across varying image quality conditions. The quality-tier analysis conducted in this study reveals that image quality significantly impacts model reliability. Specifically, poor-quality images (Tier D) resulted in an accuracy decline of 0.4727 compared to high-quality images (Tier A) while simultaneously reducing explanation entropy by 0.3533. This reduction in entropy indicates compromised predictive confidence and diminished model interpretability when processing degraded inputs.

These findings underscore the critical importance of implementing rigorous data preprocessing protocols and quality-aware stratification strategies. Furthermore, they highlight the necessity of exercising caution when interpreting model predictions on suboptimal image inputs as such cases may reflect data quality limitations rather than true diagnostic capability.

Interpretability and Transparency
Explanatory models are being treated today as one of the most genuine system's needs and not only nice fun. XAI outputs should be visually interpretable, clinically meaningful with focus on lesions or lesion-proximal areas and not unrelated artifacts. Since XAI techniques may sometimes uncover spurious correlations, the system should provide explanations only as supporting evidence, not causal proof.

Usability and Maintainability
This is not entirely hospital integration but the design solutions must be modular, maintainable. The preprocessing, classifier, XAI, and reporting modules should be decoupled so they can be updated independently, without performing a full rewrite.

Reproducibility
Each experiment should be reproducible, with defined dataset splits, uniform preprocessing, recorded hyperparameters and saved model checkpoints. Reproducibility is important for validating science, and comparing with future research.

3.2 Societal Impact

Diabetic retinopathy (DR) is the leading cause preventable blindness and population based screening is essential as early stages are often asymptomatic. But screening programmes face resource constraints, especially in low- and middle-income countries with few ophthalmologists but large numbers of patients. Computer-aided DR grading systems can alleviate the screening workload by identifying potentially sight-threatening cases earlier and supporting initial referral decisions.

The method proposed in this thesis contributes to this direction by A) enhancing the scalability of screening, and B) enabling more consistent grading through automatization. Most importantly, the proposed method is interpretable, i.e., it is of necessity to avoid resistance from clinicians and to encourage responsible treatment decisions based on healthcare predictions. Rather than raw black-box predictions,the technology satisfies a clinical transparency requirement by providing lesion-level heatmaps and structured summaries.

Nonetheless, societal influence implies risk. A false-negative result in severe or proliferative diabetic retinopathy (DR) can delay referral and significantly increase the risk of irreversible vision loss. Therefore, this system should be used strictly as a clinical decision-support tool with clearly communicated limitations and it must undergo rigorous clinical validation and prospective evaluation before being considered ready for real-world deployment.

3.3 Environmental Impact

Deep learning models, and particularly hybrid and transformer architectures, can be computationally expensive to train. This paper compares five high-throughput designs in a " one pipe way ", which is more costly on the computation scale. To mitigate environmental impact, the approach leverages strategies such as transfer learning, mixed-precision GPU training and early stopping with learning rate scheduling. These optimizations enhance the efficiency of training and the sustainability of research quality assessment.

3.4 Ethical Issues

Safety, fairness, transparency and accountability are among the ethical questions raised by medical AI. The errors in DR grading can be harmful since it has effect on the referring urgency and also for clinical management. Ethical concerns posed by this topic include issues of dataset bias, interpretability, performance and misuse.

Another issue is the bias and nonrepresentativeness of public data sets. APTOS 2019 and DR Dataset were not fully inclusive of global demographics, which may limit thegeneralizability to other populations. Moreover, in the context of imbalanced class distribution and minority severe classes being under-detected, predictions with bias can be made – not ethically desirable as it discriminates against high-riskpopulations (i.e., that are sick or injured).

Interpretability has ethical concerns. The XAI results may enhance interpretability, but they could mislead one to attribute causality. We need to treat heatmapsas signals, in favour of an idea not as proofs. This paper eliminates these risks by establishing the XAI side and exploring the explanation factor if it is not so good.

Finally, in this work we only use freely available and de-identified publicly available data and we do not process any patients' identifiable information. Ethical concerns still require that results are reported along with these statements in the hope of emphasizing that the reporting contrivance is not a diagnostic tool, and never should be used to replace expert medical decision.

3.5 Project Management Plan

The research followed a structured seven-phase methodology to ensure systematic development, rigorous evaluation, and clinical relevance:

3.5.1 Phase 1: Dataset Acquisition and Integration

Merged APTOS 2019 and Diabetic Retinopathy Detection datasets (77,576 images total), harmonized labels to five-class DR severity scale, and conducted exploratory analysis to characterize class imbalance and quality variation.

3.5.2 Phase 2: Quality Assessment and Preprocessing

Developed automated quality scoring system (four-tier A-D classification), implemented tier-specific preprocessing strategies (CLAHE, denoising, gamma correction), and established quality-stratified dataset splits.

3.5.3 Phase 3: Model Training

Trained five architectures (ConvNeXt, CoAtNet, Hybrid, MaxViT, Vision Mamba) under unified pipeline with focal loss, transfer learning, and early stopping. Maintained consistent preprocessing and hyperparameters for fair comparison.

3.5.4 Phase 4: Performance Evaluation

Evaluated models on held-out test set using Quadratic Weighted Kappa (QWK), accuracy, and per-class metrics. Generated confusion matrices and conducted comparative analysis prioritizing sight-threatening DR detection (Class 3/4 sensitivity).

3.5.5 Phase 5: Explainability Analysis

Implemented XAI techniques (Grad-CAM, Integrated Gradients, Occlusion Sensitivity), analyzed quality-tier robustness, and discovered entropy paradox (accuracy drops 47%, attribution entropy drops 35% from Tier A$\rightarrow$D).

3.5.6 Phase 6: Report Generation Framework

Designed and prototyped rule-based automated report generation module, created sample outputs for different quality tiers and DR grades, validated deterministic mapping from predictions to clinical summaries.

3.5.7 Phase 7: Documentation and Finalization

Compiled thesis documentation, conducted revision cycles based on supervisor feedback, finalized presentation materials, and organized code repository for reproducibility.

3.5.8 Quality Control:

All phases incorporated validation checkpoints (data integrity checks, reproducibility protocols with fixed random seeds, clinical relevance verification through XAI anatomical plausibility). The unified training pipeline ensured fair model comparison through identical preprocessing, loss functions, and evaluation metrics.
This phased workflow ensured that all design decisions were guided by the dataset characteristics as well as relevant clinical requirements.

3.6 Risk Management

Several risks were identified during system development:

- Severe class imbalance leading to reduced sensitivity for Class 3/4 (Severe/Proliferative DR)
- Label inconsistency and grading noise after merging APTOS 2019 and DR Dataset
- Image quality variability (Tier A–D) causing performance degradation and unreliable predictions
- Domain shift across datasets/capture conditions affecting real-world generalisation
- False-negative detection of sight-threatening DR causing delayed referral and clinical risk
- XAI explanations highlighting non-lesion regions, reducing clinical trust and interpretability reliability
- Limited prospective clinical validation, restricting readiness for real-world deployment

3.7 Economic Analysis

From economical viewpoint, systems for screening diabetic retinopathy (DR) tend to reduce the workload of physicians and patients to be monitored more frequently because DR has been found to be expensive disease. Automated DR grading methods can be cost-effective to the health systems by expediting screening and reducing workload to provide a quick response for patients with high risk content. The report writing tool in the application generates summary structured reports and saves time on typing.

But to develop and use models requires technical infrastructure and computational power. Such activities were not simulated in thisthesis but the framework illustrates a possible cost-benefit trajectory: once real-world experience AI is trained, inference is substantially cheaper than repeated manual grading, making AI-assisted screening financially viable in large scale programs.

Chapter 4

Preliminary Process and Design

4.1 Design Process

This thesis follows a quantitative, experimental research design to develop and evaluate a clinically oriented deep learning framework for five-class diabetic retinopathy (DR) grading using 2D colour fundus photographs. Diabetic retinopathy is a progressive disease that can remain asymptomatic in its early stages, making screening essential for preventing irreversible vision loss [16]. However, manual grading of fundus images is time-consuming and subject to intra- and inter-observer variability, limiting scalability in large screening programmes, especially in resource-limited environments [16]. These constraints motivate the development of automated, reliable, and interpretable DR grading systems.

The proposed framework is designed to address three real-world deployment barriers commonly observed in automated DR screening research: severe class imbalance, heterogeneous image quality, and limited interpretability of high-performing deep learning models [15]. Instead of focusing solely on classification accuracy, the system integrates robust learning strategies and explainable AI (XAI) mechanisms to improve clinical trust and diagnostic transparency.

The methodology follows an end-to-end pipeline consisting of: (i) dataset acquisition and merging, (ii) class imbalance and quality distribution analysis, (iii) preprocessing and enhancement, (iv) imbalance-aware training under a unified protocol, (v) systematic evaluation of five modern vision architectures, (vi) explainability analysis using lesion-level visualisations, and (vii) prototype diagnostic report generation. Unlike many studies that evaluate models under different preprocessing and training settings, this thesis compares all architectures under the same pipeline to ensure fair and meaningful comparison

4.2 Model Specification

The proposed system is developed as a modular architecture, where each component addresses a specific screening challenge and supports reproducibility. The design is informed by the clinical grading procedure, where DR severity depends on both local lesion evidence and global lesion distribution across the retina [12]. Therefore,

the framework combines robust feature extraction with interpretability mechanisms that allow clinicians to inspect model reasoning.

4.2.1 Design Objectives and Constraints

The primary design objectives are:

- To achieve reliable five-class DR grading aligned with clinical severity ordering.
- To improve detection sensitivity for severe and proliferative DR, which are clinically urgent stages.
- To increase robustness under heterogeneous image quality, reflecting real screening conditions.
- To integrate lesion-level interpretability through XAI methods such as Grad-CAM and attention visualisation [8].
- To generate structured, screening-oriented outputs through a prototype reporting module [9].

The system is constrained by the use of retrospective public datasets and does not include prospective clinical trials, hospital integration, or clinician user studies. Interpretability evaluation is therefore conducted through qualitative and quantitative XAI analysis rather than clinical usability scoring.

4.2.2 Architecture Selection

To ensure modern and representative evaluation, five architectures were selected:

- ConvNeXt, representing modernised convolutional networks
- CoAtNet, representing hybrid convolution-attention modelling.
- Hybrid CoAtNet–ConvNeXt, designed to combine local lesion modelling with global context reasoning.
- MaxViT, using window-based attention for scalable vision modelling.
- Vision Mamba, using state-space modelling for long-range dependency learning.

All architectures are trained under identical preprocessing, loss, and evaluation settings, enabling consistent comparison.

4.2.3 Hybrid ConvNeXt–CoAtNet Architecture with Learnable Gated Fusion

To improve DR grading robustness and ordinal agreement, this thesis proposes a hybrid model that combines a pretrained ConvNeXtSmall backbone with a CoAtNet-inspired convolution–attention branch. This design is motivated by the observation that DR severity grading depends on both (i) local lesion-level patterns (e.g., microaneurysms, haemorrhages, exudates) and (ii) broader contextual evidence such as lesion spread and distribution across retinal quadrants [12]. Convolutional networks are effective at capturing fine lesion textures, while attention-based mechanisms support global reasoning by modelling long-range spatial dependencies [8].

Dual-Branch Feature Extraction
The hybrid model processes each fundus image $I \in R^{H\times W\times 3}$ through two parallel branches:

ConvNeXt branch (local lesion representation): A pretrained ConvNeXt-Small model (ImageNet weights) is used as a deep convolutional feature extractor. The classification head is removed, producing a high-level spatial feature map. Global average pooling compresses this map into a fixed-length vector, which is projected into a 512-dimensional embedding space using a dense projection layer and layer normalization:

$$\mathbf{f}_{conv} \in R^{512}$$

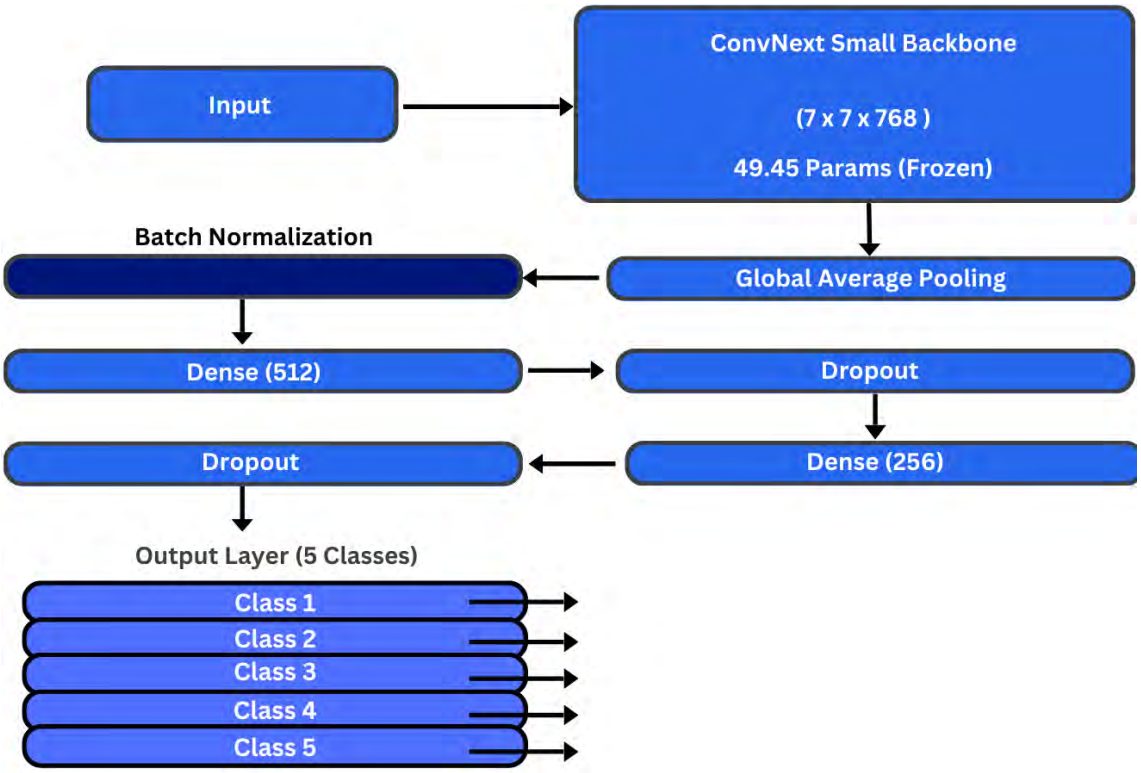

Figure 4.1: ConvNext

This branch provides strong sensitivity to lesion boundaries, vessel structures, and texture patterns that are critical for early and intermediate DR grading.

CoAtNet-inspired branch (contextual/global reasoning): The second branch begins with a convolutional stem consisting of strided convolutions, batch normalization, and GELU activations to downsample and stabilise feature extraction. A depthwise convolution block with residual connection refines local structures while preserving information flow. The resulting feature map is converted into a token sequence by reshaping:

$$(B, H, W, C) \rightarrow (B, HW, C)$$

Two Transformer encoder blocks (multi-head self-attention and feed-forward networks) are applied to model long-range dependencies and integrate evidence across retinal regions. After token-level processing, global average pooling produces a compact embedding which is projected to the same fusion dimension:

$$\mathbf{f}_{coat} \in R^{512}$$

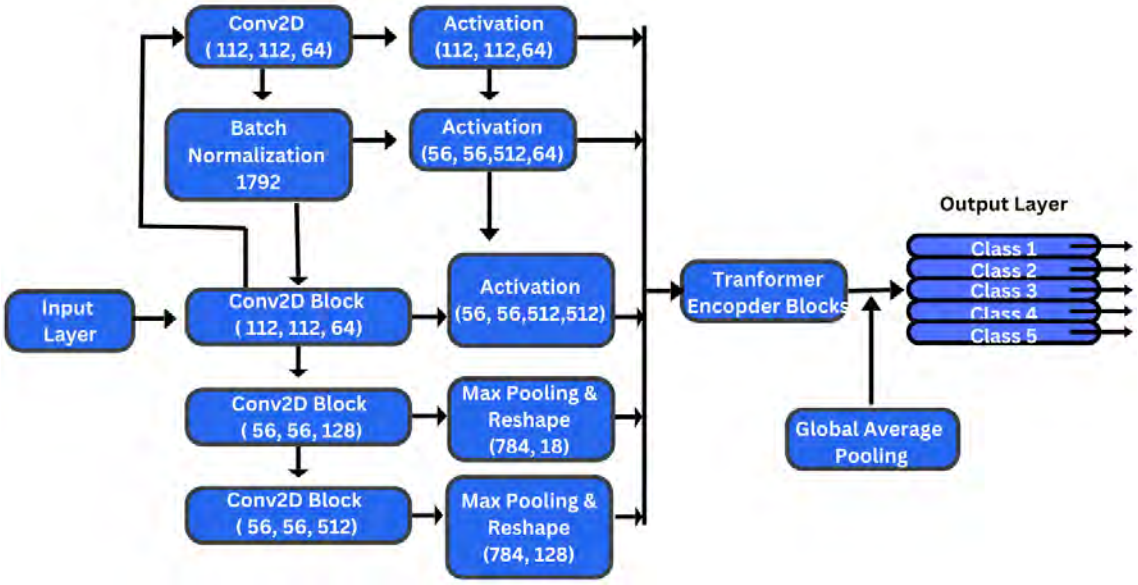

Figure 4.2: CoatNet

Learnable Gated Fusion
Instead of fixed fusion, the model uses a sigmoid gating mechanism that adaptively weights the contribution of each branch per image. The gate g is computed from the concatenated embeddings:

$$g = \sigma\left(W\left[\mathbf{f}_{conv}; \mathbf{f}_{coat}\right] + b\right), \quad g \in (0, 1) \tag{4.1}$$

The final fused representation $\mathbf{f}$ is formed as:

$$\mathbf{f} = g \cdot \mathbf{f}_{conv} + (1 - g) \cdot \mathbf{f}_{coat} \tag{4.2}$$

This design is particularly relevant for screening settings where image quality varies, allowing the model to dynamically balance local lesion sensitivity with global context reasoning.

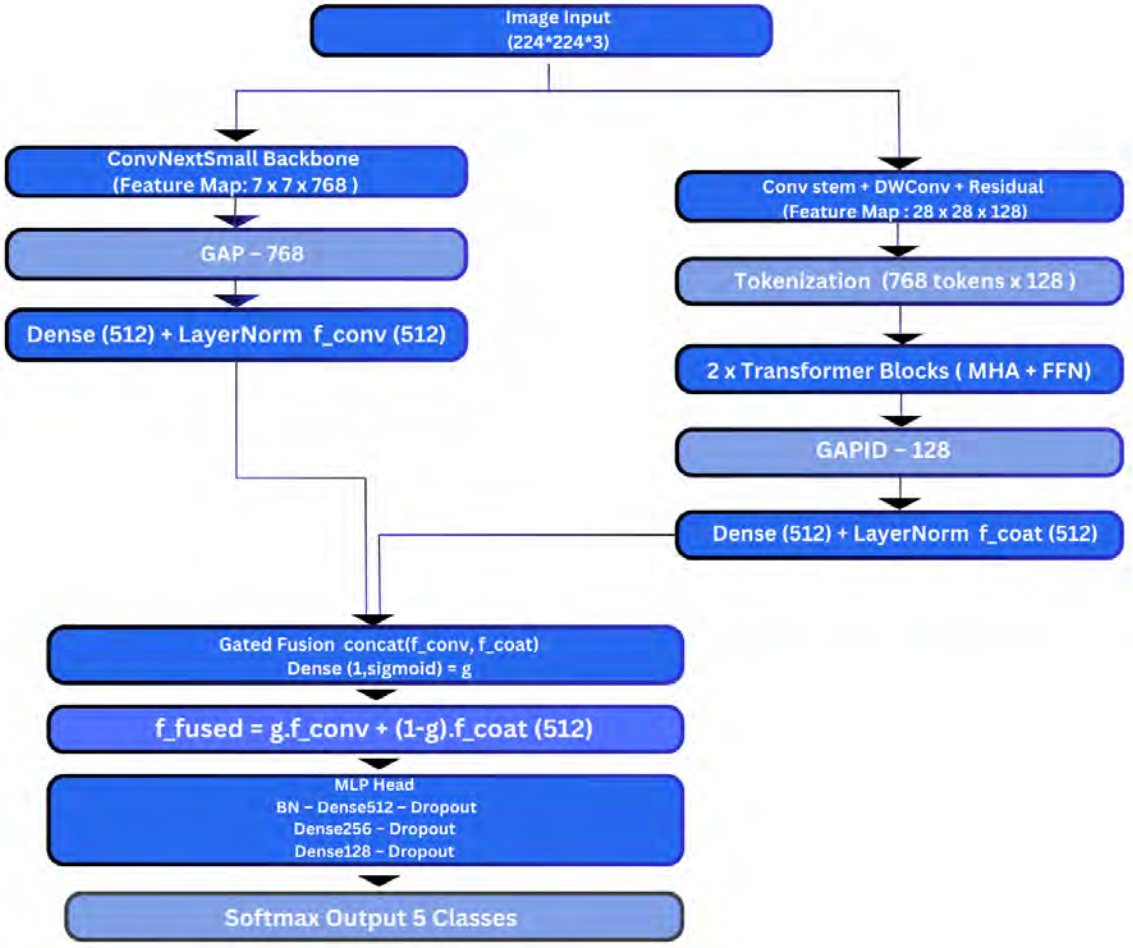

Figure 4.3: Hybrid

Classification Head The fused embedding is passed through a regularized multi-layer perceptron (batch normalization, GELU dense layers, and dropout) to reduce overfitting and produce the final five-class probability distribution using softmax.

Training Objective and Ordinal Evaluation
The hybrid model is trained using focal loss, which reduces majority-class dominance and emphasises difficult examples, improving minority-class learning. Performance

is evaluated primarily using Quadratic Weighted Cohen's Kappa (QWK) to reflect ordinal severity ordering. Under the unified pipeline, the hybrid model achieved strong ordinal agreement (QWK = 0.8944), supporting the effectiveness of gated fusion in DR grading.

4.2.4 Training Strategy Overview

Training is performed under a unified protocol across all architectures. Transfer learning is used where applicable, as pretrained models improve convergence and generalisation on medical imaging tasks [19]. Mixed-precision GPU training is employed to reduce memory overhead and accelerate training. Early stopping, checkpointing, and learning rate scheduling are applied to prevent overfitting and stabilise convergence [11]. Class-aware learning strategies (such as focal loss and imbalance-aware sampling) are used to improve performance on rare severe/proliferative classes.

4.3 Data Collection

4.3.1 Data Sources

This study uses two publicly available datasets:

- APTOS 2019 Blindness Detection dataset
- Diabetic Retinopathy Detection Kaggle dataset (DR Dataset)

Both datasets contain labelled fundus images graded on the standard five-class DR severity scale.

4.3.2 Dataset Merging and Label Harmonization

Labels are normalized to ensures consistent encoding of images for the five categories. The aggregation improves the representation of diverse retinal features and capturing situations, leading to better generalization.

4.3.3 Quality Assessment of Fundus Images:

Quality of the combined fundus dataset was carefully assessed based on various criteria such as blur, brightness, contrast and usable retinal area to ensure reliable training and validation. This quality analysis led to an identification of differences across data sets and grades of DR, reiterating the quality-aware pre-processing and robustness-focused design of this work.

4.3.4 Class Imbalance and Dataset Characteristics

The combined dataset is heavily imbalanced in terms of class, with far more No DR and Mild examples compared to Severe and Proliferative. Such discrepancy is an actual screening case distribution while introduces a learning bias and justifies the consideration of imbalance-based training [16]. In addition, photos suffer from inequality in quality due to differences between cameras, blurs, illumination changes and image artifacts; thus requiring strict preprocessing and quality-aware evaluation

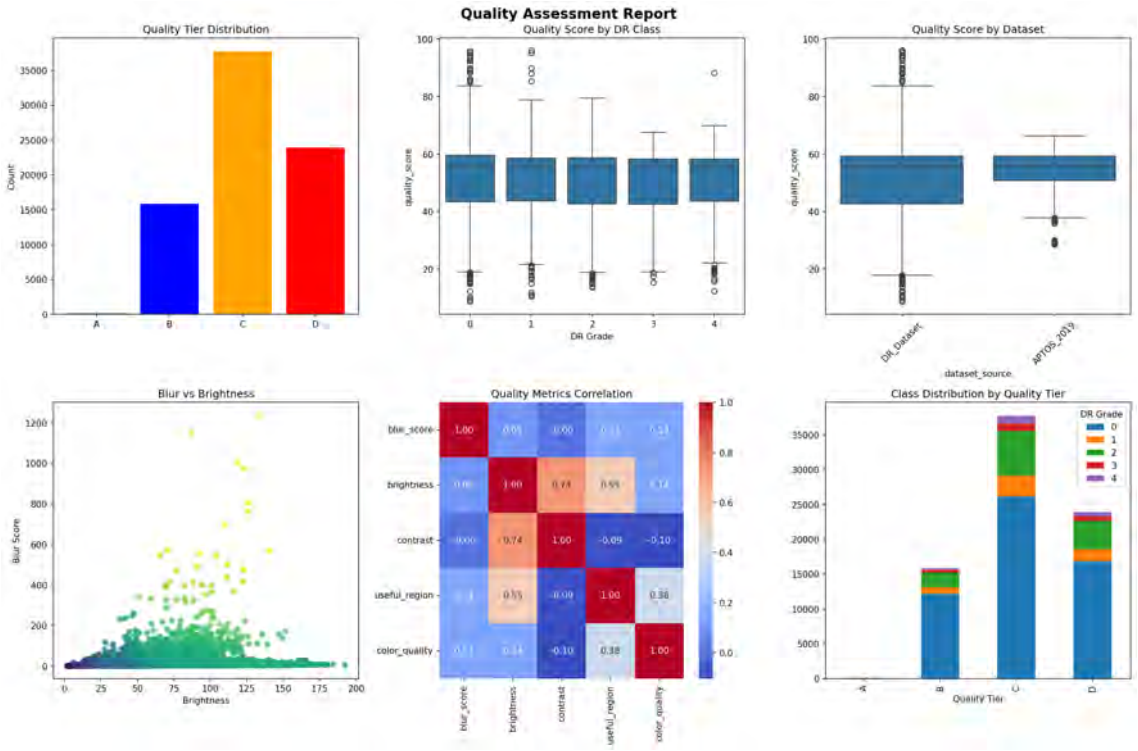

Figure 4.4: Quality Assessment Report

4.3.5 Data Cleaning

Images are validated for integrity, and corrupted/unreadable samples are removed. Label validity is verified to ensure all samples fall within the five-class grading scheme. Duplicate samples across datasets are checked to avoid leakage between training and evaluation splits.

4.3.6 Data Transformation (Preprocessing)

The included retinal fundus dataset has significant variations in illumination, clarity, contrast and color due to the differences in acquisition conditions that may reduce lesion visibility and limit model generalization. Therefore, a comprehensive preparation pipeline was adopted to normalize image appearance and enhance DR factors before training [12]. All photos were resized to a common resolution, and an automatic quality-assessment module (evaluating the blur, brightness, contrast, usable retinal region) and color quality phase was performed. Based on the determined qualityscore, the images were divided into four quality groups (A-D). Tier-specific post-processing was subsequently applied: minimal amount of CLAHE and sharpening for high-quality images, and increased denoising, gain-contrast enhancement (CLAHE/histogram equalization), gamma correction, and edge sharpening to improve visualization of vascular structures/lesions in lower quality images. These improve the validity of diabetic retinopathy grading and its readability under real-world picture variance [9], [12], [15].

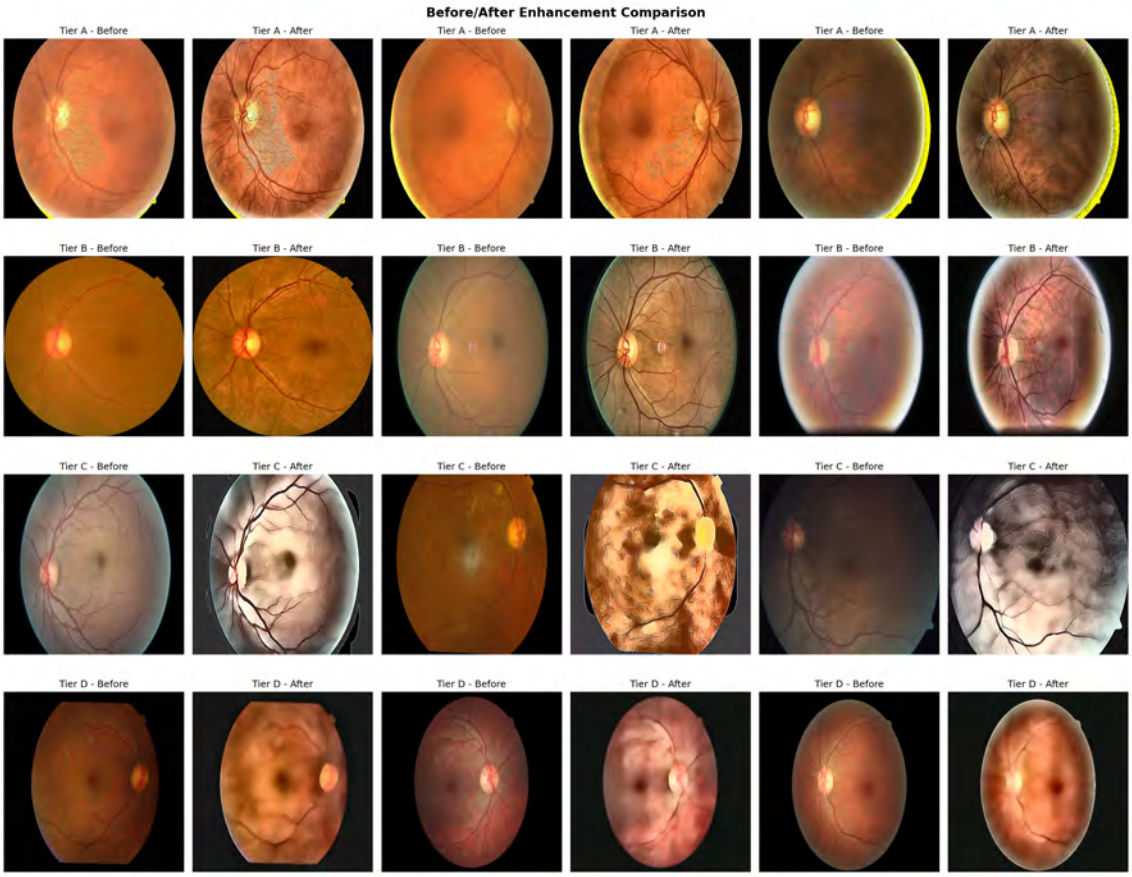

Figure 4.5: Enhancement comparison

4.3.7 Data Integration

After preprocessing, every single augmented image was linked to its own DR grade label and assigned a unique identifier for the purpose of keeping track over both set of datasets. The final training dataset was organized into class-specific directories (class0-class4) for easy and straightforward loading during model training. This embedding ensured that the input format was consistent, facilitated the use of the dataset and preserved a systematic mapping of photos, labels, quality tiers and file locations for reproducible testing [9], [12].

4.3.8 Summary of Preprocessed Data

The processed dataset contains scaled and quality-augmented retinal fundus images with their associated paired (i) DR severity labels, (ii) dataset source, (iii) computed quality metrics, and (iv) path to the final processed image. The multi-level enhancement approach ensures that the low-quality scans would be further enhanced, whereas high-quality ones would not be excessively processed for enabling robust training, and thus more clinically interpretable explanations in later XAI [12], [15].

4.4 Implementation of Selected Design

4.4.1 Training Implementation

In order to make sure all the models are being fairly judged, they went through similar settings. The training part is using focal loss, GPU acceleration and QWK for imbalance, mixed precision and metrics. Training callbacks, including checkpointing, early stopping, and learning rate scheduling, facilitate stable optimization and reproducibility.

4.4.2 Model Evaluation and Comparative Analysis

Profiling results exposed performance delta between different topologies in the grading task. CoAtNet ranked first based on a QWK score of 0.902 showing superior ordinal agreement. This was closely followed by the hybrid CoAtNet–ConvNeXt with a score of 0.8944, showing strong severity grading ability. ConvNeXt also outperformed the reported QWK of Vision Mamba (0.6983) and MaxViT (0.1111). In general, convolution-attention hybrids are superior to classical architectures in such a mixed dataset scenario.

4.4.3 Explainability (XAI) Implementation

Interpretation is being done with the help of attention-based visualization so that it highlights the lesion regions to further help on the predictions [18]. After being implemented, the results show promising decrease through A to D by 0.4727 while the decrease through attribution entropy by 0.3533, entailing the prediction reliability of the low-quality images.

Chapter 5

Performance Evaluation, Comparative Analysis and Discussion

5.1 Performance Evaluation Framework

In this chapter, we have presented the performance of our method on to what extent our method works well for five-class classification for DR grading. For a fairer comparison, we will evaluate using the same pipeline across 5 most-recent architectures : ConvNeXt, CoAtNet, Hybrid CoAtNet-ConvNeXt, MaxViT and Vision Mamba. As DR grading is an ordinal classification task, we need to measure the correctness of misclassification severity. For instance, predicting "No DR" in a case of proliferative DR is much worse than confusing neighboring levels of severity.

Therefore, QWK will be of high priority as the primary evaluation metric in this thesis. QWK is frequently used in DR grading challenge benchmarks since it penalizes large grade differences between scores more than near-grade confusion, which has clinical significance. In addition to the QWK, overall accuracy, class-wise precision/recall/F1-score and confusion matrix have been reported analysis for an easy interpretation of model behavior against all severity categories.

5.2 Test setup and protocol

All the models are trained in an ensemble using APTOS 2019 and DR Dataset data concatenated and evaluated with identical preprocessing, training schedule, metric computation method. To maintain the distribution of severity and to prevent sampling bias, we split the dataset into training, validation and testing. We employ the held-out test set only for final evaluation to guarantee fair comparability.

To address imbalance in training, we adopt class-aware learning methods with the help of focal loss. This reduces the over-sampling of high-represented classes and encourages learning from minority or challenging samples. Training was monitored using validation QWK, and best model checkpoints were selected according to the highest validation kappa.

5.3 Overall Performance Results

The relative performance of the five architectures investigated is largely presented QWK wise. Ordering agreement The results (Table 3) show that there are substantial differences in the relationship of ordinal agreement:

- For QWK indicator in Table 3, the highest value is achieved by CoAt-Net (best at ordinal consistency and overall grade reliability), which means its grading capability to differentiate progression types.
- Hybrid CoAtNet-ConvNeXt achieved a score of QWK = 0.8944 which means once again very close to the best performance and high sensitivity for high-grade DR.
- ConvNeXt was able to reach a QWK of 0.8607, which is an indicator for high performance but less than attention enhanced hybrids.
- Vision Mamba QWK was 0.6983, suggesting moderate agreement and poor discrimination in early-stage.
- MaxViT attained a QWK of 0.1111, representing weak ordinal consistency and random category behaviour.
- There results strongly imply that architectures translating convolutional inductive bias along with attention based global modeling (CoAtNet and Hybrid) are more proper ones for five-class DR grading on a merged dataset environment.

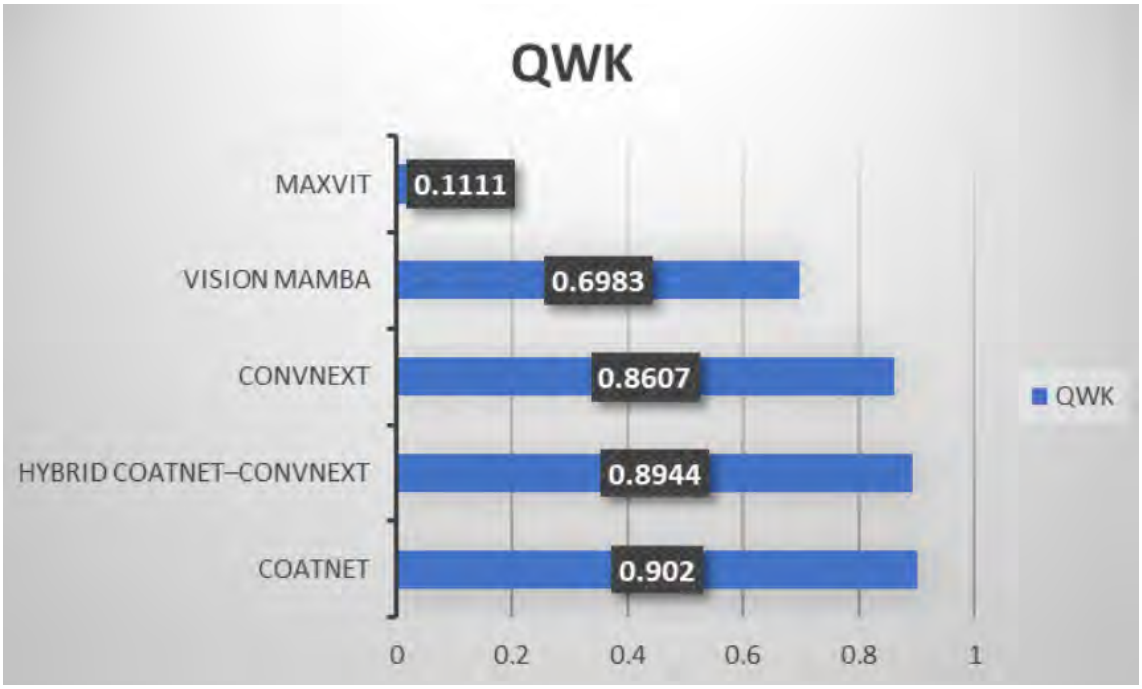

Figure 5.1: Overall QWK

These results strongly suggest that architectures combining convolutional inductive bias with attention-based global modelling (CoAtNet and Hybrid) are better suited for five-class DR grading in the merged dataset setting.

5.4 Model-Wise Detailed Performance Analysis

5.4.1 Hybrid CoAtNet-ConvNeXt (QWK=0.8944)

The Hybrid model is effective overall, it achieves 80.32% of accuracy on the test set with 0.8100 averaged F1-score (macro). Its per-class statistics indicated particularly good performance on clinically important epochs.

Class 3 (Severe DR): precision = 0.9833, recall = 0.8619, and F1 = 0.9186.

Class 4 (Proliferative DR): precision = 0.9716, recall = 0.8935 and F1 = 0.9309.

This is a highly pertinent characteristic for a screening setting as these classes represent sight-threatening DR that requires urgent referral. The confusion matrix also indicates that the misclassifications are mainly for contiguous grades, which can be useful from the clinical point.

The role of the particular classes follows-
Class 2 (Moderate DR) has a fairly high recall (0.7862) but relatively low precision (0.5636), which means the model identifies moderateDR more by "overfitting" than under-fitting in borderline situations. This is a common problem in the grading of DR because severity and distribution of lesions overlap.

The interpretation of the confusion matrix also supports this: many Class 1 samples are predicted as Class2 (858 samples) and few Class 2 samples are predicted as Class 3 (340 samples). This shows that the Hybrid model is extremely sensitive to lesion data and that it may sometimes overestimate severity, which in a clinical setting is safer than underestimating disease severity.

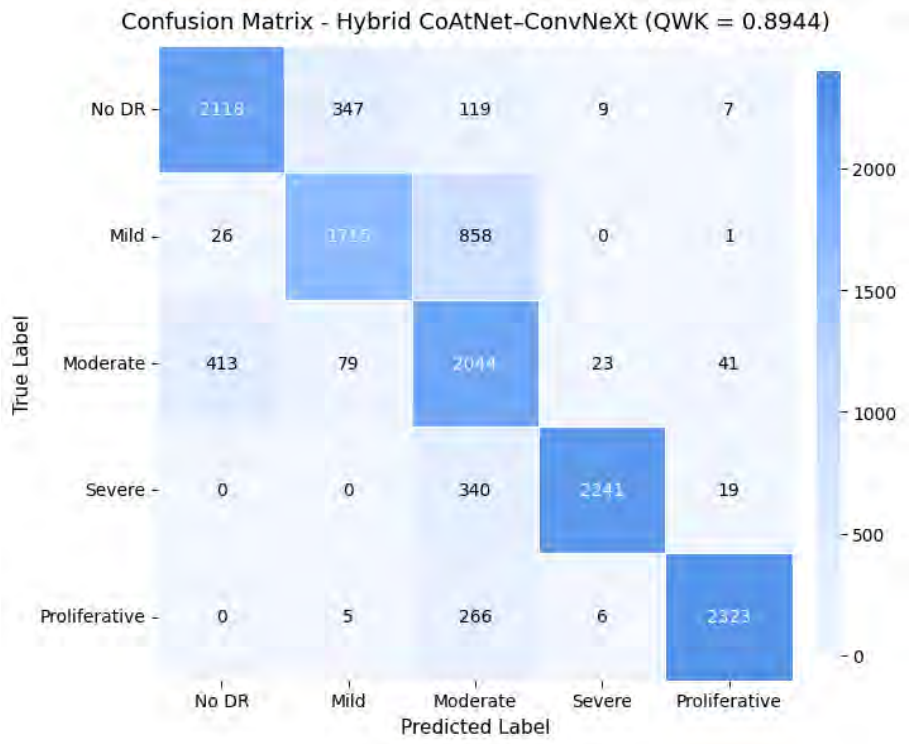

Figure 5.2: Confusion Matrix: Hybrid CoAtNet-ConvNeXt

As a whole, the Hybrid model provides one of the best trade-offs between sensitivity for high-grade DR and ordinal-consistent grading.

5.4.2 ConvNeXt (QWK = 0.8607)

ConvNeXt was surprisingly better for the task with an accuracy of 78.72% and a macro-average F1-score of0.7912. The model is doing great in high-grade DR classes.

Class 3 (Severe DR): precision = 0.9578, recall = 0.8727, F1 = 0.9133.

Class 4 (Proliferative DR): precision = 0.9245, recall = 0.9038, F1 = 0.9146.

These results show ConvNeXt's better ability to capture lesion-level patterns, vascular changes and retinal texture abnormalities. The confusion matrix demonstrates that most misclassifications are between neighboring grades; especially between Class 1 and Class 2, reflecting conflicting diagnostic interpretations of mild and moderate DR in practice.

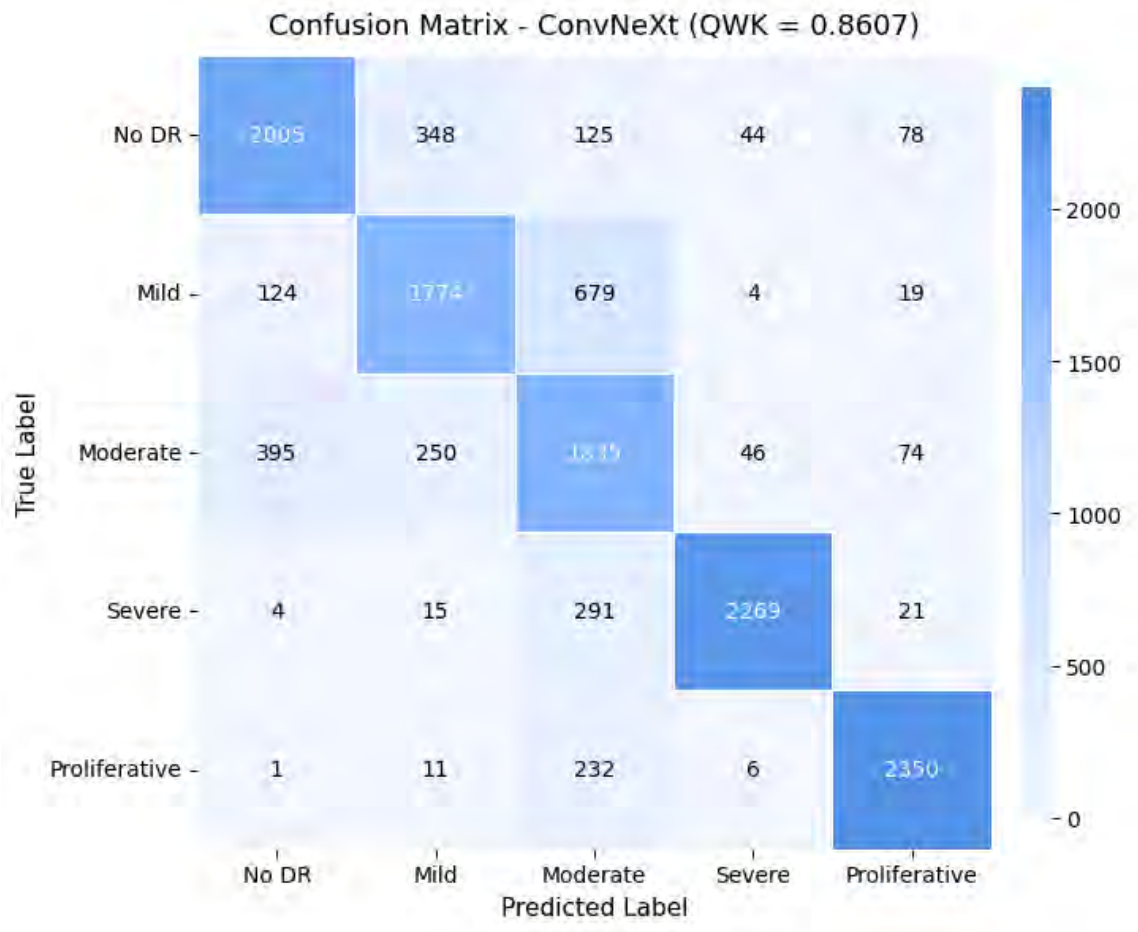

Figure 5.3: Confusion Matrix: ConvNeXt

ConvNeXt has a moderate lower ordinal agreement comparing with the Hybrid model (QWK 0.8607 vs 0.8944), suggesting that incorporating attention-based context can help grade more consistently across the range of severity.

5.4.3 Vision Mamba (QWK = 0.6983)

In the case of Vision Mamba, he was a so-so performer with 64.48% accuracy and macro-average F1-score set at 0.6339. The greatest weakness is that it failed to detect Mild DR efficiently.

Class 1 (mild DR): precision = 0.5592, recall = 0.2854, F1 = 0.3779

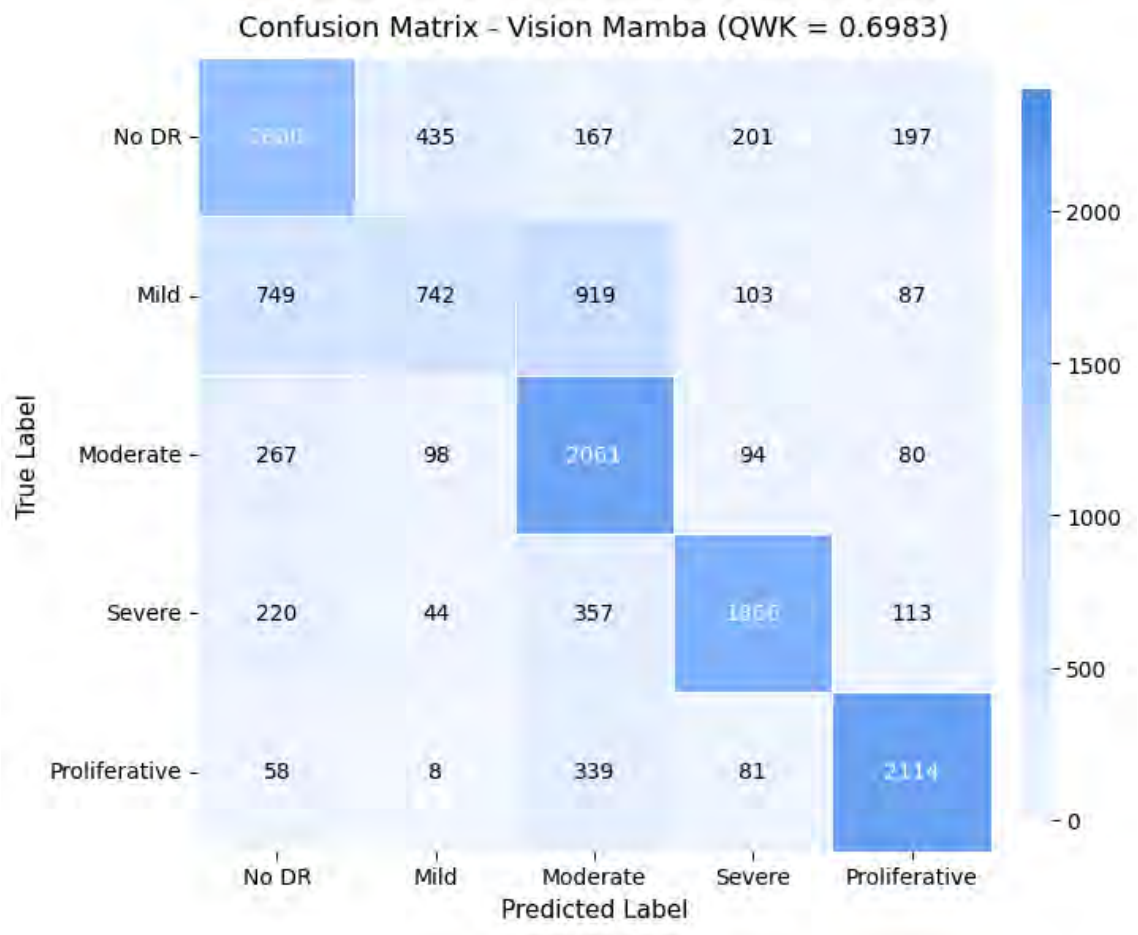

Figure 5.4: Confusion Matrix: Vision Mamba

This indication implies that the model does not well predict a large proportion of Class 1 (DR patients at early stage), and tend to give it as No DR or Moderate DR in most cases; The confusion matrix reveals that class 1 samples are usually predicted as class 0 (749 examples) or class 2 (919 examples). This behavior is not desirable for screening as it is necessary to detect mild DR in order to allow prevention and follow-up.

But Mamba does better at later stages. Class 4 (Proliferative DR): recall = 0.8131, F1 = 0.8145. Class 3 (severe DR): recall = 0.7177, F1 = 0.7547.

5.4.4 MaxViT (QWK = 0.1111)

In particular, MaxViT performed badly in this evaluation setting. It achieved only 25.45% accuracy and the macro-average F1-score was 0.2245, suggesting that the model failed to learn discriminative representations for DR grading. The confusion matrix shows large misclassification among all classes, and upscattered desired values for unrelated grades

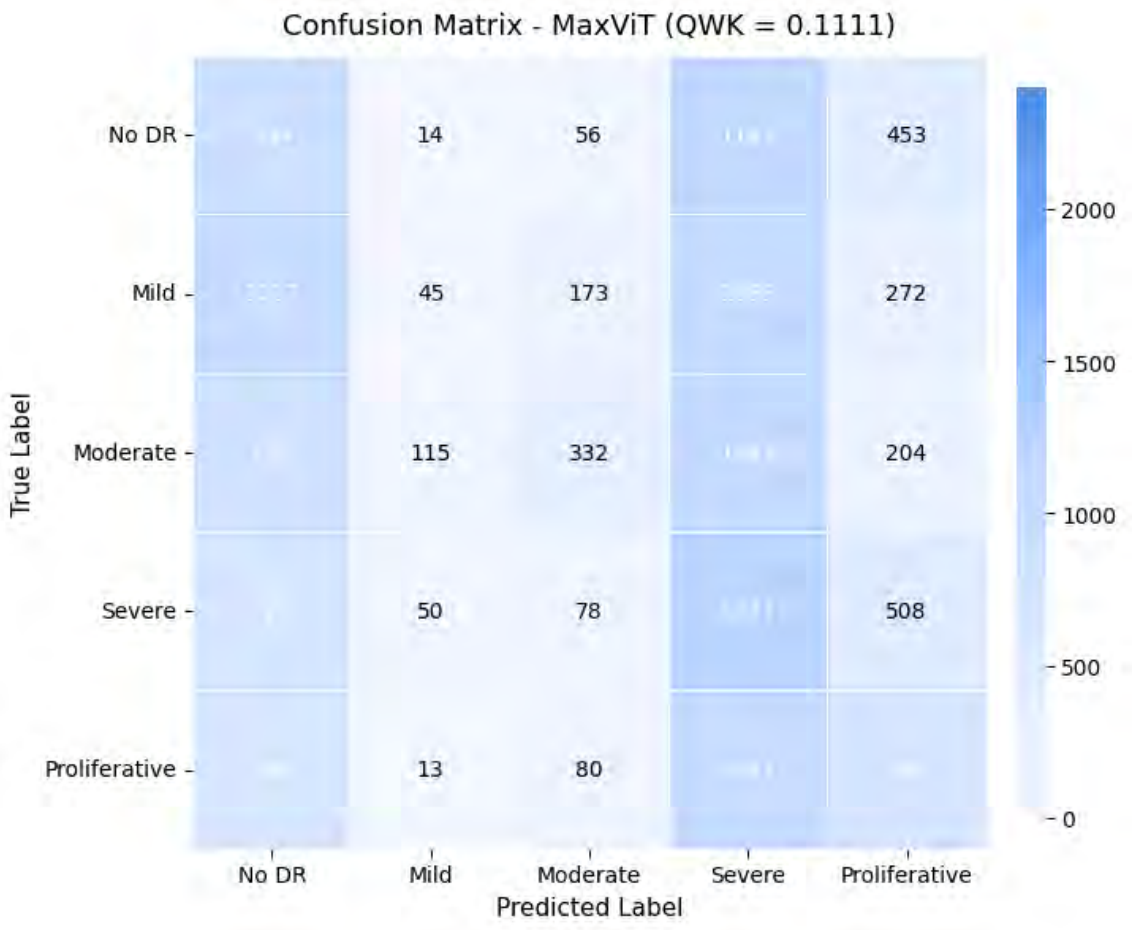

Figure 5.5: Confusion Matrix: MaxViT

The poor recall for Class 1 (0.0173) reflects that the model barely detects moderate DR, and relatively low intermediate degree learning is not achieved as well from its Class 2 recall=0.1277. The QWK 0.1111 suggests that MaxViT does not maintain ordinal structure under the current training setting.

This observation suggests that MaxViT may need different hyperparameter settings, longer training, or alternative preprocessing to challenge for DR grading.

5.4.5 CoAtNet (QWK = 0.902)

CoAtNet was the best overall performer (QWK = 0.902), demonstrating the greatest ordinal agreement among all models examined. This indicates that CoAtNet is

most efficient for producing severity-consistent predictions, while limiting the clinical harm of misclassification.

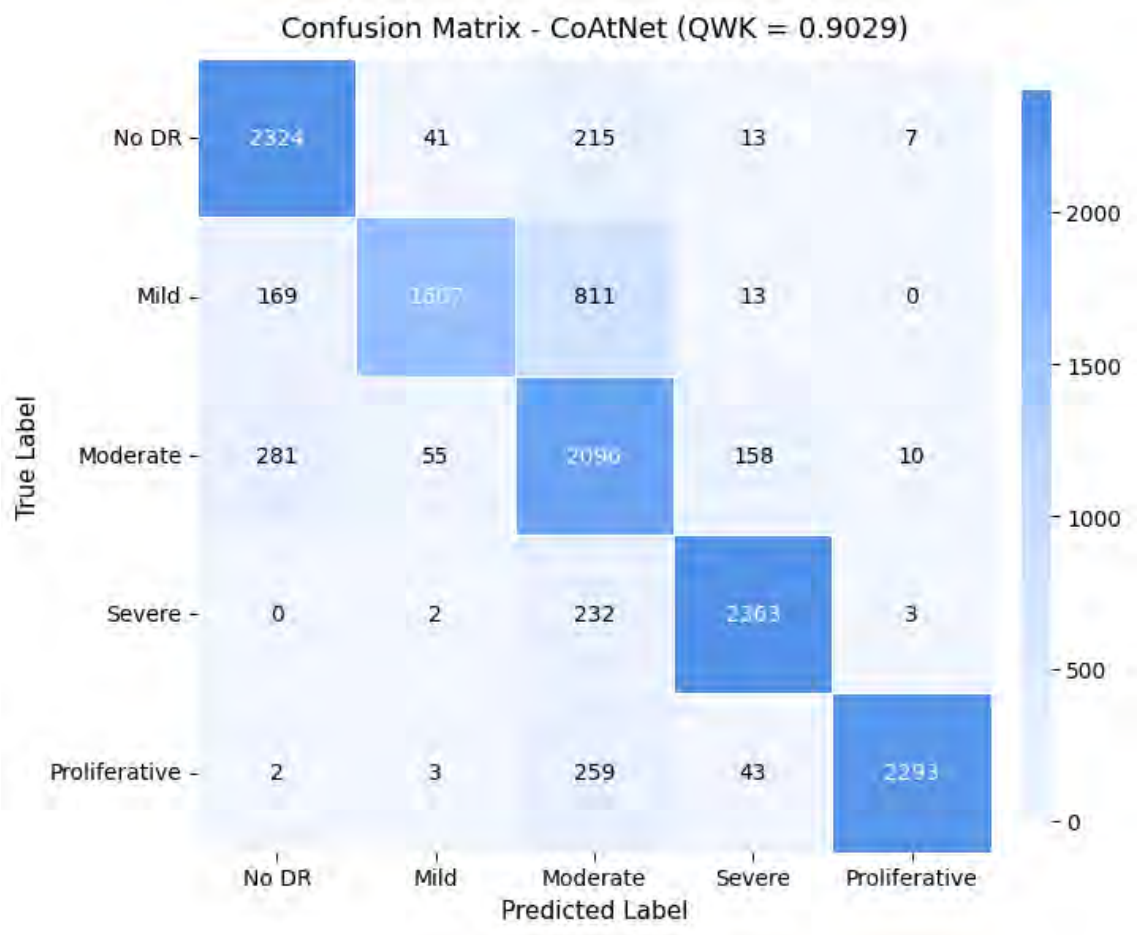

Figure 5.6: Confusion Matrix: CoAtNet

The improved performance can be attributed to the design of CoAtNet, which synthesizes convolutional inductive bias for characterizing the lesion-level representation and attention mechanisms for modeling global context. This is what makes CoAtNet suitable for the task of DR grading wherein severity classification depends not only on local lesions, but also their spatial scattering.

5.5 Comparison of Misclassification Patterns

For all models, the most common errors are found between consecutive levelsof severity, in particular

Mild to Moderate (Class 1 → Class 2)

mild vs severe (Class 2 → Class 3).

This is expected as DR grading boundaries are clinically indistinct, especially when lesion numbers and features approach threshold condition. Significantly, the top-

performing models (CoAtNet and Hybrid) result in higher separation for higher-grade DR (Class 3 and 4), which is crucial for screening safety.

5.6 Explainability and Quality-Tier Robustness Assessments

Further than that, this thesis applys XAI approaches to the models interpretability to different situations of image quality at end, despite predicting accuracy performances. Explanations were generated for all quality levels (Tier A to Tier D) and both explanation and performance reliability was analyzed.

Two important findings were observed:

The accuracy for Tier A is 0.4727 lower than Tier D.

The average attribution entropy for users decreases 0.3533 from Tier A to Tier D.

The reduction in accuracy illustrates that degradation of model trustworthiness is observed when low quality fundus photo. The decreasing attribution entropy provides empirical evidence that explanation maps are concentrated in low-quality images or that the model may over-rely on limited or noisy features. This supports the thesis's argument that variability of quality has implications on both predictive confidence and interpretability coherence.

5.7 Discussion

Experimental results demonstrate under merged data settings that the hybrid convolution-attention architectures can potentially outperform five-tier DR grading. CoAtNet frames the highest ordinal agreement (QWK 0.902), Hybrid does second best (QWK 0.8944). Units of these architectures generally showed good recall for vision-threatening and proliferative DR, suggesting that they were more appropriate for a diagnostic setting in which the absence of high-grade retinopathy is clinically unacceptable.

ConvNeXt also performed well, suggesting that existing convolutional designs are still very powerful for lesion-level grading. But CConvNeXt Vs CoAtNet/Hybrid difference does hint is powerful on global context modeling would help promote severity-consistent classification.

We observed that Vision Mamba had a middle level of capacity but poor at early-stage differentiation, and also that the use of MaxViT failed in the present pipeline to reveal that not all very recent architectures are good generalizers when it comes to fundus grading without fine-tuning.

Chapter 6

Explainable AI

6.1 Explainability-Driven Report Generation Module

Automated diabetic retinopathy (DR) classifiers often achieve high grading accuracy but their clinical utility in the realworld goes beyond numerical predictions. In a testing environment, clinicians want actionable evidence that engenders confidence in the evidence, allows for error checking and speeds documentation. The majority of deep learning models can only give approximated label, which impedes clinical application on account of black-box characteristics in most modern models. This limitation is addressed by this Thesis introducing an explainability-driven reporting component which transforms model output into a structured diagnostic summary in the style of medical screening.

We present the design and development of a prototype automatic report generation module that incorporates three important components: (i) predicted DR grade, (ii) model uncertainty and (iii) lesion-level visual evidence with Evidential Gradients using Explainable AI(XAI)-heatmaps. The reporting module is designed to aid screening workflows, by generating standardized documents and presenting clear evidence for every prediction. Importantly, this module is not intended to replace clinician decision-making but to provide a decision support tool that increases transparency and reduces manual reporting workload.

6.2 Design Rational and Clinical Motivation.

In routine public screening programs, doctors usually grade the severity of DR based on the patterns and distribution of lesions first, then record a clinical impression and suggest follow-ups. Manually reporting is time-consuming and requires cognitive load – work that you don't necessarily want your staff to be doing, especially when it's busy. Then, it seems that the following two applied advantages can be achieved by organizing reports:

Standardization serves to maintain uniformity with documentation style and minimize subjective variability in reporting.

Moreover, the integration of XAI evidence mitigatesone of healthcare AI adoption barrier: Health practitioners are more likely to have confidence in model predictions if they can observe that the model concentrates on relevant anatomical regions such as hemorrhages, exudates, and microaneurysms rather than irrelevant artifacts.

6.3 Inputs and Outputs of the Reporting Module

6.3.1 Inputs

The report generation module takes the following inputs from the trained grading pipeline:

- Predicted DR grade $\hat{y} \in \{0, 1, 2, 3, 4\}$.
- Confidence score $p(\hat{y})$, obtained from the softmax output.
- XAI heatmap(s) generated using Grad-CAM and/or attention visualization.
- Quality tier indicator (Tier A–D), derived from image quality assessment.
- *Optional:* top-k class probabilities, to capture uncertainty between neighboring grades.

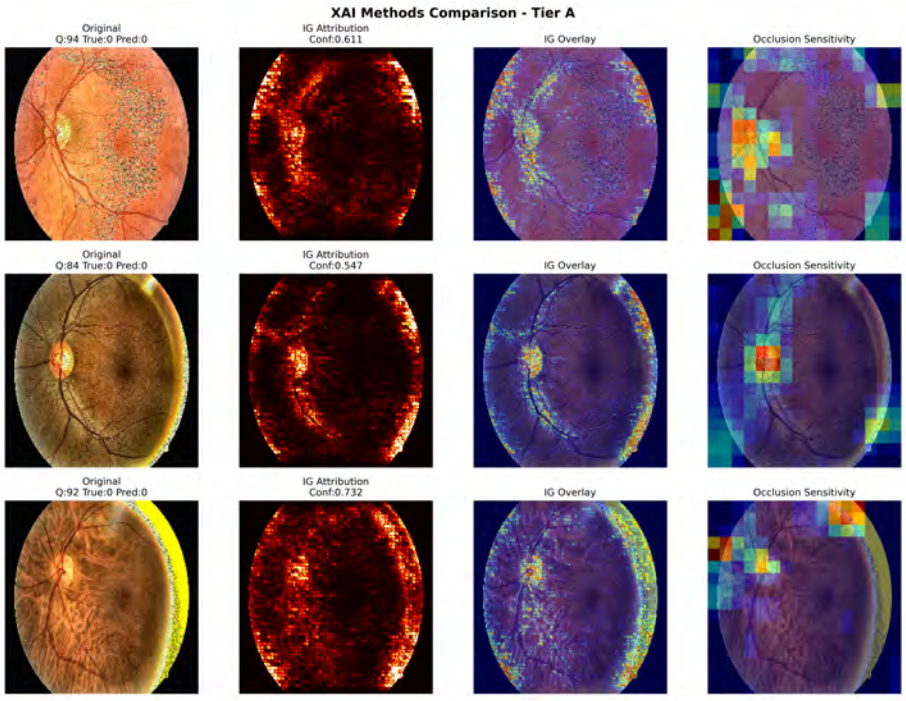

Figure 6.1: XAI models comparison on Tier A

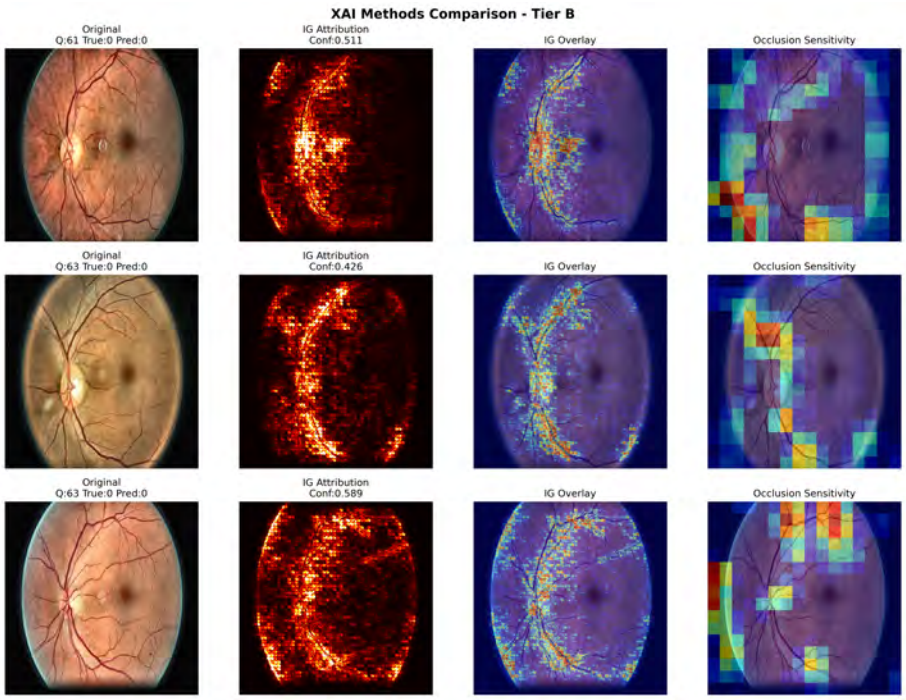

Figure 6.2: XAI models comparison on Tier B

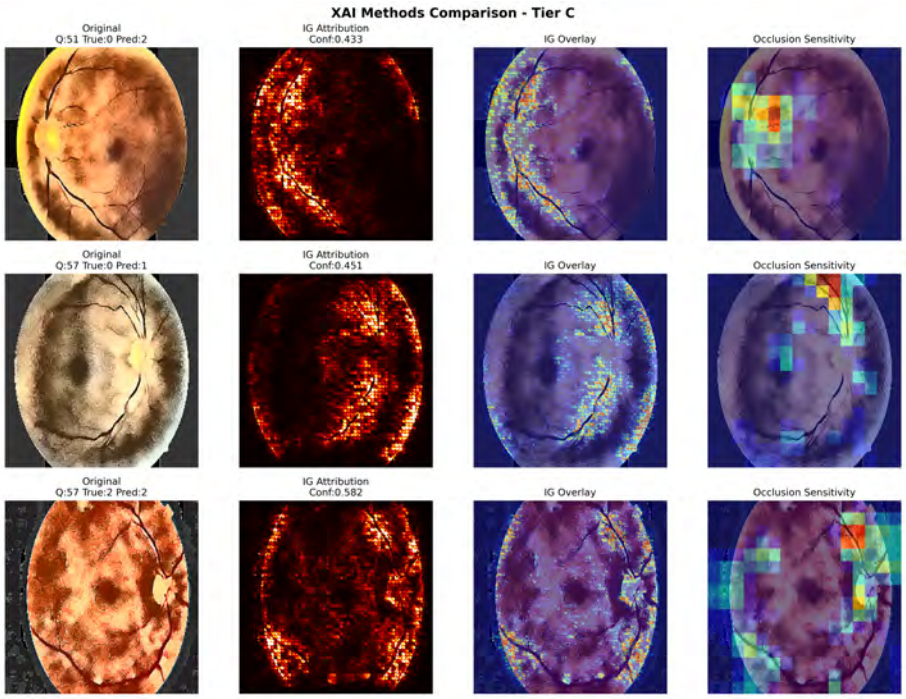

Figure 6.3: XAI models comparison on Tier C

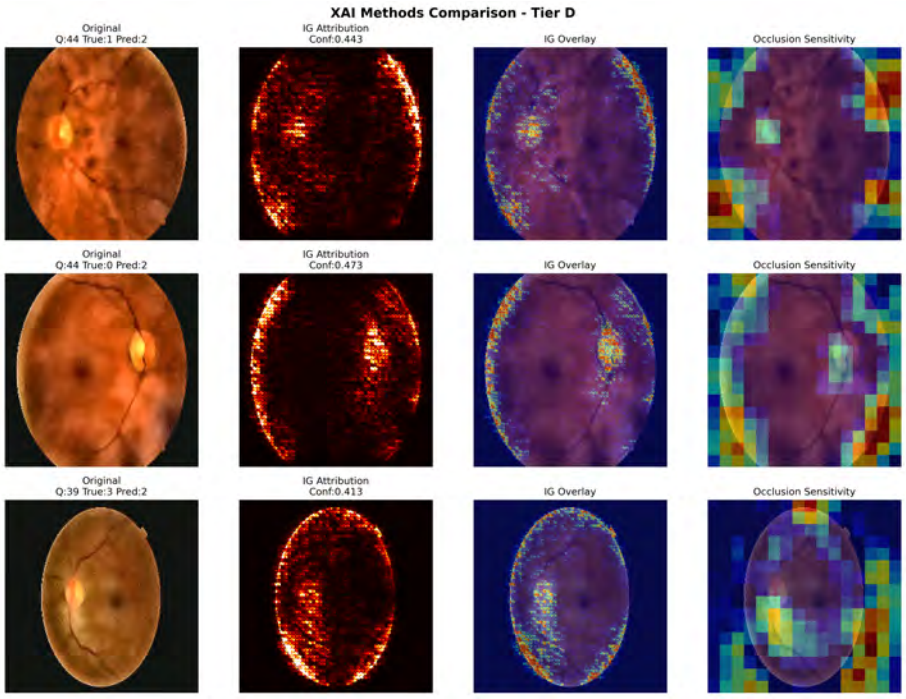

Figure 6.4: XAI models comparison on Tier D

6.3.2 Outputs

The module generates a structured textual summary containing:

- Patient screening outcome (predicted DR severity grade)
- Confidence estimate
- XAI-driven evidence statement (regions highlighted)
- Lesion-informed interpretation (rule-based inference)
- Follow-up recommendation (screening-oriented)

The output is formatted to resemble a short clinical screening report.

6.4 Report Generation Strategy

The batch generator is a rule-based natural language template system for translating model outputs into texts. We chose this approach because it offers deterministic, controllable report structure and avoids the supernatural dangers of fully generative language models. It also assures reproducibility and standardization between patients.

There are four steps in the reporting logic

Step 1: Severity Interpretation

- The expected grade is equivalent to a clinical severity label: 0 is No DR, 1 is Mild NPDR, 2 is Moderate NPDR, Severe NPDR, Proliferative DR.

Step 2: Confidence Categorization

- Confidence score values are also mapped to qualitative terms.
- $p \geq 0.85 \rightarrow$ High confidence.
- $0.65 \leq p < 0.85 \rightarrow$ Moderate confidence level.
- $p < 0.65$ implies low confidence.
- This enhances the physician's understanding and identifies questionable cases for manual review.

Step 3: XAI Evidence Extraction

- Those retinal locations are analyzed in the XAI heatmaps to understand which regions of the retina informed the decision. Evidence statements include:
- highlighted the center macular area.
- optic disc/peripapillary focus.
- Diffuse peripheral activation.
- vessel arcade-related focus.
- Avoid the classic problem of fancy filtering looking for disorders: we are not to assure that you'll be given direct lesion segmentation, but it works with activation focus and site of clinical interest.

Step 4: Synthesis of recommendations.

- A grade-based screening recommendation is included.
- Grade 0–1: regular follow-up screening.
- Grade 2: close / watching and likely referral according to the rules.
- Grades 3-4: recommended immediate referral because of the risk to vision.

6.5 Role of XAI in Report Generation

There are two objectives fulfilling the Explainable AI in this module:

- Justification of Evidence: Heatmaps provide visual evidence for expected grades.
- Context: The narrative support for the evidence regions is provided within the report as short summaries.

For better reliability the report module has quality-tier awareness. As XAI study demonstrated that robustness of explanation is reduced as image quality decreases, the module flags low-quality cases:

The cognate retrieval is lower with a decrement of 0.4727 from Tier A to Tier D. There's a reduction of 0.3533 in the attribution entropy between Tier A and D. For images of Tier C/D, the study cautions against potential deterioration of prediction and explanation reliabilities due to the image quality. "Manual verification is recommended." This ensures that interpretability is appropriately leveraged and aligns with the focus of the thesis on quality-aware screening.

6.6 Illustrated Example Report Template

Following structure is constructed by the above module:

- DR Screening Summary Report.
- Predicted grade: Class 3, Severe NPDR.
- Confidence is high (0.91).
- Visual Evidence (XAI): The activation was concentrated around vascular arcades and the mid-peripheral retina; highlighted areas are haemorrhagic/exudative patterns.
- Predictive features of full lesions are predicted by the model, which is typical for advanced NPDR.
- Recommendations: Urgent referral for ophthalmology review is indicated.
- This pattern also illustrates how the module enlists both classification and interpretability to provide a clinically useful summary.

```
================================================================================
DIABETIC RETINOPATHY — AI-ASSISTED SCREENING REPORT
================================================================================

Tier: Tier A (Excellent)

1) Predicted Diagnosis
   • Predicted Grade: Class 3 (Severe NPDR)
   • Confidence: High (89%)

2) Visual Evidence (XAI-highlighted regions)
   • Vascular arcades
   • Mid-peripheral retina

3) Lesion Patterns
   • Hemorrhages
   • Hard exudates

4) Clinical Interpretation
   The ConvNeXt model detected extensive retinal vascular changes consistent
   with Severe Non-Proliferative Diabetic Retinopathy. Multiple hemorrhages
   and exudates indicate significant microvascular damage.

5) Recommendation
   URGENT REFERRAL (2-4 weeks)
   Action: Refer to ophthalmology for:
    • Comprehensive dilated examination
    • Consider fluorescein angiography
```

Figure 6.5: Report Template Of Tier A

```
================================================================================
DIABETIC RETINOPATHY — AI-ASSISTED SCREENING REPORT
================================================================================

Tier: Tier B (Good)

1) Predicted Diagnosis
   • Predicted Grade: Class 2 (Moderate NPDR)
   • Confidence: Moderate (76%)

2) Visual Evidence (XAI-highlighted regions)
   • Macular region
   • Superior arcade

3) Lesion Patterns
   • Microaneurysms
   • Mild hemorrhages

4) Clinical Interpretation
   The ConvNeXt model detected retinal changes consistent with Moderate Non-
   Proliferative DR. Close monitoring and follow-up recommended to prevent
   progression.

5) Recommendation
   Close monitoring, consider referral
   Action: Schedule follow-up screening in 6-12 months
```

Figure 6.6: Report Template Of Tier B

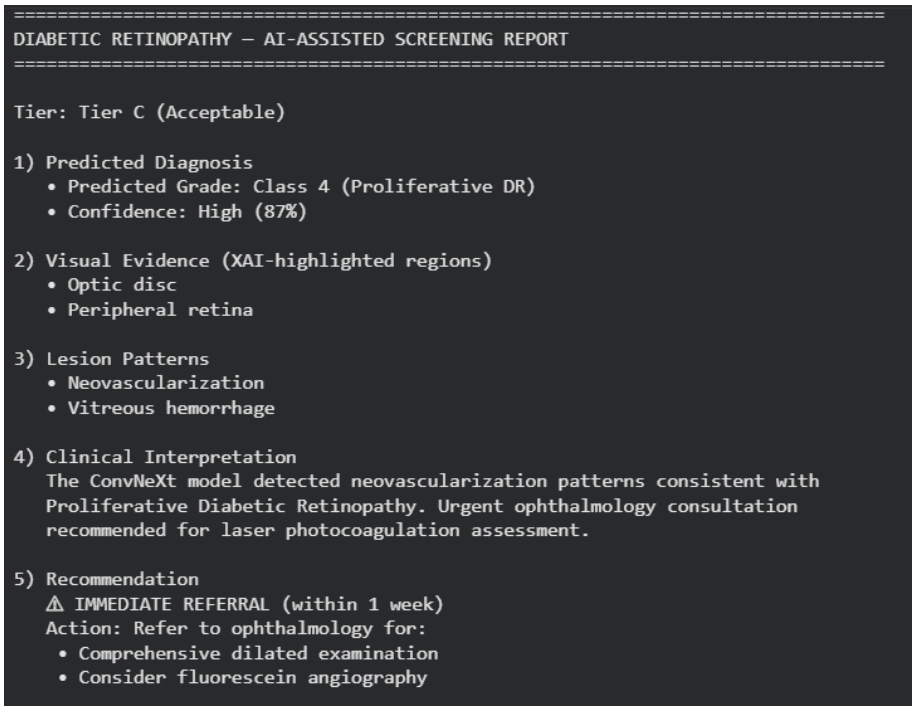

Figure 6.7: Report Template Of Tier C

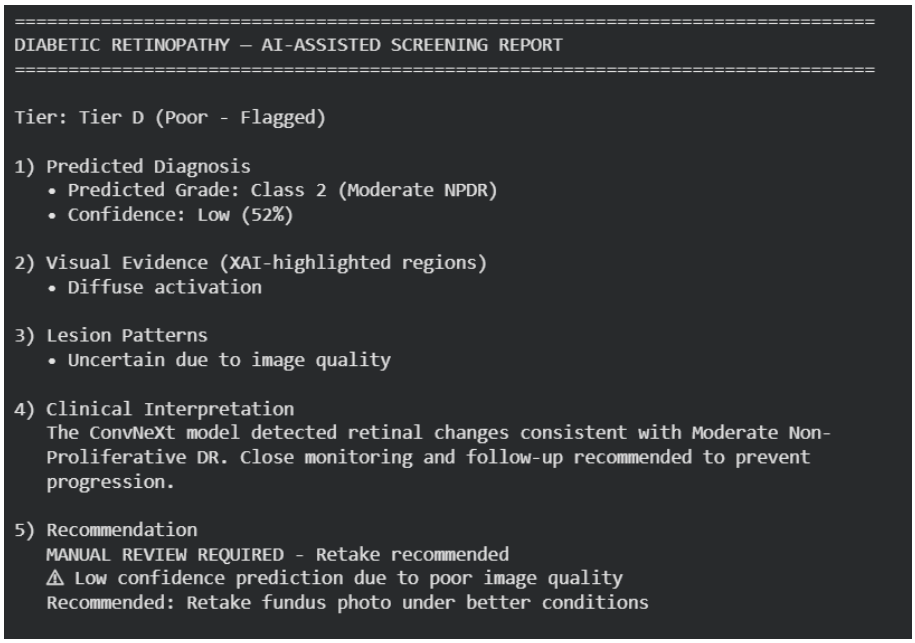

Figure 6.8: Report Template Of Tier D

6.7 Shortcomings of the Reporting Module

- Despite the module enhancing readability/documentation, however, there are a few downsides:
- Reports have been generated from rule-based templates without harnessing true free-text clinical reasoning.
- XAI heatmaps yield attention-based evidence, yet they may not offer causal lesion localization.
- No clinical user research was performed to evaluate usability, readability or integrating that into real-world workflow.
- This module also takes only fundus photos and introduces no patient information (HbA1c, duration of diabetes) that may influence any clinical decision.
- Despite these constraints, the module demonstrates that structured summary reports can be generated from the output of a deep learning algorithm and this is an important milestone for clinically useful DR screening AI.

6.8 Excluded Explainable AI Methods: Rationale and Limitation

While multiple explainable AI (XAI) techniques exist for interpreting deep learning predictions in medical imaging, not all methods are equally suited to retinal fundus analysis and clinical deployment. This section discusses two widely adopted XAI approaches—LIME (Local Interpretable Model-Agnostic Explanations) and SHAP (SHapley Additive exPlanations)—that were considered but ultimately excluded from this work due to practical and clinical limitations.

6.8.1 LIME (Local Interpretable Model-Agnostic Explanations)

LIME is a model-agnostic explainability technique that interprets individual predictions by locally approximating the model's decision boundary using simplified surrogate models. The method operates by perturbing segments of the input image (typically superpixels) and observing the resulting changes in the model's output probability distribution.

Rationale for Exclusion Despite its widespread adoption in explainable AI research, LIME was not implemented in this thesis due to the following limitations in the context of diabetic retinopathy (DR) grading:

- **Coarse Spatial Resolution and Lesion Misalignment:** LIME relies on superpixel segmentation which often fails to align with fine-grained retinal lesions such as microaneurysms, hemorrhages, and exudates. This results in coarse explanations that highlight clinically irrelevant regions.
- **Instability and Non-Reproducibility:** LIME explanations are stochastic due to random perturbation sampling. Repeated runs on the same image yield different explanations, which is problematic for medical applications requiring reproducibility.
- **Computational Overhead:** LIME requires 500-1000 model evaluations per explanation, resulting in 10-30 seconds inference time per image, which is impractical for large-scale screening.
- **Limited Anatomical Interpretability:** LIME produces abstract superpixel importance scores rather than anatomical region labels, reducing clinical utility.

LIME was excluded in favor of Grad-CAM, which provides stable, class-specific, spatially coherent explanations aligned with clinical interpretation.

6.8.2 SHAP (SHapley Additive exPlanations)

SHAP is a model-agnostic explainability framework grounded in cooperative game theory. It assigns importance values to input features by estimating their contribution through Shapley value approximations.

Rationale for Exclusion Although SHAP offers strong theoretical guarantees, it was not utilized due to practical limitations:

- **Computational Complexity:** Even approximate SHAP variants require hundreds of model evaluations, resulting in 20-60 seconds per image—prohibitive for screening workflows.
- **Diffuse Explanations:** SHAP produces pixel-level importance maps that lack spatial coherence, making it difficult to identify specific lesion evidence.

- **Limited Compatibility:** Extending SHAP to hybrid attention-based models introduces complexity, as attention mechanisms don't decompose into additive contributions.
- **Limited Clinical Validation:** Few studies demonstrate SHAP's clinical utility for fundus image interpretation compared to gradient-based methods.

SHAP was excluded in favor of Grad-CAM, Integrated Gradients, and Occlusion Sensitivity, which provide efficient, spatially localized, anatomically interpretable explanations.

Chapter 7

Summary of Findings, Contributions and Future Work

7.1 Summary of findings

This thesis developed and evaluated a quality-sensitive, explainable deep learning model for five-class diabetic retinopathy grading using open source fundus data(corresponding to the APTOS 2019 and DR Dataset). The approach was designed to tackle three obstacles toward clinical adoption of DR AI: class imbalance, heterogeneous image quality and lack of interpretability.

A lesson learning pipeline was used on Five state of the art architectures. Results from the experiments demonstrate that hybrid convolution-attention architectures are superior to classical ordinal grading. CoAtNet achieves the best overall agreement with ground-truth (QWK = 0.902), closely followed by Hybrid CoAtNet-ConvNeXt architecture(QWK = 0.8944). ConvNeXt also did well (QWK = 0.8607), while Vision Mamba showed rather moderate agreement (Qwk = 0.6983). MaxViT performed badly in the current experiment (QWK = 0.1111), suggesting more or less a lack of suitability without extra tuning.

In addition to the predictive performance, lesion-level visual explanations were obtained by explainable AI algorithms employed here in the thesis. Our quality-tier robustness analysis identified that model reliability and interpretability degraded with worse picture quality, where accuracy decreased by 0.4727 and attribution entropy declined by 0.3533 from A to D, respectively; such observations underscore the importance of quality-aware screening pipelines and careful interpretation of AI predictions under suboptimal imaging conditions.

Ultimately a proof-of-concept report generation module was developed to translate model predictions and XAI evidence into structured screening-style diagnostic summaries, increasing transparency and relevance to the workflow.

7.2 Contributions in the Field

This thesis contributes to DR screening research in several aspects:

- Merged, curated dataset formulation
- By integrating APTOS 2019 and DR Dataset together, the unified DR dataset is greatly enlarged in scale and variety at the same time makes it systematically possible to investigate into imbalance and quality variation.
- Unified benchmarking for modern architectures
- Five state-of-the-art vision architectures were trained with the same pipeline settings and compared for five-class DR grading.
- High-performance hybrid model design
- A gated-fusion Hybrid ConvNeXt-CoAtNet network was adopted to incorporate lesion-level feature extraction and global context reasoning, attaining excellent ordinal agreement.
- Interpretability and quality-tier assessment
- XAI representations were tested at different image qualities and quantitatively examined to show types of degradation that informs the responsible use.
- Explainability-driven reporting prototype
- A structured reporting system was developed to translate model outputs into clinician-friendly reports, promoting transparency and utility of the screening process.

7.3 Advice for Future Work

While the proposed system is high-performing and interpretable, many research avenues are open to further improve:

- External validation of clinical datasets.
- In the future, we need to evaluate the trained models using hospital-grade datasets from multiple populations and acquisition devices to verify our model generalization.
- Assessment with clinician in the loop.
- User studies with ophthalmologists should assess the utility, interpretability, and credibility of XAI heatmaps and automated reports.
- Advanced quality modeling.
- Image quality estimation could be directly incorporated into training as an auxiliary task, helping models cope with low-quality inputs in a flexible manner.
- Lesion-specific segmentation supervision
- Integration of lesion-level annotations may help interpretability and enable true lesion localization over heatmap-based attention evidence.

- Deployment-oriented optimization
- Light-weight formulations of the best performing architectures could even be developed for mobile or edge deployment in low-resource screening setups.

Chapter 8

Conclusion

In summary, in this thesis, to answer an urgent demand in scalable DR screening, we developed a quality-aware interpretable and high-performing deep learning system for five-class-severity grading DR from retinal fundus images. The work established a practical experimental base for the benchmark of real-world screening performance, by integrating public competition datasets (APTOS 2019 and DR Dataset) and investigating class imbalance and image-quality difference in sophisticated manners. A common unified training and testing pipeline was employed to benchmark 5 recent architectures, and among them CoAtNet yielded the best ordinal agreement (QWK = 0.902) closely followed by Hybrid CoAtNet-ConvNeXt model (QWK= 0.8944). Notably, interpretability was considered to be a system requirement rather than a stretch goal; XAI-based lesion-level visualizations were embedded into the system to highlight clinically relevant retinal image areas affecting predictions and thus enhance transparency and trust by clinical users. Reliability of predictions and behavior of explanations also could be affected by image quality, as shown in robustness analysis where the averaged reliability dropped 0.4727 while the averaged entropy was lowered by 0.3533 from tier A to D (high- and low-quality images). Last, the proposed architecture was complemented by a prototype automatic report-generation module that outputs predicted grades and visual evidence as organised diagnostic summaries in order to enhance workflow suitability and document uniformity. In conclusion, the thesis introduces a complete pipeline that balances accuracy, robustness and transparency of such systems for practical AI-assisted DR screening in clinical practice, especially in high through-put settings with limited resources.

Bibliography

[1] M. D. Abràmoff, J. C. Folk, D. P. Han, *et al.*, "Automated analysis of retinal images for detection of referable diabetic retinopathy," *JAMA ophthalmology*, vol. 131, no. 3, pp. 351–357, 2013.

[2] M. D. Abràmoff, Y. Lou, A. Erginay, *et al.*, "Improved automated detection of diabetic retinopathy on a publicly available dataset through integration of deep learning," *Investigative ophthalmology & visual science*, vol. 57, no. 13, pp. 5200–5206, 2016.

[3] F. Nawaz, M. Ramzan, K. Mehmood, H. U. Khan, S. H. Khan, and M. R. Bhutta, "Early detection of diabetic retinopathy using machine intelligence through deep transfer and representational learning.," *Computers, Materials & Continua*, vol. 66, no. 3, 2021.

[4] M. M. Butt, D. A. Iskandar, S. E. Abdelhamid, G. Latif, and R. Alghazo, "Diabetic retinopathy detection from fundus images of the eye using hybrid deep learning features," *Diagnostics*, vol. 12, no. 7, p. 1607, 2022.

[5] M. S. Ali and M. Islam, "A hyper-tuned vision transformer model with explainable ai for eye disease detection and classification from medical images," *BS thesis, Faculty of Engineering and Technology Islamic University*, 2023.

[6] İ. Kayadibi and G. E. Güraksın, "An explainable fully dense fusion neural network with deep support vector machine for retinal disease determination," *International Journal of Computational Intelligence Systems*, vol. 16, no. 1, p. 28, 2023.

[7] B. Lalithadevi, S. Krishnaveni, and J. S. C. Gnanadurai, "A feasibility study of diabetic retinopathy detection in type ii diabetic patients based on explainable artificial intelligence," *Journal of Medical Systems*, vol. 47, no. 1, p. 85, 2023.

[8] V. S. Pete, "Xai-driven cnn for diabetic retinopathy detection," 2023.

[9] Y. Yang, H. Wang, C. Ji, and Y. Niu, "Artificial intelligence-driven diagnostic systems for early detection of diabetic retinopathy: Integrating retinal imaging and clinical data," *SHIFAA*, vol. 2023, pp. 83–90, 2023.

[10] K. A. Alavee, M. Hasan, A. H. Zillanee, *et al.*, "Enhancing early detection of diabetic retinopathy through the integration of deep learning models and explainable artificial intelligence," *IEEE Access*, vol. 12, pp. 73 950–73 969, 2024.

[11] D. Bhulakshmi and D. S. Rajput, "A systematic review on diabetic retinopathy detection and classification based on deep learning techniques using fundus images," *PeerJ Computer Science*, vol. 10, e1947, 2024.

[12] P. Dixit, S. Sharma, and P. Vaishnav, "Deep learning and ensemble techniques with xai in diabetic retinopathy: A performance-driven review," in *2024 IEEE 11th Uttar Pradesh Section International Conference on Electrical, Electronics and Computer Engineering (UPCON)*, IEEE, 2024, pp. 1–8.

[13] B. Lalithadevi and S. Krishnaveni, "Diabetic retinopathy detection and severity classification using optimized deep learning with explainable ai technique," *Multimedia Tools and Applications*, pp. 1–65, 2024.

[14] K. Mridha, M. Wang, and L. Zhang, "Ai-driven diagnostics in ophthalmology: Tailored deep learning models for diabetic retinopathy with xai insights," in *Proceedings of the 16th International Conference on*, vol. 101, 2024, pp. 73–82.

[15] U. Posham and S. Bhattacharya, "Diabetic retinopathy detection using deep learning framework and explainable artificial intelligence technique," in *2024 14th International Conference on Cloud Computing, Data Science & Engineering (Confluence)*, IEEE, 2024, pp. 411–415.

[16] G. Rajarajeshwari and G. C. Selvi, "Application of artificial intelligence for classification, segmentation, early detection, early diagnosis, and grading of diabetic retinopathy from fundus retinal images: A comprehensive review," *IEEE Access*, 2024.

[17] T. Shahzad, M. Saleem, M. S. Farooq, S. Abbas, M. A. Khan, and K. Ouahada, "Developing a transparent diagnosis model for diabetic retinopathy using explainable ai," *IEEE Access*, 2024.

[18] H. K. Vasireddi, K. S. Devi, and G. R. Reddy, "Dr-xai: Explainable deep learning model for accurate diabetic retinopathy severity assessment," *Arabian Journal for Science and Engineering*, vol. 49, no. 9, pp. 12 899–12 917, 2024.

[19] M. Kannan, S. Nikitha, S. Akshay, *et al.*, "Explainable diabetic retinopathy detection using enlightengan images with deep learning technique," *Procedia Computer Science*, vol. 258, pp. 1586–1597, 2025.

[20] N. Sureja, V. Parikh, A. Rathod, P. Patel, H. Patel, and H. Sureja, "Explainable artificial intelligence based deep learning for retinal disease detection," *Journal of Electronics, Electromedical Engineering, and Medical Informatics*, vol. 7, no. 2, pp. 471–483, 2025.