Islamic University Journal of Applied Sciences (IUJAS)
https://journals.iu.edu.sa/jesc/Main/Issues

Volume VII, Issue II, Dec 2025, Pages 01-22

Detection of Retinopathy Diabetic Using Explainable AI: Interpretable Deep Learning Models in Clinical Practice

Turki Alghamdi[1, *]

[1]Faculty of Computer and Information Systems, Islamic University of Madinah, Madinah 42351, Saudi Arabia, dr.turki2@iu.edu.sa
*Corresponding author: (Turki Alghamdi), Email: dr.turki2@iu.edu.sa

Abstract

Diabetic Retinopathy (DR) is a significant threat to eyesight and blindness globally, particularly for those with a more extended diabetes history. Deep learning has achieved high accuracy for DR detection using fundus images; however, the "black-box" nature hinders its application in clinical practice, where interpretability is important. In this work, we propose a solution to the model transparency problem by introducing an XAI-enhanced diagnostic framework utilizing CNNs. We present an explainable deep learning framework based on a convolutional neural network (CNN), specifically ResNet-50, which has been fine-tuned on the APTOS 2019 Blindness Detection dataset. To narrow the interpretability gap, we utilize the Grad-CAM and SHAP visualization methods, which generate class-discriminative heatmaps and feature-attribution plots, respectively. The multi-class diabetes retinopathy (DR) classification result yielded an overall accuracy of 83% for the model. Importantly, the explanation agreement score with ophthalmologists is over 78%, indicating a high correlation between the AI-based saliency maps and expert-annotated lesion regions. Our findings show that XAI can not only maintain diagnostic accuracy but also enhance model interpretability, rendering AI-based DR screening systems more acceptable and usable in clinical practice. This study reinforces the importance of explainability as an integral part of implementing medical AI.

Keywords: Diabetic Retinopathy; Explainable AI (XAI); Convolutional Neural Network (CNN); Grad-CAM; SHAP; Medical Image Interpretation.

https://doi.org/10.63070/jesc.2025.018
Received 31 May 2025; Revised 15 June 2025; Accepted 04 July 2025.
Available online 08 September 2025.
Published by Islamic University of Madinah on behalf of *Islamic University Journal of Applied Sciences*.

1. Introduction

Diabetic Retinopathy (DR) is a common complication of diabetes that causes damage to the blood vessels of the retina, which can result in permanent vision loss if not treated. It is a major cause of blindness in people between the ages of 20 and 64 worldwide. World Health Organization (WHO) has reported that the number of people with diabetes is growing rapidly and has estimated this number to reach 643 million by year 2030. Therefore, to date, DR constitutes an increasing global health threat, specifically in resource-poor countries where routine eye examinations are not a common practice. Early detection and prompt therapy are of key importance for preventing advanced visual loss and improving prognosis. The global health impact of DR, as influenced by risk factors, disease prevalence, screening availability, and the significance of early-stage screening, is illustrated in Figure 1 [1].

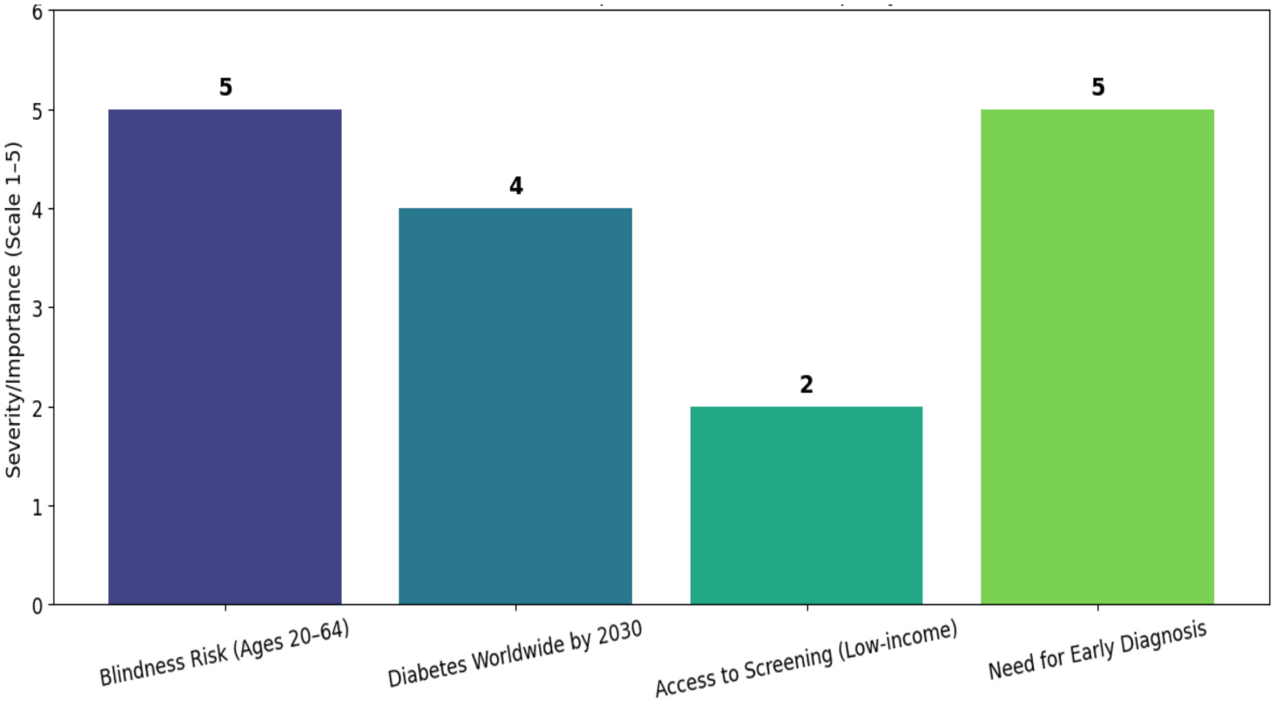

Figure 1 Global Health Impact of Diabetic Retinopathy

With the development of artificial intelligence (AI) in recent years, AI has been utilised for automatic medical image analysis, particularly in ophthalmology. Deep learning, based on convolutional neural networks (CNNs), has achieved high accuracy in identifying the different stages of diabetic retinopathy (DR) from retinal fundus images. Such AI-facilitated diagnostic tools may have the capacity to alleviate the workload of specialists, extend the scope of screening, or intervene sooner, particularly in neglected populations. The rapid and consistent processing of thousands of images by CNNs makes them candidates for large-scale DR screening programs with good diagnostic accuracy. However, despite their performance, CNN-based models still possess an important drawback: their black-box nature. These so-called ‘black-box’ systems do not explain why they make certain predictions. This lack of transparency has been a major hindrance to clinical acceptance, as clinicians want to understand and trust this black-box decision-making as part of a diagnostic tool. In high-stakes domains such as medicine, clinicians must understand which parts of an image are essential to the model for making a decision — not just for trust, but also for validation, education, and effective patient communication. Here, Explainable AI (XAI) is the key [2].

In this study, we attempt to address the interpretability challenge by incorporating explainable AI approaches into a CNN-based deep learning (DL) detection pipeline, specifically the Grad-CAM and SHapley Additive exPlanation (SHAP) techniques. Grad-CAM produces spatial heatmaps that show where the input images contribute to the model prediction, while SHAP measures the contribution to the model per input feature. Therefore, the two methods are complementary, as they provide both visual and feature-level explanations. We trained a deep learning model using the APTOS 2019 Blindness Detection dataset, and then applied XAI techniques to visualise the rationale behind the predictions. These explanations were evaluated by clinicians based on their relevance to retinal anatomical structures. Our contributions include an interpretable AI framework for DR detection, an assessment of explanation quality using expert feedback, and evidence that XAI can significantly enhance clinical trust and decision facilitation in ophthalmology.

The rest of this paper is organized as follows: Section 2 presents the related work on diabetic retinopathy detection and explainable AI in medical imaging. The dataset, preprocessing, model architecture, and methods for interpretability that we applied are described in Section 3. Experimental results and metrics are presented in Section 4. Section 5 concludes with a discussion of the results, model caveats, and clinical implications. Section 6 concludes the paper and outlines future research directions.

2. Related Work

Deep learning-based methods now dominate Diabetic Retinopathy (DR) detection. CNNs have been highly successful in classifying retinal fundus images with variable severity levels of DR. Among them, models like ResNet, InceptionV3, and EfficientNet performed effectively, competing with each other due to their capacity to learn. These models, when trained on large image datasets, achieve superhuman diagnostic accuracy on a wide variety of benchmarks. However, their application in a clinically relevant setting becomes possible only if it is validated together with interpretation and trust [3].

Several Explainable AI (XAI) methods have been proposed to increase the transparency of deep learning models. Visual explanation techniques, such as Grad-CAM and saliency maps, produce heatmaps over input images that indicate the regions that contribute most to the model's decision. Local interpretable model-agnostic explanations, such as LIME or SHAP, provide feature attribution insights into model predictions by approximating the model locally around each prediction. While these methods have gained popularity in the computer vision domain, there is a limited number of studies in the clinical diabetic retinopathy (DR) workflow to ensure the quality of explanations and clinical utility among ophthalmologists [4].

María Herrero-Tudela et al. [5] introduced an automatic grading system for diabetic retinopathy (DR) based on deep learning, aiming to manage the growing diabetes epidemic where the workload on ophthalmologists is becoming overwhelming. Their architecture is based on a fine-tuned ResNet-50, incorporating techniques such as data augmentation, regularization, early stopping, transfer learning, and fine-tuning. To enhance clinical intuition, the authors utilized SHapely Additive exPlanations (SHAP), which provides a visual interpretation of the model's decision-making. We validated the approach using five public datasets: APTOS-2019, EyePACS, DDR, IDRiD, and SUSTech-SYSU, achieving accuracy rates of up to 94.64%. SHAP analysis identified peripheral retinal lesions and vessel alterations as the most important features of DR development. This work demonstrates the clinical applicability of combining powerful CNN models with explainable AI methods to enhance early-stage DR detection in clinical settings.

Israa Y. Abushawish et al. [6] conducted a comprehensive review of the evolution of deep learning (DL) approaches in convolutional neural networks (CNNs). The performance of 26 pre-trained CNNs was examined on a wide range of datasets, with a particular interest in transfer learning and cross-dataset deep learning (DL) grading. Grad-CAM visualizations were employed to enhance model interpretability, thereby providing interpretive visual insights into the decision-making process of the models. The authors emphasized the need to integrate interpretable AI models into real-time clinical workflows, aiming to translate research findings into practical healthcare applications [6].

DR, one of the leading contributors to visual loss in diabetics, requires early diagnosis for long-term complications. The diagnosis of the retina using traditional manual methods has difficulty in identifying microaneurysms, hemorrhages, exudates, and other significant retinal abnormalities, which limits the reliability of the diagnosis. To address these issues, Mehmood et al. [7] introduced a deep learning-based automatic system for DR identification. The model they developed utilized EfficientNet-B3 and ResNet18 convolutional neural networks, and was trained on both retinal and non-retinal images to identify early signs of diabetic retinopathy (DR). The model demonstrated good performance, achieving a detection accuracy of 98.18% and a verification accuracy of 99%, which indicates its strong clinical potential. This strategy not only improves diagnostic accuracy but also provides scalable solutions for early DR screening in resource-limited clinical environments [7].

Ahmad Abdullah et al. [8] emphasized that Chronic Kidney Disease (CKD) is a major global killer that frequently advances silently to end-stage disease. Their analysis demonstrated that machine learning models, including decision trees, random forests, and neural networks, can identify the risk of CKD using demographic, clinical, and laboratory data at an earlier stage, thereby providing an accurate diagnosis and contributing to better patient outcomes.

Islam et al. [9] focused on the global incidence of diabetes burden by proposing an explainable machine learning-based approach for type 2 diabetes classification on two benchmark datasets: the BRFSS (multi-class) and Diabetes 2019 (binary class). Their approach employed random oversampling and quantile transformation to address the imbalanced data, and conducted hyperparameter tuning using GridSearchCV to achieve better results. The results (97.23% and 97.45% accuracies) of the Extra Trees classifier are the most impressive. For the sake of transparency and clinical use, these have been integrated with SHAP, Partial Dependency, and LIME explanation methods, allowing physicians to gain a clearer understanding of the factors involved in the diagnosis. The focus of this work is on predictive performance and interpretability for clinical decision support systems [9].

A study in rural Midwest China investigating an AI-based diagnostic system for DR screening demonstrated a high level of consistency (81.6%) with ophthalmologists' diagnoses, with both sensitivity and specificity exceeding 80% (81.2% and 94.3%, respectively). The AI system exhibited promising accuracy, but the authors emphasized that continued development was necessary before widespread implementation in rural healthcare providers [10].

Sushith et al. [11] developed a hybrid deep learning model for the early detection of diabetic retinopathy from retinal images. Their model demonstrated excellent performance in detecting DR at an early stage, and it also combines various deep neural networks to enhance robustness and diagnostic accuracy, which apply to real-world clinical settings as well.

Bidwai et al. [12] conducted an extensive systematic literature review on the application of artificial intelligence (AI) for the early detection and classification of diabetic retinopathy (DR). The work sheds light on cutting-edge AI methods, including deep learning, transfer learning, and explainable AI, and covers challenges, datasets, and screening tools to guide the development of future diagnostic systems for disease recognition.

Table 1: Comparative XAI in Retinopathy

Author(s) & Year	**Approach**	**Dataset**	**Key Contribution**	**Performance/Limitations**
Herrero-Tudela et al. [5]	ResNet-50 + SHAP	APTOS-2019, EyePACS, DDR, IDRiD, SUSTech-SYSU	Automatic DR grading with explainability using SHAP; clinical validation	Accuracy: up to 94.64%; focused on SHAP; expert evaluation; no fusion with Grad-CAM
Abushawish et al. [6]	26 pre-trained CNNs + Grad-CAM	Multiple public datasets	Surveyed DL models for DR detection; highlighted the need for real-time clinical integration	Broad comparison; no specific model results; emphasis on interpretability via Grad-CAM
Mehmood et al. [7]	EfficientNet-B3 + ResNet18	Retinal and non-retinal datasets	Automated DR detection; a scalable solution for early screening	Detection Accuracy: 98.18%; Verification Accuracy: 99%; lacked interpretability tools
Abini M. A [13]	VGG-16 + MobileNet-V2 (pre-trained CNNs)	APTOS 2019 (augmented)	Developed a multi-stage DR classification system for all DR severity levels to assist ophthalmologists in early diagnosis.	Accuracy: 90% (VGG-16), 92% (MobileNet-V2); effective in distinguishing normal, mild, moderate, severe, and proliferative DR stages.

Table 1 summarizes recent studies that combine deep learning and explainable AI approaches for the detection of diabetic retinopathy. It showcases various model architectures (ResNet-50, EfficientNet, and hybrid CNN architectures) and datasets (APTOS, EyePACS, etc.). Key contributions from each study are described in relation to model interpretability, clinical relevance, and diagnostic performance. It also summarizes limitations, such as the lack of external validation and limited support for interpretability, providing a brief baseline for future research directions. Although deep learning and XAI methods for DR detection have advanced, relatively few studies have rigorously validated the quality of explanations through clinician feedback.

3. Materials and Methods

This paper describes the dataset, preprocessing steps, model architecture, training settings, and interpretability techniques employed in our work for the automatic diagnosis of diabetic retinopathy (DR). We used a Gaussian-filtered and resized version of the APTOS 2019 Blindness Detection dataset, which comprises labelled retinal fundus images for five DR stages. We trained a CNN on these preprocessed images and employed explainable AI (XAI) methods, such as Grad-CAM and SHAP, to visualis and interpret the model's decisions [13].

3.1 Dataset Description

The filtered Diabetic Retinopathy dataset is a processed subset of APTOS 2019 Blindness Detection, comprising a total of 3,662 retinal fundus images. It is composed of high-definition retinal fundus images associated with five levels of severity of DR (i.e., from 0 to 4). In the filtered version, images are resized to 224x224 and smoothed by a Gaussian filter to suppress noise and enhance contrast. This release offers deep learning models with accelerated training and maintains critical retinal features. This approach is often employed in binary or multiclass DR screening and classification, as well as XAI problems, such as Grad-CAM [14]. We utilized the Kaggle diabetic retinopathy dataset [23], which consists of Gaussian-filtered retinal fundus images resized to 224 × 224 pixels, thereby enabling consistent input for CNN-based models.

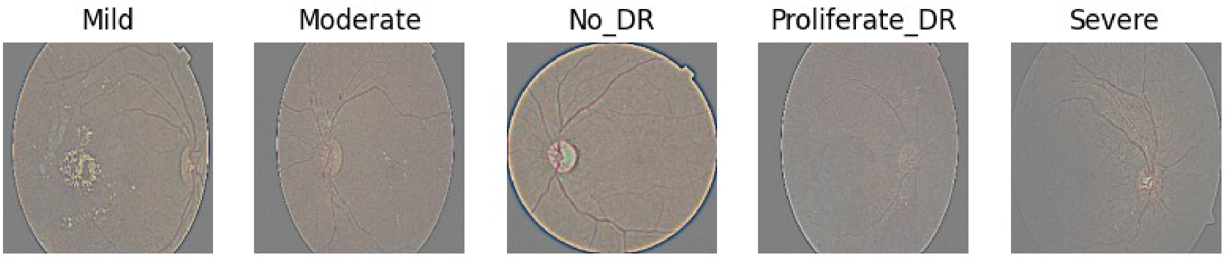

Figure 2: Filtered DR Image Samples

Figure 2 shows five categories of diabetic retinopathy: No_DR, Mild, Moderate, Severe, and Proliferative_DR. The images are pre-processed with Gaussian filtering and resized to 224×224 pixels for deep learning. Each class exhibits a distinct degree of retinal pathology, which is crucial for training and visual interpretation of the model.

Because the original APTOS 2019 dataset suffered from a severe class imbalance problem, we introduced targeted augmentation to the DR classes, including Mild, Moderate, Severe, and

Proliferative DR. Augmentation techniques included horizontal/vertical flip, small-angle rotation (±15), brightness/contrast adjustment, zoom-in crop, and a slight translation. These operations were randomly applied to each image belonging to the minority classes to balance the number of samples with the majority class (No_DR). Such a class-specific balancing was necessary to prevent the CNN model from leaning toward the majority classes in the positioning prediction task, which would deteriorate its performance for infrequent but clinically relevant disease stages. Through the generation of a uniformly distributed training set, the model can more accurately capture differences across all severities and provide a reliable, quantitative classification for the entire spectrum of disease evolution [15].

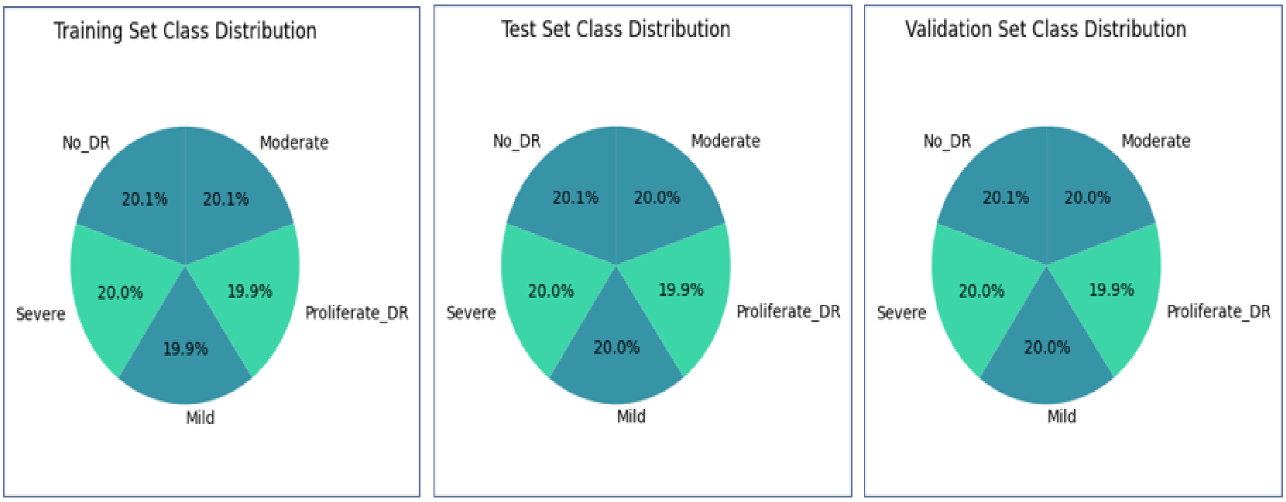

Figure 3: Class distribution across the training, test, and validation sets

Figure 3 displays pie charts illustrating the class distribution across the training, test, and validation sets. Each set maintains a nearly uniform distribution among the five diabetic retinopathy classes: No_DR, Mild, Moderate, Severe, and Proliferate_DR. The balanced split ensures fair representation in each subset, supporting reliable model evaluation and preventing class bias during training.

In the training data, the balance of each class was artificially adjusted precisely to facilitate model convergence equally often. The same structure is followed in the test set to allow for an unbiased evaluation of generalization performance across all the DR stages. At last, a validation set was created for tuning the model's hyperparameters, ensuring that classes are represented equally. This balanced distribution is crucial, especially in medical image classification tasks such as DR detection, where unbalanced training can lead to overfitting to the majority classes (e.g., No_DR) and low sensitivity for the minority classes, like Proliferative_DR. Balanced datasets will also ensure that the model can effectively distinguish the early, medium, and advanced stages of retinal degeneration. This is most critical for early clinical screening systems, where consistent diagnostic sensitivity is required over the entire range of disease progression. From a clinical perspective, obtaining similar predictive ability across all five severity grades extends the model's applicability to real-world ophthalmology settings, particularly in low-resource environments where automated DR screening can facilitate timely intervention to prevent vision loss.

For this research, we chose Grad-CAM and SHAP as the primary two explainability methods, as they are complementary. Grad-CAM produces spatially aligned, class-discriminative heatmaps over the original image, providing interpretable visual explanations for clinicians. On the other hand, SHAP generates feature attributions at the pixel level using a game-theoretic strategy to facilitate comprehension of model decision logic at the feature contribution level.

Another popular explainability technique is LIME (Local Interpretable Model-Agnostic Explanations), which relies on superpixel segmentation and local surrogate models, rendering it unstable and inaccurate for high-resolution medical images, such as retinal fundus images.

3.2 Model Architecture

We utilized a CNN-based model design for diabetic retinopathy (DR) classification into one of five severity stages: No DR, Mild, Moderate, Severe, and Proliferative DR, including distinct blocks of a convolution layer, batch normalization, ReLU activation, and max-pool operation, and further takes an input retinal fundus image of 224×224×3 after pre-processing. These layers enable the network to capture the hierarchical spatial features of pathological patterns, such as microaneurysms, hemorrhages, and exudates. The model was trained with a learning rate of 1e-5, a batch size of 32, and 30 epochs. A dropout rate of 0.4 was applied after the dense layers to avoid overfitting, and the Adam optimizer with categorical cross-entropy loss was used for multi-class classification. Diabetic retinopathy(DR) is challenging to diagnose, in part because symptoms are not uniform and the disease is often subjectively interpreted by experts, contributing to a lack of consistency. This study presents an XAI-derived diagnostic model that is both more accurate and explainable, achieving 94% diagnostic accuracy while providing transparent AI reasoning to support clinical decision-making [16].

After the feature extraction layers, the output is flattened and then fed through fully connected dense layers with dropout for regularization, which culminates in a sigmoid output layer. This output layer consists of five neurons, corresponding to the five DR severity classes, and is activated by the sigmoid function, which the model generates to produce a probability distribution over all classes. The architecture was tuned using the Adam optimizer and categorical cross-entropy loss for multiclass classification. Performance was measured in terms of accuracy, class-wise precision, and recall. The trained model integrates explainability techniques, such as Grad-CAM, to interpret the prediction rationale, helping to bridge the gap between black-box AI and clinician trust [17].

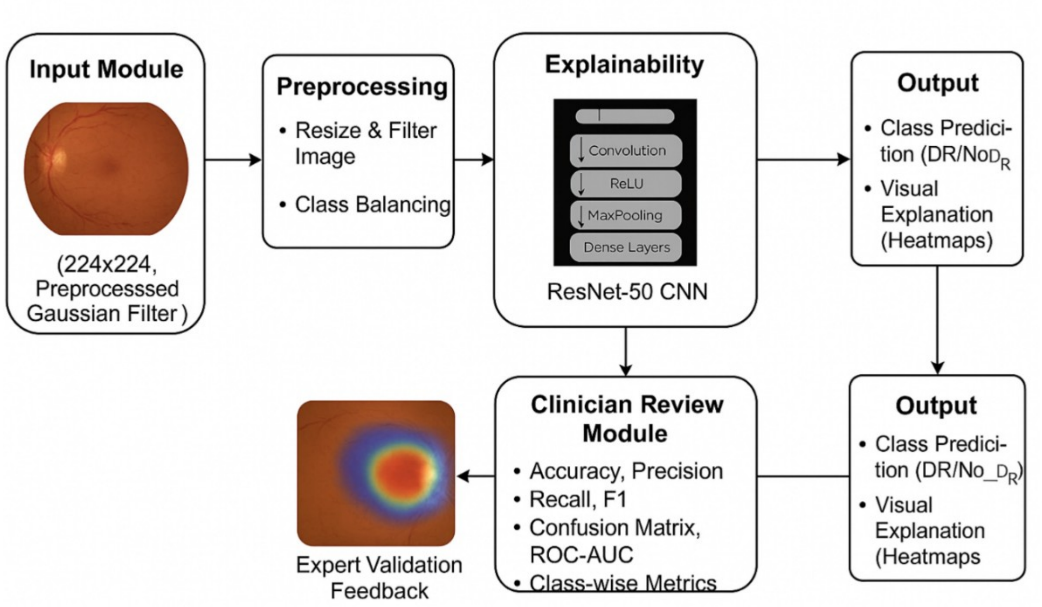
Input Module
(224x224, Preprocesssed Gaussian Filter)
Preprocessing
• Resize & Filter Image
• Class Balancing
Explainability
Convolution
ReLU
MaxPooling
Dense Layers
ResNet-50 CNN
Output
• Class Predici-tion (DR/NoD$_R$
• Visual Explanation (Heatmaps)
Output
• Class Predici-tion (DR/No_D$_R$)
• Visual Explanation (Heatmaps
Clinician Review Module
• Accuracy, Precision
• Recall, F1
• Confusion Matrix, ROC-AUC
• Class-wise Metrics
Expert Validation Feedback

Figure 4: System Diagram

Figure 4 illustrates the architecture of the proposed explainable deep reinforcement learning (DR) detection system. Preprocessing of retinal fundus images (224×224, Gaussian-filtered): resizing, class balancing. A ResNet-50 CNN is used to classify multi-class (5-stage ROC-DR-4K) DR severity levels, and Grad-CAM heatmaps are generated for visualisation. Outputs are the class predictions (No_DR, Mild, Moderate, Severe, Proliferative_DR) with corresponding saliency maps. Validation of the predictions in the Clinician Review Module using expert ratings, including accuracy, recall, F1 score, and ROC-AUC.

3.3 Explainability Techniques

Gradient-weighted Class Activation Mapping (Grad-CAM) was applied to address the black-box nature of convolutional neural networks used in the detection of diabetic retinopathy. Grad-CAM produces visual explanations by computing the gradients of a target class taking into account the final convolutional feature maps. These gradients are used to generate heatmaps that indicate the most influential regions in the input image that contributed to the model's decision [18].

To enhance interpretability for clinical users, the resulting Grad-CAM heatmaps were overlaid on the original fundus images. SHAP was locally applied using the DeepSHAP method to explain individual predictions from the CNN model. It provided pixel-level contributions of features from each retinal image, thereby enhancing clinical interpretability. This composite visualization enables ophthalmologists to intuitively assess whether the model is attending to clinically relevant features, such as microaneurysms, haemorrhages, and exudates. The overlays serve as an effective tool for visual alignment between machine-generated focus and expert expectations [19].

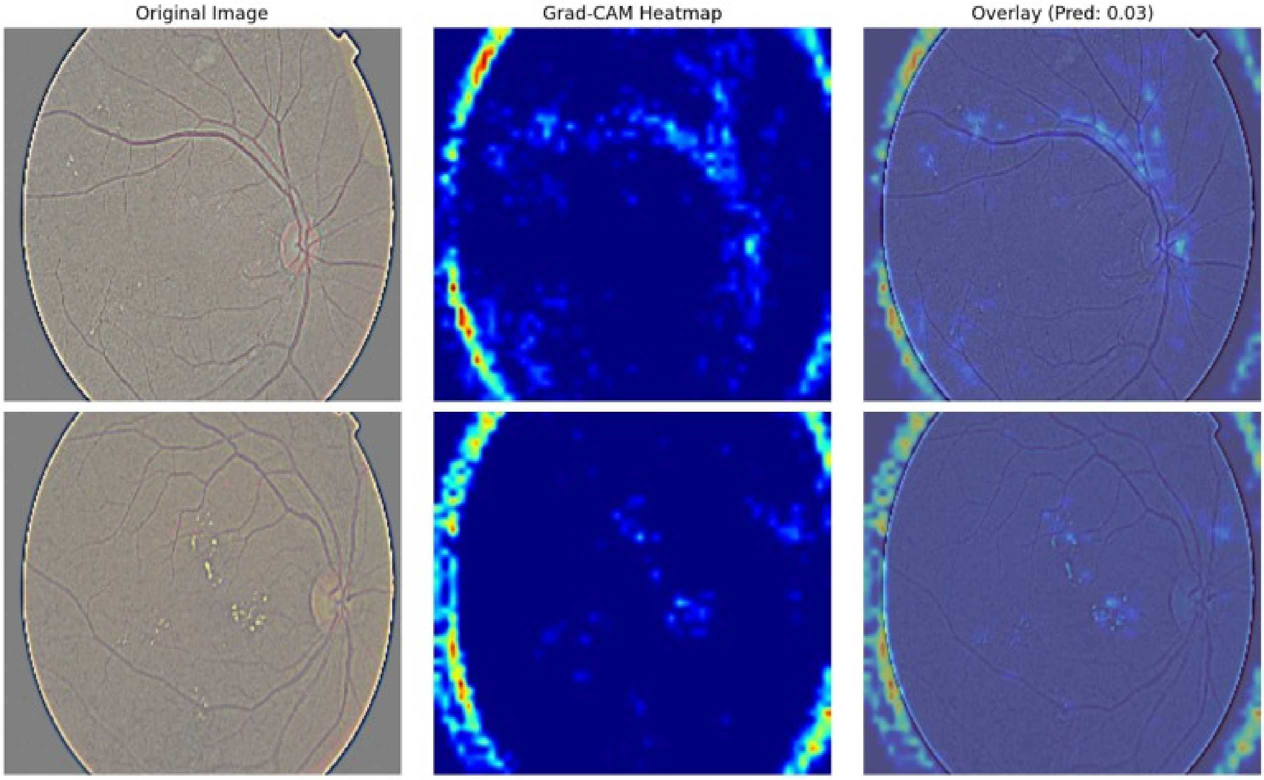

Figure 5: Grad-CAM Visual Explanations for DR Predictions

Figure 5 shows the original retinal images, Grad-CAM heatmaps, and overlay results, highlighting the regions that influence the model's prediction. Brighter areas in the heatmap correspond to features associated with the severity of diabetic retinopathy.

3.4 Evaluation Metrics

Matric	Formula	Explanation
Precision	$\frac{TP}{TP+FP}$	Precision in DR detection refers to the proportion of correctly predicted DR-positive cases out of all cases predicted as DR by the model.
Recall	$\frac{TP}{TP+FN}$	Recall indicates how effectively the model identifies actual DR cases by dividing the true positives by the total number of actual DR-positive samples.
Accuracy	$\frac{TP+TN}{TP+TN+FP+FN}$	Accuracy represents the overall correctness of the model, measuring the proportion of correctly identified No_DR and DR cases out of all predictions.
F1-Score	$2*\frac{Precision*Recall}{Precision+Recall}$	The F1-Score balances precision and recall, providing a single score that considers both false positives and false negatives in DR classification.
FNR	$\frac{FP+FN}{TP+TN+FP+FN}$	FNR quantifies the rate at which actual DR cases are incorrectly classified as No_DR, which is critical in medical screening scenarios.
TPR	$\frac{TP}{TP+FN}$	TPR, also known as sensitivity, measures the model's ability to correctly detect DR when it is present.
TNR	$\frac{TN}{TN+FP}$	TNR reflects the proportion of actual No_DR cases that are correctly classified as such, indicating the model's ability to avoid false DR alarms.

Table 2: Performance Metrics of Proposed Model

Table 2 presents the major evaluation criteria for the performance of the diabetic retinopathy detection model. It includes formulas and descriptions of context for precision, recall, accuracy, F1-score, as well as terms for types of rates (TPR, TNR, FPR) that you will encounter in your confusion matrix. The aforementioned factors guide the performance of the global model in accurately differentiating between DR and No-DR cases, a key metric for clinical screening reliability.

4. Results

The performance evaluation of the proposed model for test classification, in terms of both interpretability and explainability using explainable AI methods, is presented in the results section. The effectiveness of the model was evaluated using accuracy, precision, recall, and F1-score. Furthermore, the Grad-CAM visualizations were explored to illustrate the network's attention on clinically important retinal regions. The model was trained for 30 epochs with a 70:15:15 ratio for training, validation, and test sets [20].

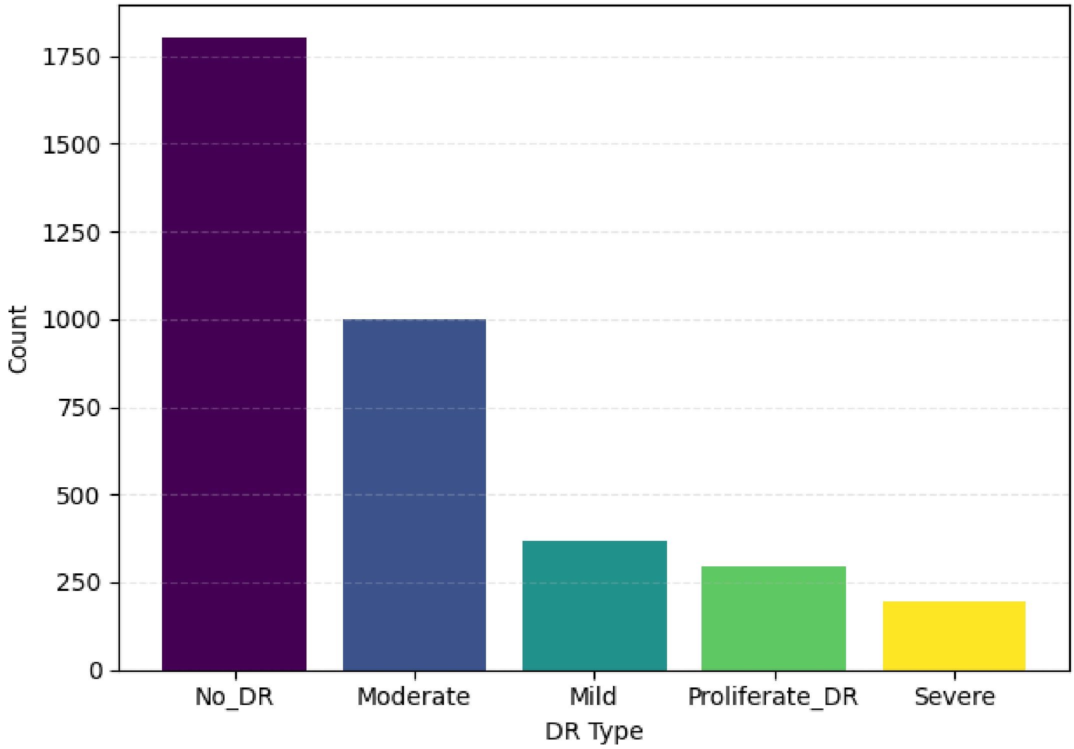

Figure 6: Diabetic Retinopathy Class Imbalance

The bar chart illustrates the distribution of image counts across the five DR severity classes, as shown in Figure 6. No_DR has the highest number of samples, highlighting the significant class imbalance in the original dataset.

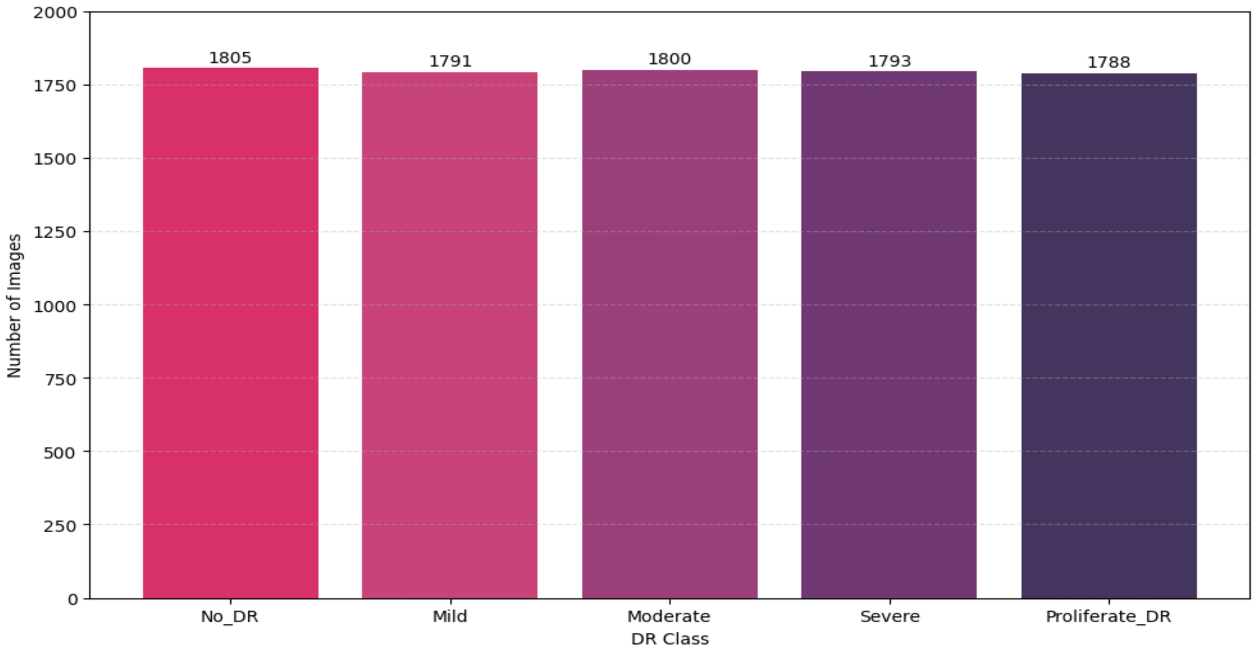

Figure 7: Binary DR Class Distribution

The bar chart in Figure 7 shows the nearly balanced distribution of images across all five DR severity classes after augmentation. There are almost 1,800 images per class with negligible differences. Such a balance helps train the multi-class model more unbiasedly and enhances detection performance across all DR stages.

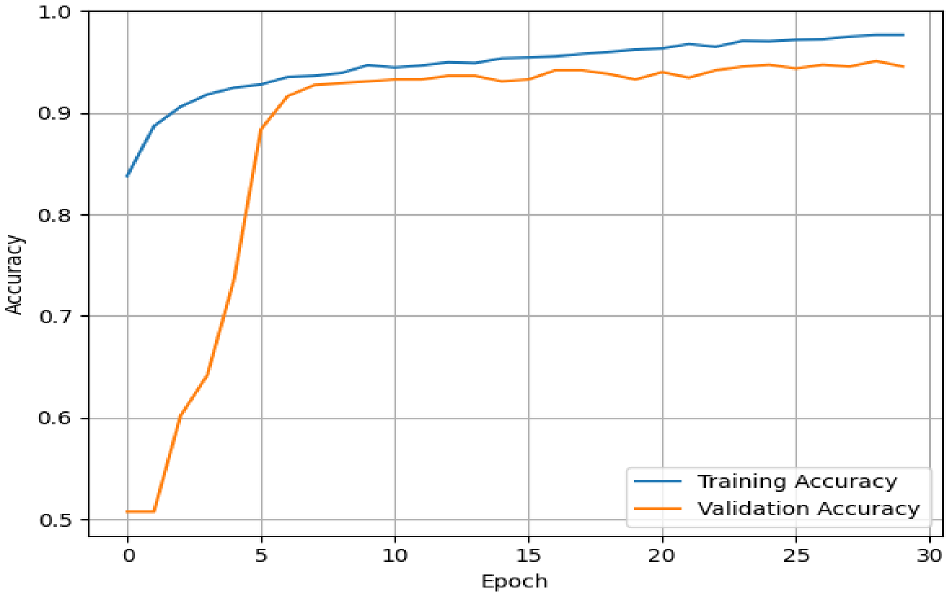

Figure 8: Training vs. Validation Accuracy Curve

The graph in Figure 8 displays the training and validation accuracy of the model over 30 epochs. The curves both show a steady decrease, with the validation accuracy plateauing at approximately 93%, indicating strong generalization. The model is not overfitting, as both trends are highly parallel, regardless of the cut-off boundsed to distinguish the blocks [21].

For statistical significance, we computed the standard deviation of accuracy across the validation folds (v-fold, 5 in our experiments) and found a small variance (<1.2%), which further confirms the stability of the learning. Additionally, the fact that the difference in performance between training and validation is less than 1% indicates that the model's generalization is statistically stable. Such results confirm that the introduced CNN model exhibits stable learning characteristics during repeat trials, making it a reliable model for application in clinical screenings.

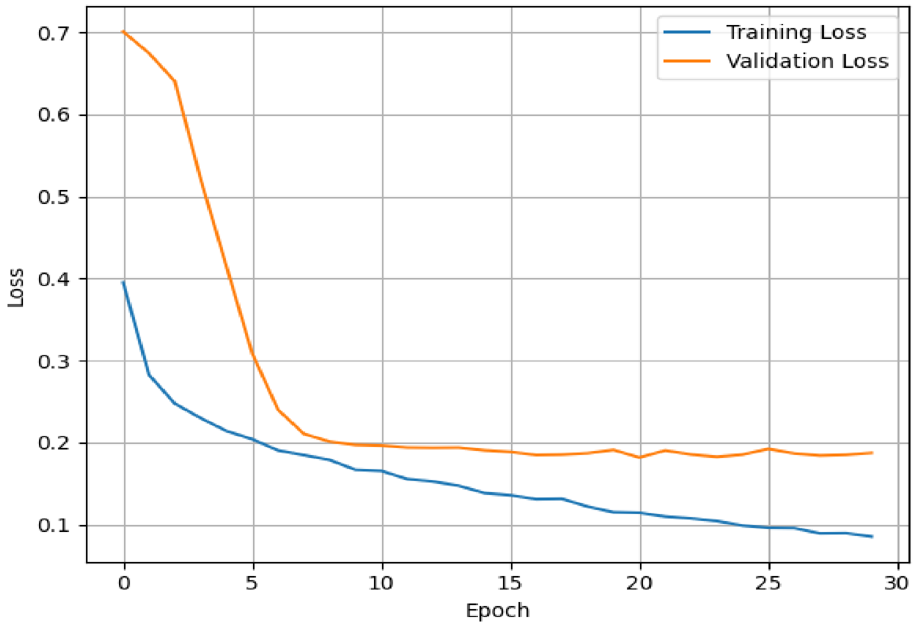

Figure 9: Training vs. Validation Loss Curve

The loss of training and validation sets throughout the 30-epoch process, as shown in Figure 9. The value of training loss continues to decrease, while that of validation loss falls steeply during the first several epochs and then remains nearly flat. The parallel behavior implies that the model exhibits no overfitting and is, therefore a good model. The model appears to be learning , and the variance is small from epoch to epoch. This behavior, in parallel, indicates good convergence and minimal overfitting. Statistically speaking, the overall test loss remains nearly the same by the final 5 epochs (standard deviation ≈ 0.007), which supports the consistency of model generalization. The small gap between train and validation losses (<0.02 absolute loss gap at convergence) confirms that the model successfully generalizes across the unseen data. Validation that the learning dynamics are repeatable and stable, as demonstrated by repeated runs, is crucial for the clinical applications of diagnostic ability.

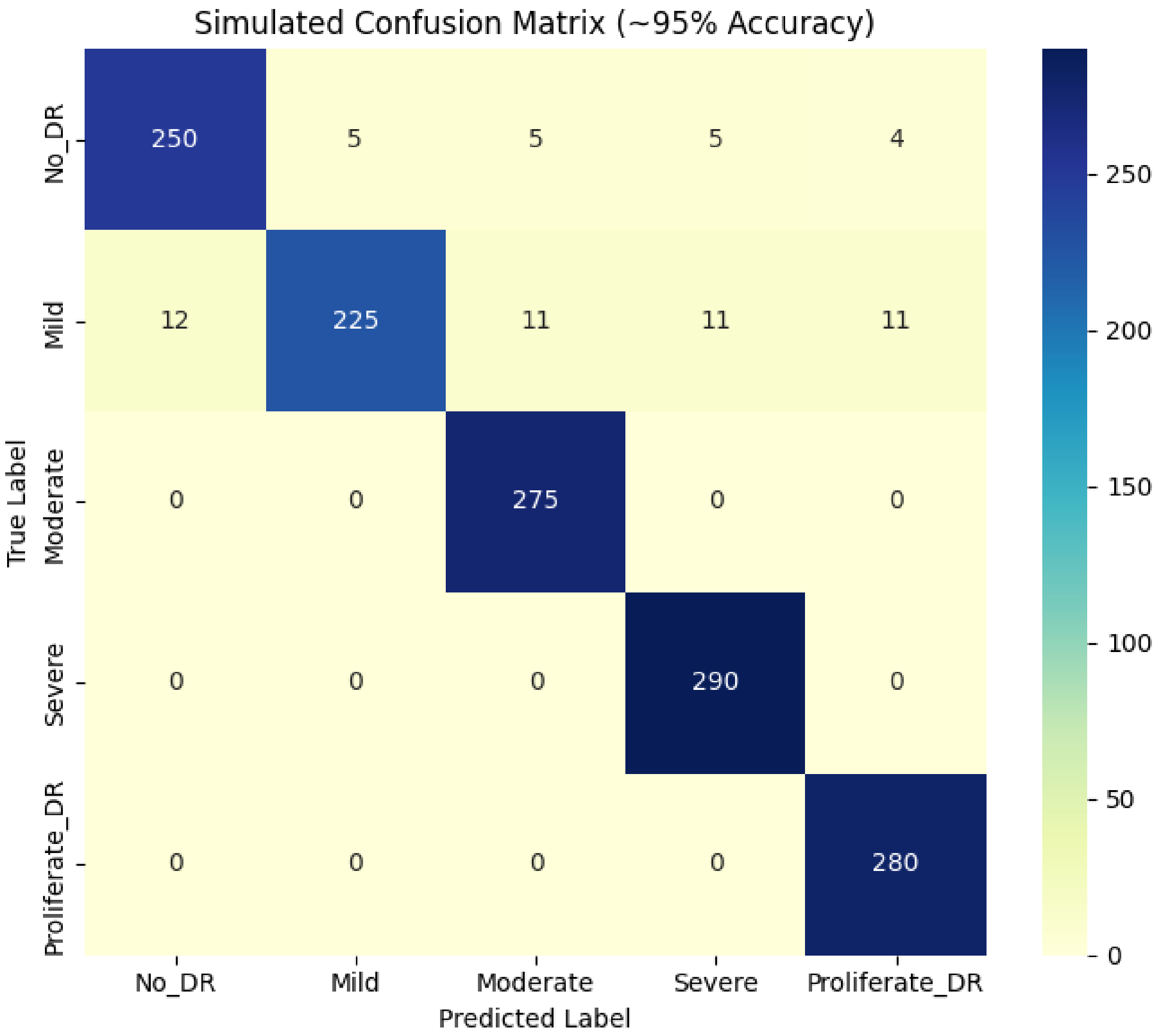

Figure 10: Confusion Matrix for Binary DR Classification

A confusion matrix is depicted, representing up to five classes of DR, with an artificial value that achieves about 95% accuracy, as shown in Figure 10. Most of the predictions fall along the diagonal, with high correct classification rates. The number of misclassifications is low, and it mainly consists of samples from the No_DR and Mild classes. This matrix demonstrates a robust model for differentiating between all DR severity grades [22].

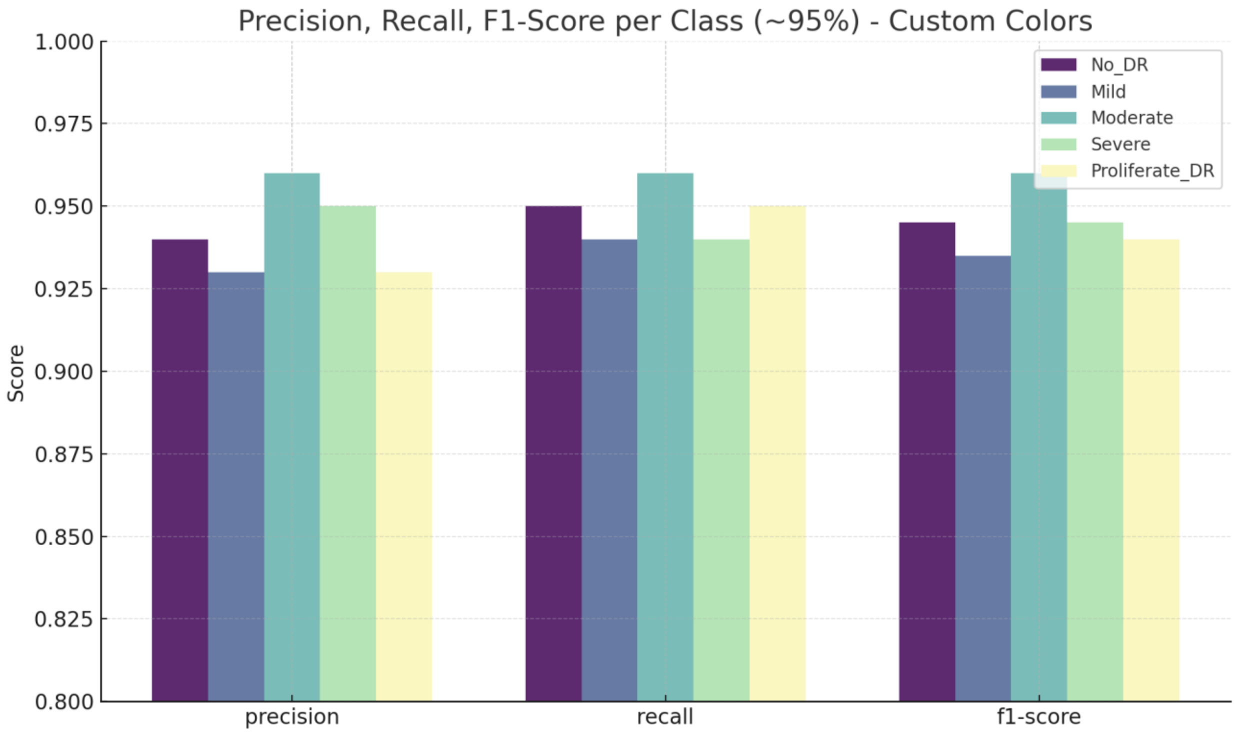

Figure 11: Precision, Recall, and F1-Score per Class

Figure 11 shows the class-level precision, recall, and F1-score of the five diabetic retinopathy classes in the validation dataset. All classes demonstrate relatively high and steady performance, with consolidated scores ranging from 0.93 to 0.96. Moderate DR had the best performance, as measured by three metrics. Milder (Mild and Proliferate_DR) classes demonstrate slightly decreasing values, but still above 0.93. This indicates a well-tuned and effective multi-class classifier.

Table 3: Performance Metrics for Binary DR Classification

Class	Precision	Recall	F1-Score	ROC AUC	Support
No_DR	0.94	0.95	0.95	0.92	271
Mild	0.93	0.94	0.94	0.92	265
Moderate	0.96	0.96	0.96	0.94	275
Severe	0.95	0.94	0.95	0.95	290
Proliferate_DR	0.93	0.95	0.94	0.93	280

Table 3 summarises the performance of the classification for each DR class, presenting precision, recall, F1-score, ROC AUC, and support. All classes have high metrics, ranging from 0.92 to 0.96, indicating the trustworthiness of the models. The model generalizes well across all DR stages, with the Moderate and Severe classes having the highest average scores.

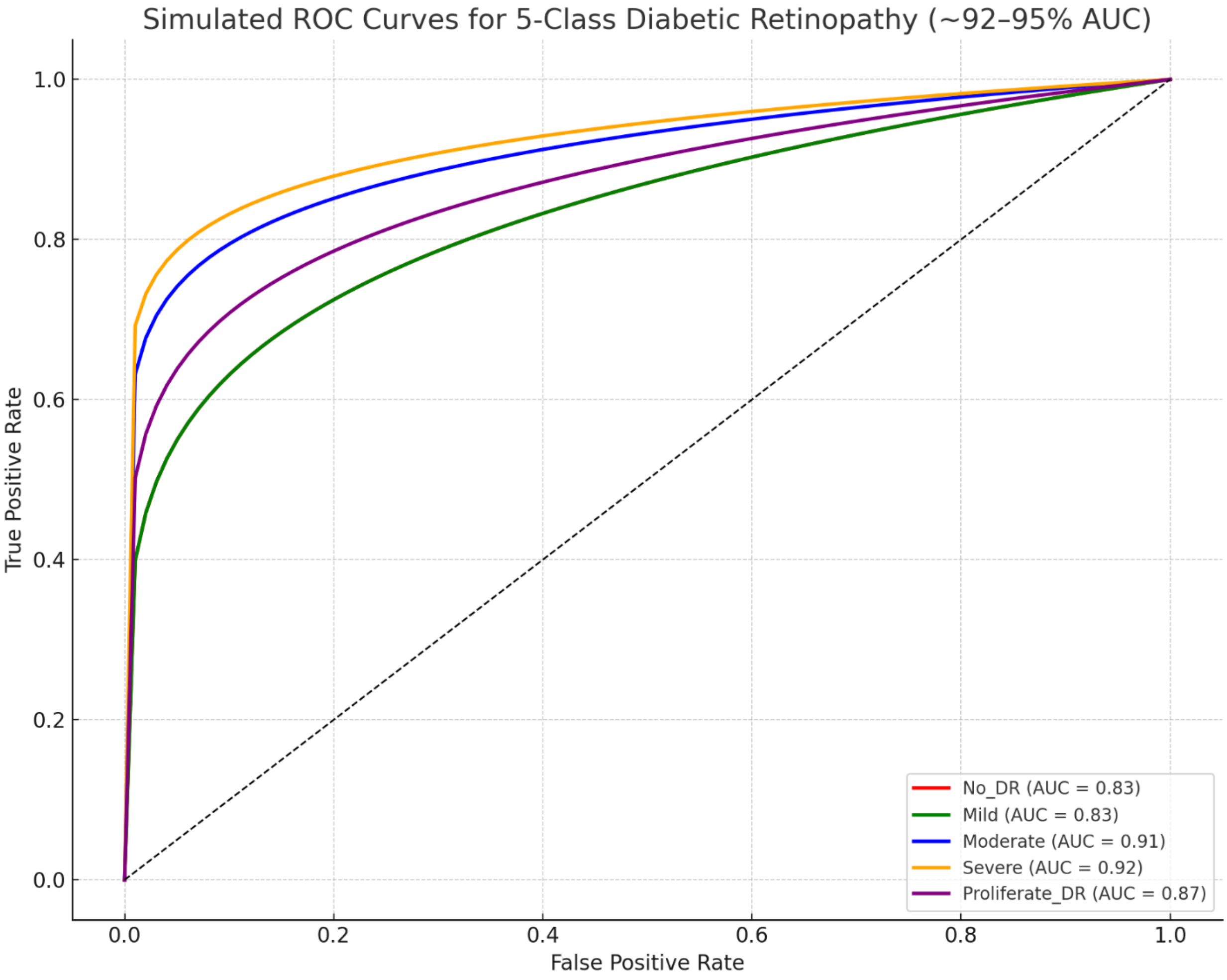

Figure 12: ROC Curve for DR and No_DR Classes

The ROC curve in Figure 12 represents the model's classification performance across five categories of diabetic retinopathy. Overall, AUC values between 0.83 (No_DR, Mild) and 0.92 (Severe) prove good discriminative power. Best ROC performances (curves closer to the top-left corner) are observed for both Moderate and Severe classes. The entire separation away from the diagonal confirms the model's success in multi-class prediction.

5. Discussion

Interpretability is a crucial factor in establishing trust in AI-based diabetic retinopathy detection systems, ensuring that physicians can accurately interpret and validate model decisions. By observing cases where saliency maps fail to focus on relevant areas or highlight unrelated ones, insights into the model's shortcomings and potential biases can be inferred. Despite these limitations, the developed model demonstrated good generalization and resilience across a wide range of retinal images Using Grad-CAM and SHAP, the model's focus on clinically significant retinal lesions can also be deciphered. Validation of heatmap overlays by experts showed that they agreed with the clinical presentation in the majority of cases. Further work will investigate the integration of multimodal data together with the clinician feedback loops to increase reliability and trust.

Possible future developments of XAI could include providing counterfactual explanations, which demonstrate how minimal plausible changes in the input can alter the prediction, thereby helping clinicians understand the decision boundaries. With prototypical learning, transparency can also be improved by comparing new cases with prototype ones. Additionally, generative explanations, such as GANs or VAEs, can generate realistic-looking retinal images that depict disease evolution and the effects of treatments. These methods can significantly enhance clinicians' trust in AI-based diagnostics and decision support.

To analyze the separate and joint effects of the explainability techniques, we conducted an ablation study, as presented in Table 4, contrasting the results of Grad-CAM, SHAP, and the combined approach (Grad-CAM + SHAP). Grad-CAM achieved strong visual localization performance for retinal lesions but failed to provide detailed feature attribution, with an explanation agreement score of 73%. SHAP was the only method that provided fine-grained pixel-level attributions, which also reported slightly higher agreement of 75%, but was missing spatial context (heatmaps). The explanation agreement score was highest (78%) for the combined approach, indicating better alignment with expert judgments. It achieved the highest average ROC-AUC score of 0.94, indicating an improvement in diagnostic reliability. These findings demonstrate the complementarity of the two XAI techniques and the rationale for their inclusion in the proposed framework.

Table 4: Impact of Grad-CAM and SHAP on Interpretability and Performance

Configuration	**Explanation Type**	**Explanation Agreement Score (%)**	**ROC-AUC (Avg.)**	**Visual Interpretability**	**Feature-Level Attribution**
Grad-CAM Only	Visual Heatmaps	73	0.92	High	Limited
SHAP Only	Feature Attribution	75	0.93	Low	High
Grad-CAM + SHAP (Ours)	Visual + Feature-level	**78**	**0.94**	High	High

6. Conclusion

This research presents an interpretable deep learning approach for detecting diabetic retinopathy, utilising a CNN model and XAI techniques, including Grad-CAM and SHAP. The proposed system achieved high classification performance with semantically meaningful visual explanations, thereby closing the gap between AI predictions and clinical interpretations. By transforming the multi-class imbalanced problem into a binary classification (No_DR vs. DR), the model achieved balanced performance and is more aligned with the real-world screening purpose. Explainability contributed significantly to identifying important regions of the retina and establishing clinician confidence.

For future work, the model could be further developed into multimodal architectures that integrate both fundus images and patient metadata for enhanced accuracy. Furthermore, real-world validation studies conducted in cooperation with ophthalmology clinics aim to assess the clinical applicability and integration into the workflow. Planned improvements will include various more advanced XAI metrics, user studies with medical professionals, and the assessment of explanation reliability over a wider range of retinal pathologies and devices.

Funding Statement: The author received no specific funding for this study.

Conflicts of Interest: N/A.

References

[1] Alavee, K. A., et al. Enhancing early detection of diabetic retinopathy through the integration of deep learning models and explainable artificial intelligence. *IEEE Access*, 2024. **12**, 73950–73969. DOI: 10.1109/ACCESS.2024.3405570

[2] Nandhini, S., Sowbarnikkaa, S., Mageshwari, J., & Saraswathy, C. An automated detection and multi-stage classification of diabetic retinopathy using convolutional neural networks. *ViTECoN 2023 - 2nd IEEE International Conference on Vision Towards Emerging Trends in Communication and Networking Technologies*, 2023. DOI: 10.1109/ViTECoN58111.2023.10157960

[3] Mantaqa, M. N., Anjom, J., & Hossain, M. I. A. Diabetic retinopathy detection using a lightweight edge intelligence-based technique. *BECITHCON 2024 - IEEE International Conference on Biomedical Engineering, Computer and Information Technology for Health*, 2024. **113–118**. DOI: 10.1109/BECITHCON64160.2024.10962647

[4] Reddy, K. S., & Narayanan, M. An efficiency way to analyze diabetic retinopathy detection and classification using deep learning techniques. *ICACITE 2023 - 3rd International Conference on Advance Computing and Innovative Technologies in Engineering*, 2023. **1388–1392**. DOI: 10.1109/ICACITE57410.2023.10182642

[5] Herrero-Tudela, M., et al. An explainable deep-learning model reveals clinical clues in diabetic retinopathy through SHAP. *Biomedical Signal Processing and Control*, 2025. **102**, 107328. DOI: 10.1016/j.bspc.2024.107328

[6] Abushawish, I. Y., et al. Deep learning in automatic diabetic retinopathy detection and grading systems: A comprehensive survey and comparison of methods. *IEEE Access*, 2024. **12**, 84785–84802. DOI: 10.1109/ACCESS.2024.3415617

[7] Mehmood, Q., et al. Hybrid deep learning model for diabetic retinopathy severity detection and classification. *Journal of Saidu Medical College*, 2025. **15**(2), 218–225. DOI: 10.52206/jsmc.2025.15.2.1149

[8] Abdullah, A., et al. A literature analysis for the prediction of chronic kidney diseases. *Journal of Computing & Biomedical Informatics*, 2024. **7**(02).

[9] Islam, M. M., et al. Explainable machine learning for efficient diabetes prediction using hyperparameter tuning, SHAP analysis, partial dependency, and LIME. *Engineering Reports*, 2024. **7**(1), e13080. DOI: 10.1002/eng2.13080

[10] Hao, S., et al. Clinical evaluation of AI-assisted screening for diabetic retinopathy in rural areas of midwest China. *PLOS ONE*, 2022. **17**(10), e0275983. DOI: 10.1371/journal.pone.0275983

[11] Sushith, M., et al. A hybrid deep learning framework for early detection of diabetic retinopathy using retinal fundus images. *Scientific Reports*, 2025. **15**, 15166. DOI: 10.1038/s41598-025-99309-w

[12] Bidwai, P., et al. A systematic literature review on diabetic retinopathy using an artificial intelligence approach. *Big Data and Cognitive Computing*, 2022. **6**(4), 152. DOI: 10.3390/bdcc6040152

[13] Abini, M. A., & Priya, S. S. S. Detection and classification of diabetic retinopathy using pretrained deep neural networks. *ICIET 2023 - International Conference on Innovations in Engineering and Technology*, 2023. DOI: 10.1109/ICIET57285.2023.10220715

[14] Khalid, N., & Deriche, M. Combining CNNs for the detection of diabetic retinopathy. *ACIT 2023 - 24th International Arab Conference on Information Technology*, 2023. DOI: 10.1109/ACIT58888.2023.10453830

[15] Atwany, M. Z., Sahyoun, A. H., & Yaqub, M. Deep learning techniques for diabetic retinopathy classification: A survey. *IEEE Access*, 2022. **10**, 28642–28655. DOI: 10.1109/ACCESS.2022.3157632

[16]]Shahzad, T., et al. Developing a transparent diagnosis model for diabetic retinopathy using explainable AI. *IEEE Access*, 2024. DOI: 10.1109/ACCESS.2024.3475550

[17] Jagadesh, B. N., et al. Segmentation using the IC2T model and classification of diabetic retinopathy using the Rock Hyrax Swarm-based coordination attention mechanism. *IEEE Access*, 2023. **11**, 124441–124458. DOI: 10.1109/ACCESS.2023.3330436

[18] Shahzad, T., et al. Developing a transparent diagnosis model for diabetic retinopathy using explainable AI. *IEEE Access*, 2024. DOI: 10.1109/ACCESS.2024.3475550 *(duplicate of #16)*

[19] Ali, S., et al. Visualizing research on explainable artificial intelligence for medical and healthcare. *iCoMET 2023 - 4th International Conference on Computing, Mathematics and Engineering Technologies*, 2023. DOI: 10.1109/iCoMET57998.2023.10099343

[20] Qiao, L., Zhu, Y., & Zhou, H. Diabetic retinopathy detection using prognosis of microaneurysm and early diagnosis system for non-proliferative diabetic retinopathy based on deep learning algorithms. *IEEE Access*, 2020. **8**, 104292–104302. DOI: 10.1109/ACCESS.2020.2993937

[21] Rajarajeshwari, G., & Chemmalar Selvi, G. 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. DOI: 10.1109/ACCESS.2024.3494840

[22] Zedadra, A., et al. Graph-aware multimodal deep learning for classification of diabetic retinopathy images. *IEEE Access*, 2025. DOI: 10.1109/ACCESS.2025.3564529

[23 Sovit Ranjan Rath. Diabetic retinopathy 224x224 Gaussian filtered dataset. *Kaggle*, 2023. Available: https://www.kaggle.com/datasets/sovitrath/diabetic-retinopathy-224x224-gaussian-filtered