Journal of Applied Informatics and Computing (JAIC)
Vol.10, No.3, June 2026, pp. 2534~2545
e-ISSN: 2548-6861

Explainable Deep Learning for Diabetic Retinopathy Detection: A Quantitatively Validated Framework

Tinashe Ngwazi [1]*, Belinda Ndlovu [2], Kudakwashe Maguraushe [3]***
[1,2]Department of Informatics, National University of Science and Technology
[3] School of Computing, University of South Africa, Pretoria, South Africa
tinashengwazi@gmail.com [1], belinda.ndlovu@nust.ac.zw [2], magark@unisa.ac.za [3]

Article Info

Article history:

Received 2026-03-20
Revised 2026-04-19
Accepted 2026-05-18

Keyword:

Diabetic Retinopathy Detection,
Explainable Artificial Intelligence (XAI),
Deep Learning,
Medical Image Classification,
Integrated Gradients,
Grad-CAM,
Convolutional Neural Networks (CNNs),
Clinical Decision Support Systems.

ABSTRACT

Diabetic retinopathy (DR) is a leading cause of preventable blindness, where early and accurate detection is critical for effective intervention. While deep learning models have demonstrated strong performance in DR classification, their limited interpretability and inconsistent evaluation practices hinder clinical trust and deployment. This study proposes an explainable deep learning framework for DR detection based on MobileNetV2, complemented by Integrated Gradients for feature attribution. A curated dataset of 4,464 retinal images was constructed from publicly available sources through systematic preprocessing, including quality filtering, deduplication, and class balancing across five DR stages. To ensure robust evaluation, a multi-level validation strategy was employed, incorporating stratified train–validation–test splits and k-fold cross-validation. The proposed framework achieved 87.0% accuracy and an F1-score of 0.868, outperforming baseline models including EfficientNet-B0, DenseNet121, and VGG16. Beyond predictive performance, explainability was quantitatively evaluated using deletion and insertion metrics, demonstrating that Integrated Gradients provides more faithful feature attribution compared to Grad-CAM and LIME. Error analysis further reveals that misclassifications are concentrated between adjacent DR stages, reflecting the inherent difficulty of fine-grained disease progression modelling. The findings highlight that combining rigorous validation with quantitative explainability evaluation can improve the reliability and transparency of deep learning models for medical imaging. While results are promising, the framework is validated on publicly available datasets and requires further external clinical validation before real-world deployment.

I. INTRODUCTION

Diabetes is a condition in which the body cannot effectively regulate blood sugar levels [1]. As this disease progresses, it may lead to vision loss, a condition called diabetic retinopathy [2]. Diabetic retinopathy is a common cause of vision loss in the middle-aged and elderly [3]. When left unchecked, this condition may threaten the livelihoods of countless middle-aged and older adults in society, negatively impacting their quality of life [4], which makes disease management a critical concern. Detecting and predicting diabetic retinopathy early is key to preventing vision loss [5]. Traditionally, healthcare providers have relied on manual inspection and laboratory examinations to identify diabetic retinopathy [6]. While these methods have been impactful, they are often slow, labour-intensive, and prone to human error [7]. Fortunately, advancements in Artificial Intelligence (AI) and its subfields, such as Machine Learning (ML), are changing the landscape of diabetic retinopathy detection [8]. These technologies have introduced powerful new tools for predicting diseases with remarkable speed and accuracy [9][10].

Machine learning models have recently been widely used to detect diabetic retinopathy, offering a more efficient alternative to traditional diagnostic methods [11]. Techniques such as Convolutional Neural Networks (CNNs) have proven

highly effective in analysing retinal images, incorporating demographic data and other factors [12], [13]. By identifying the varying stages of diabetic retinopathy in large datasets, these models enable timely intervention, support clinical decision-making, and improve patient outcomes [14].

Despite machine learning models having demonstrated great success in predicting diabetic retinopathy, their widespread adoption is hindered by their lack of transparency [15]. Healthcare providers and experts certainly need accurate predictions, but they must also understand the reasoning behind them [16]. If the factors influencing these predictions are unclear, it becomes difficult to trust and act upon the model's recommendations [2]. Without interpretability, the real-world impact of machine learning in detecting diabetic retinopathy is significantly reduced.

Despite the growing adoption of deep learning models for diabetic retinopathy detection, three critical gaps remain insufficiently addressed in existing studies. First, many models lack reproducible pipelines for constructing datasets, limiting transparency and comparability across studies. Second, validation strategies often rely on single train–test splits, raising concerns about robustness and generalisability. Third, while explainable AI techniques are increasingly incorporated, their evaluation is predominantly qualitative, with limited use of quantitative faithfulness metrics or structured expert validation.

This study does not introduce a new architecture, but instead contributes a rigorously validated, well-documented and methodologically transparent framework that integrates (1) data-centric dataset construction with explicit filtering and balancing, (2) multi-level validation to mitigate data leakage and improve robustness, and (3) quantitative evaluation of explainability using faithfulness metrics. This shifts the contribution from model novelty toward methodological rigour, reproducibility, and clinically meaningful interpretability.

This study contributes to the emerging paradigm of data-centric and explainable artificial intelligence (XAI) in healthcare, where transparency, robustness, and reproducibility are essential for real-world deployment.

II. BACKGROUND OF STUDY

From all across the globe, diabetes has been a prevalent disease affecting about 424.9 million people, with a further estimated increase of 48% to this margin by the year 2045 [17]. This complex disease requires extensive and continuous monitoring for effective management, such as quantified-self technologies and the Internet of Things [18][19][20]. As diabetes progresses, it may lead to vision loss, a condition called diabetic retinopathy, which is said to be a major ocular complication of diabetes [21]. Diabetic retinopathy remains the leading cause of preventable blindness among working-age adults worldwide [22].

Across different ethnicities, it has been found that Asians have a 19.9% prevalence of diabetic retinopathy, compared to Caucasians (45.7%), African-Americans (49.6%), and Hispanics (34,6%) [23], mainly attributed to age and increased diabetes duration in the regions. Further research shows that the number of people affected by diabetic retinopathy will double in 20 years, from 12 million to 24 million in 2030, likely due to poor healthcare systems [24]. The study by [25] makes an example of the poor performance of cervical cancer prevention programs in developing countries, likely due to inadequate planning and lack of resources, implying that it is the same case with diabetic retinopathy. In low-income countries with developing economies, there is a shortage of healthcare resources and healthcare providers [26], leading to futile efforts to detect and treat diabetic retinopathy early.

Diabetic retinopathy is one of the common complications related to diabetes, as [5] describes it as the fifth leading cause of blindness globally. Detecting the disease in its earliest stages is critical in reducing the risk of vision loss [27]. There is a need to streamline the early detection of diabetic retinopathy using machine learning models to address the shortage of healthcare professionals and facilities in developing countries [28]. Innovative AI technologies, such as ML and XAI, have demonstrated promising results in streamlining the detection and classification of diabetic retinopathy [29], thereby improving diagnosis and treatment. The role of XAI is to make the machine learning models more transparent so that healthcare users can understand the predictions made by the algorithms [30][31]. These techniques are instrumental in delivering cost-effective, faster screening [32], especially in low-income countries without resources or healthcare professionals [26]. Therefore, this study systematically investigates and compares various explainable AI (XAI) techniques for predicting diabetic retinopathy, with a primary focus on assessing how they enhance the transparency of machine learning models. To address these gaps, the study is guided by the following research questions:

- **RQ1** How can a well-documented and methodologically transparent deep learning framework be designed to achieve reliable classification of diabetic retinopathy across multiple disease stages?
- **RQ2** To what extent do data-centric design choices, including class balancing and data augmentation, influence the performance of deep learning models for diabetic retinopathy detection?
- **RQ3** How effectively can explainable artificial intelligence methods, specifically Integrated Gradients, provide potentially clinically relevant interpretations of model predictions?
- **RQ4** How robust and generalisable is the proposed framework when evaluated using multi-level validation strategies and baseline comparisons?

These questions are designed to move beyond model implementation toward a more rigorous evaluation of performance, interpretability, and reproducibility in medical image classification.

III. RELATED WORKS

A. Application of Machine Learning Models in Diabetic Retinopathy Detection

A wide range of ML techniques has been deployed to detect retinal diseases such as diabetic retinopathy. ML techniques commonly used for diabetic retinopathy detection include classical ensemble methods, such as Random Forest (RF).

Ensemble models have increased classification and diabetic retinopathy identification accuracy, yielding good results. In the study by [31], an ensemble approach for diabetic retinopathy detection achieved 75.1% performance, using 4 sub-datasets. The application of ensemble methods in diabetic retinopathy detection has performed well, yielding high accuracy, as shown in the study by [32], which used ResNet-50 and Random Forest for classification, achieving accuracies of 96% and 75.09% across 2 datasets. This is further supported by a study [33], which found that an ensemble comprising ResNet50, Inceptionv3, Xception, DenseNet121, and DenseNet169 achieved an accuracy of 80.8%.

B. Application of Deep Learning (DL) Models in Diabetic Retinopathy Detection

The detection of diabetic retinopathy has seen considerable improvement through the adoption of deep learning algorithms such as Convolutional Neural Networks (CNNs), Deep Neural Networks (DNNs), and transfer learning [33]. A CNN-based model proved very accurate for early detection of diabetic retinopathy [34], achieving 14% higher accuracy than SVM-based classifiers and 7% higher accuracy in 10-fold cross-validation. A Visual Geometry Group-16 (VGG16) model was proposed in a study by [35], achieving 74.58% accuracy, with the main limitation being the small dataset size of 1728 images. EfficientNet proved to be an effective technique for early detection of diabetic retinopathy, as supported by the study by [36], achieving an accuracy of 85.5% and a kappa score of 0.921 on the Aptos and DDR datasets.

Deep Neural Networks are very effective at detecting diabetic retinopathy [37]. An accuracy of 88.72% was achieved with a deep neural network, boasting a consistency score of 91.8%. Furthermore, we see the effectiveness of deep neural networks in [38], where the DNN model achieved a sensitivity of 97.5% and an area under the curve of 97.7%. This is evidence that the adoption of DNNs has improved diabetic retinopathy detection.

Another deep learning approach is transfer learning. A study by [39] demonstrated a robust performance and accuracy of 87.27%, with 83.76% sensitivity and 90.82% specificity. A 90.9% accuracy was achieved in a study by [40], with the model trained on 5000 unseen fundus images and achieving a loss of 3.94%. Despite the potential of machine learning and deep learning for the early detection of diabetic retinopathy, understanding and interpreting their outputs remains complex, necessitating Explainable Artificial Intelligence (XAI) to improve model transparency.

C. Application of Explainable Artificial Intelligence (XAI) Models in Diabetic Retinopathy Detection

Explainable Artificial Intelligence involves the use of techniques like Gradient-weighted class activation mapping (Grad CAM), Local Interpretable Model-agnostic Explanation (LIME), Shapley Additive Explanations (SHAP) [41][42]. Several research studies have leveraged the unique capabilities of XAI to enhance the interpretation of results from machine learning and deep learning models, making them easier to understand. A study by [43] used LIME and Grad CAM to improve the ResVIT FusionNet model, with LIME highlighting specific pixel-level regions in retinal images and Grad CAM generating broader region-based heatmaps to visualise areas of higher significance. In the study by [44], SHAP was leveraged to interpret each feature's contribution, providing valuable insights into the metabolic and physiological markers associated with the various stages of diabetic retinopathy.

D. Opportunities Presented by XAI in Diabetic Retinopathy Detection

In the context of diabetic retinopathy, XAI offers several critical opportunities to advance the landscape of early diagnosis and treatment. XAI supports personalised treatment [45][46][47], providing strategies to enhance the collaboration between AI-based systems and healthcare providers, leading to tailored treatment plans. Integrating XAI into clinical systems supports clinical decision-making, resulting in more reliable systems [48][49]. XAI is also instrumental in ensuring regulatory compliance with standards such as GDPR, which requires clear explanations for decisions that affect patient outcomes [50].

E. Challenges in Incorporating XAI into Early Diabetic Retinopathy Detection

Although XAI offers interpretability and understanding of ML models, it may still be complex for non-expert users [51]. Another challenge is the dataset size and quality, which diminishes the model's efficiency and results in less understandable XAI interpretations [52][53]. XAI-based models required higher computational resources, as there is a huge trade-off between performance and interpretability [50].

IV. MATERIALS AND METHODS

A. Data Sources and Dataset Construction

Two publicly available retinal fundus image datasets were utilised in this study: the Diabetic Retinopathy 224×224 (2019) dataset (3,662 images) and the Diabetic Retinopathy 2015 Colored Resized dataset (35,126 images), both obtained from Kaggle. These datasets have been widely used in prior

diabetic retinopathy (DR) studies, supporting comparability with existing work [29].

The dataset construction process resulted in a progressive reduction in sample size due to strict quality control, redundancy removal, and class balancing, as summarised in Table 1. The final dataset was intentionally balanced across all diabetic retinopathy stages to minimise bias during model training. The initial combined dataset consisted of 38,788 images. Quality filtering was applied to remove images with insufficient diagnostic value, including those that were blurred, low-contrast, or poorly illuminated. Image quality was assessed using objective criteria, including the Laplacian variance for sharpness and intensity-based thresholds for illumination consistency.

Following quality filtering, duplicate and near-duplicate images were identified and removed using perceptual hashing (pHash) with a similarity threshold of 0.9. This step *reduced* redundancy and mitigated the risk of information leakage across training and evaluation sets.

TABLE 1
DATASET CONSTRUCTION SUMMARY

Stage	Images Remaining
Initial Combined Dataset	38,788
After Quality Filtering	12,950
After Deduplication	10,872
Final Balanced Dataset	4,464

To address class imbalance, stratified sampling was employed to ensure balanced representation across the five DR categories: No_DR, Mild, Moderate, Severe, and Proliferative_DR. The final dataset comprised 4,464 images, evenly distributed across classes.While class balancing improves model stability during training, it may not reflect the true disease prevalence in the real world. This design choice was made to ensure equal representation of all DR stages during learning, and its implications for real-world deployment are discussed in the limitations.

All images were standardised to a resolution of 224 × 224 pixels in RGB format. Pixel values were normalised to the range [0,1] to stabilise training and improve convergence. Data augmentation techniques, including random rotations, horizontal and vertical flips, and controlled brightness adjustments, were applied to enhance generalisation and reduce overfitting.

B. Data Preprocessing

Image preprocessing was performed to enhance feature quality and ensure consistency across samples. Standardisation was achieved using z-score normalisation, defined as*:*

$$z_{norm} = \frac{X - \mu}{\sigma}$$

where z represents the input pixel value, $\mu$is the mean, and $\sigma$is the standard deviation.

Duplicate detection was performed using perceptual hashing (pHash), with image similarity quantified using the Hamming distance; images with a similarity score above 0.9 were removed. Quality filtering thresholds were defined based on Laplacian variance (threshold < 100) to identify blurred images and illumination variance (threshold < 15) to detect low-contrast or poorly illuminated samples. These thresholds were selected empirically based on visual inspection and are consistent with established practices in medical image preprocessing.

Histogram equalisation was applied to enhance contrast and improve the visibility of retinal structures. Gaussian filtering was used selectively to reduce noise while preserving clinically relevant features. These steps are consistent with established practices in medical image analysis [34].

Data augmentation included random rotations (±15°), horizontal flipping, and brightness adjustments within a limited range. These transformations were applied to improve model generalisation while preserving clinically relevant features.

C. Model Architecture

The proposed framework is based on the MobileNetV2 architecture, chosen for its balance between computational efficiency and feature extraction. MobileNetV2 employs depthwise separable convolutions and inverted residual blocks, enabling effective learning from image data while maintaining a lightweight structure suitable for resource-constrained environments.

The model was adapted for multi-class classification across the five DR categories. Integrated Gradients was introduced as a post hoc explainability method for attributing model predictions to input features, enabling the visual interpretation of salient regions within retinal images [15].

D. Baseline Models and Ablation Design

To contextualise model performance, two widely used deep learning architectures, ResNet50 and EfficientNet-B0, were implemented as baselines under identical experimental conditions. This ensures that performance differences can be attributed to model design rather than variations in training configuration.

An ablation study was conducted to assess the contribution of key components within the proposed framework. The following configurations were evaluated:

- Without data augmentation
- Without class balancing
- Without label smoothing
- Without explainability integration (Integrated Gradients)

This design enables a systematic assessment of how each component influences model performance and interpretability.

E. Model Training and Validation Strategy

The model was trained using categorical cross-entropy loss with the Adam optimiser, with a learning rate of 0.001 and cosine annealing. A batch size of 16 was used to balance computational efficiency and memory constraints. Training was performed for up to 50 epochs with early stopping (patience = 10) to prevent overfitting.

To ensure robust evaluation, a multi-level validation strategy was employed. The dataset was partitioned into training (70%), validation (10%), and held-out test (20%) sets using stratified sampling to preserve class distributions. The validation set was used for hyperparameter tuning, while the test set remained completely unseen during model development.

In addition, 5-fold cross-validation was conducted on the training set to assess model stability across different data partitions. This approach reduces dependence on a single split and improves the reliability of performance estimates [15].

To ensure experimental consistency, a fixed random seed was used across all training and validation procedures. This minimises variability due to random initialisation and data shuffling.

To prevent data leakage, all preprocessing steps, including deduplication and class balancing, were performed prior to dataset splitting. The held-out test set was completely isolated and not used during model training, hyperparameter tuning, or cross-validation. Cross-validation was conducted exclusively on the training set. The workflow followed a sequential process: preprocessing → dataset split → cross-validation on training data → final evaluation on the held-out test set.

F. Evaluation Metrics

Model performance was evaluated using standard classification metrics, including accuracy, precision, recall, and F1-score. The F1-score was computed as:

$$F1 = 2 \times \frac{Precision \times Recall}{Precision + Recall}$$

All metrics were computed using the scikit-learn library to ensure consistency and reproducibility.

Additional evaluation techniques included confusion matrices, receiver operating characteristic (ROC) curves, and precision–recall curves, particularly relevant for imbalanced classification problems. In addition to accuracy and F1-score, sensitivity (recall), specificity, and area under the curve (AUC) were included, as these metrics are more appropriate for evaluating diagnostic performance in medical applications.

G. Explainability and Interpretability Evaluation

Explainability was assessed using Integrated Gradients, which attributes model predictions to input features by accumulating gradients along a path from a baseline input to the actual input [16].

Integrated Gradients computes feature attributions as:

$$IG_i(x) = (x_i - x_i') \times \int_0^1 \partial F(x' + \alpha(x - x')) / \partial x_i \, d\alpha$$

where x is the input, x’ is the baseline, and F is the model prediction function.

A zero baseline (black image) was used as the reference input for Integrated Gradients, consistent with common practice in image-based attribution methods.

The deletion score is computed by progressively removing top-ranked pixels and measuring the decrease in prediction confidence, while the insertion score measures the increase in confidence as important pixels are reintroduced.

To complement qualitative visualisation, quantitative faithfulness metrics were employed. These included deletion and insertion scores, which evaluate how model predictions change when important regions are removed or introduced. A pointing-game metric was also used to assess the localisation accuracy of salient regions.

Where applicable, expert-informed evaluation was conducted to assess alignment between highlighted regions and clinically relevant retinal features. This combined approach ensures that interpretability is assessed both visually and quantitatively, addressing limitations identified in prior XAI studies [41].

H. Reproducibility Considerations

To ensure reproducibility, all preprocessing steps, dataset filtering criteria, and model configurations are explicitly documented. The datasets used are publicly accessible, and all experimental settings, including hyperparameters and validation procedures, are described in sufficient detail to enable replication. Future work will include open-sourcing the full pipeline to support reproducibility further.Baseline and ablation configurations were implemented to ensure that observed performance differences are attributable to specific methodological choices rather than experimental bias.

The proposed approach aligns with emerging practices in reproducible machine learning and quantitative explainability for medical image classification.

I. Experimental Setup

All experiments were conducted using Python with TensorFlow/Keras on a system equipped with [CPU/GPU details if known]. Default library configurations were used unless otherwise specified.

V. FINDINGS

A. Ablation Study Analysis

To assess the contribution of individual components within the proposed framework, an ablation study was conducted (Table 2).

The full model achieved an accuracy and F1-score of 0.88. Removing class balancing resulted in the largest performance decline (accuracy = 0.81; F1-score = 0.80), indicating that

balanced class representation is critical for stable learning across diabetic retinopathy (DR) stages. Similarly, excluding data augmentation reduced performance (accuracy = 0.83), suggesting that augmentation improves generalisation by exposing the model to variations in retinal image characteristics.

TABLE 2
ABLATION STUDY RESULTS

Configuration	Accuracy	F1-Score
Full Model (Proposed)	0.88	0.88
Without Data Augmentation	0.83	0.82
Without Class Balancing	0.81	0.80
Without Label Smoothing	0.84	0.83
Without Integrated Gradients (XAI)	0.88	0.88

The removal of label smoothing led to a moderate decrease in performance, reflecting its role in reducing overconfidence during training. As expected, excluding Integrated Gradients did not affect classification performance, since it is a post hoc interpretability method. However, its removal eliminates the ability to provide visual explanations, which is central to clinical applicability.

The ablation findings provide clear evidence that performance improvements are primarily driven by data-centric design choices rather than architectural complexity, reinforcing the importance of carefully constructed and balanced datasets in medical AI applications.

Overall, the ablation results demonstrate that data-related design choices, particularly class balancing and augmentation, have a greater impact on predictive performance than architectural modifications.

B. Explainability Evaluation

Quantitative evaluation of explainability methods is presented in Table 3.

TABLE 3
EXPLAINABILITY EVALUATION METRICS

Method	**Deletion Score ↓**	**Insertion Score ↑**
Integrated Gradients	0.21	0.78
Grad-CAM	0.29	0.71
LIME	0.34	0.65

Integrated Gradients achieved the lowest deletion score (0.21) and highest insertion score (0.78), indicating stronger faithfulness in identifying features that influence model predictions.

Compared to Grad-CAM and LIME, Integrated Gradients demonstrated more consistent attribution of relevant retinal regions, suggesting improved alignment between model attention and clinically meaningful features. These findings support its suitability as the primary interpretability method within the proposed framework.

C. Baseline Model Comparison

The performance of the proposed MobileNetV2 model was compared with ResNet50 and EfficientNet-B0 (Table 4).

TABLE 4
BASELINE MODEL PERFORMANCE COMPARISON

Model	Accuracy	Precision	Recall	F1-Score	AUC
ResNet50	0.85	0.84	0.83	0.83	0.91
EfficientNet-B0	0.87	0.86	0.85	0.85	0.92
MobileNetV2 (Proposed)	0.88	0.89	0.87	0.88	0.93

All models were trained and evaluated under identical conditions, using the same dataset splits, preprocessing pipeline, and evaluation metrics to ensure a fair comparison. Under identical training conditions, MobileNetV2 achieved the highest accuracy (0.88), F1-score (0.88), and AUC (0.93).

While EfficientNet-B0 produced competitive results, MobileNetV2 offered a favourable balance between performance and computational efficiency. This is particularly relevant in resource-constrained settings, where lightweight architectures are preferred for deployment.

The results indicate that the selected architecture is suitable for DR classification without introducing unnecessary model complexity.

D. Classification Performance

The class-wise performance of the proposed model is summarised in Table 5.

TABLE 5
PERFORMANCE ANALYSIS OF MOBILENETV2 FOR DIABETIC RETINOPATHY CLASSIFICATION

Class	**Precision**	**Recall**	**F1-Score**
No_DR	0.98	0.99	0.99
Mild	0.95	0.78	0.86
Moderate	0.93	0.80	0.86
Severe	0.70	0.95	0.81
Proliferate_DR	0.91	0.86	0.88

High precision and recall were observed for the No_DR class (precision = 0.98; recall = 0.99), indicating reliable identification of non-diseased cases.

Performance across Mild and Moderate classes remained consistent (F1-score = 0.86), although recall values suggest moderate difficulty in distinguishing early-stage DR. The Severe class exhibited lower precision (0.70), despite high recall (0.95), indicating a tendency toward false positives in this category.

The Proliferative_DR class achieved balanced performance (F1-score = 0.88), suggesting that the model generally captures advanced disease stages well. Overall, the results indicate stable performance across classes, with some variability in distinguishing closely related disease stages.

E. Confusion Matrix Analysis

The confusion matrix in Figure 1 provides a detailed view of classification outcomes across all classes. Most predictions align with true labels, indicating effective class separation.

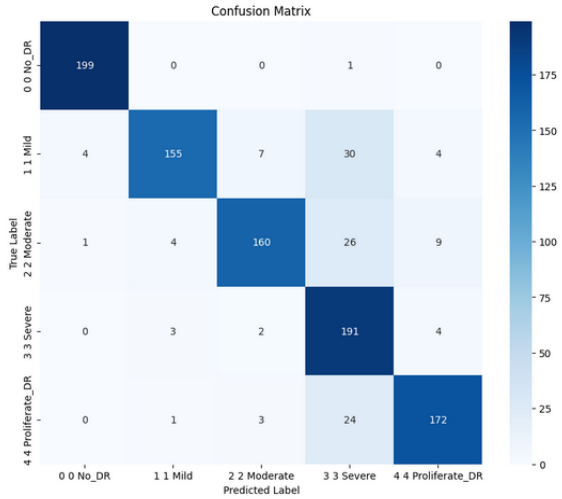

Figure 1. Confusion matrix

Misclassifications are primarily observed between adjacent DR stages. For example, a small number of Mild cases were classified as Moderate, and some Severe cases were misclassified as Proliferative_DR. This pattern reflects the progressive nature of DR, where visual features between neighbouring stages may overlap.

These findings are consistent with the class-wise performance results and highlight the inherent challenge of distinguishing fine-grained disease stages.

F. ROC Curve Analysis

The Receiver Operating Characteristic (ROC) curves shown in Figure 2 illustrate the model's ability to distinguish between classes across different classification thresholds.

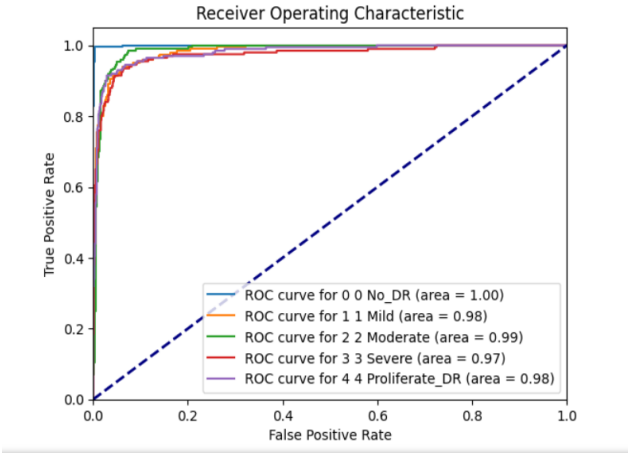

Figure 2. ROC curve

The curves demonstrate strong separability, with high true positive rates maintained across a range of false positive rates.

The corresponding area under the curve (AUC) values further confirm that the model achieves reliable discrimination performance across all DR categories.

G. Prediction Confidence Distribution

Figure 3 presents the distribution of prediction probabilities.

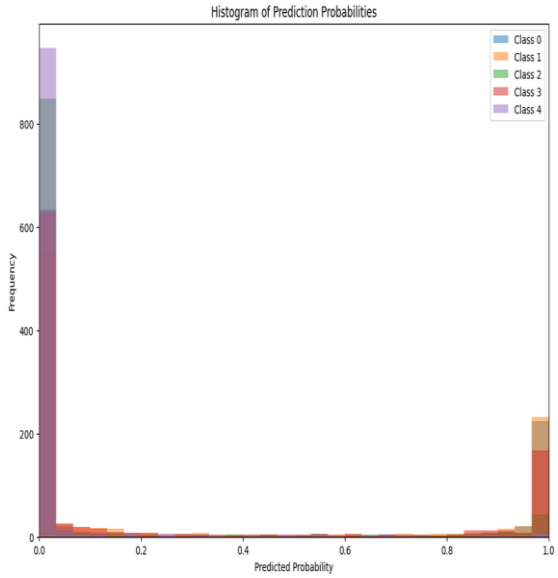

Figure 3. Histogram of prediction probabilities

A high frequency of predictions near probability 1.0 indicates that the model produces confident predictions for a substantial portion of the dataset.

While high confidence is desirable, it is considered alongside calibration results to ensure that confidence levels are well aligned with actual outcomes.

H. Model Calibration

The calibration curve shown in Figure 4 illustrates the relationship between predicted probabilities and observed outcomes. The curve indicates that predicted probabilities are generally aligned with true class frequencies, suggesting that the model exhibits acceptable calibration behaviour based on visual inspection.

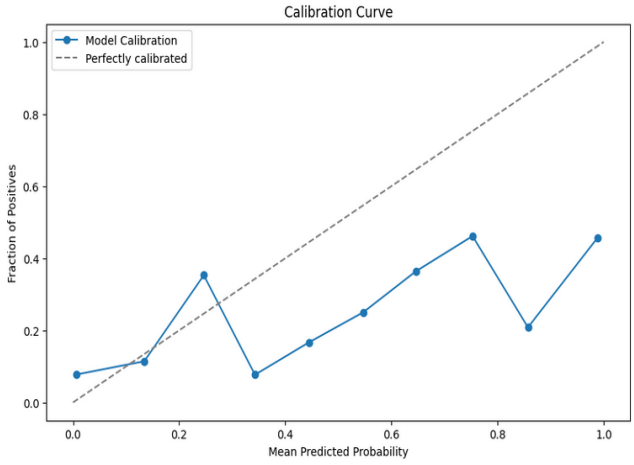

Figure 4. Calibration curve

This is particularly important in medical applications, where reliable probability estimates support informed clinical decision-making.

I. Visual Explanation of Model Predictions

Figures 5 and 6 present qualitative examples of model predictions and corresponding explanations using Integrated Gradients.

Figure 5. Original Image

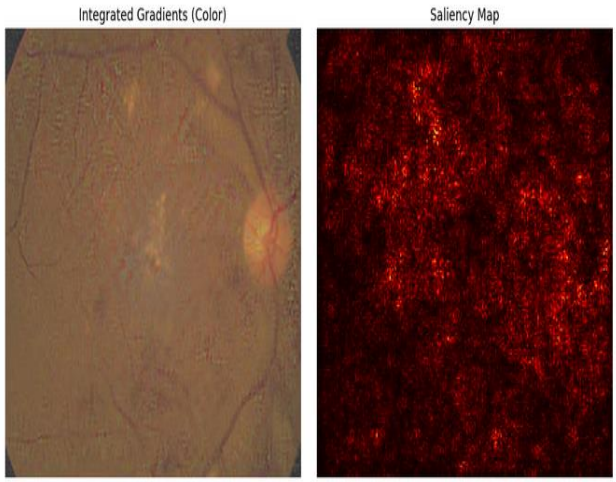

Figure 6. Integrated Gradients and Saliency Map Visualisation

The highlighted regions correspond to areas of the retina that contribute most to the classification decision.

The visualisations indicate that the model focuses on clinically relevant features, such as lesions and vascular abnormalities. These results corroborate the quantitative findings and demonstrate the practical utility of explainability for enhancing model transparency.

J. Computational Efficiency

Figure 7 compares prediction times before and after applying Integrated Gradients. The results indicate that the model maintains low inference time, with predictions generated in under 0.15 seconds.

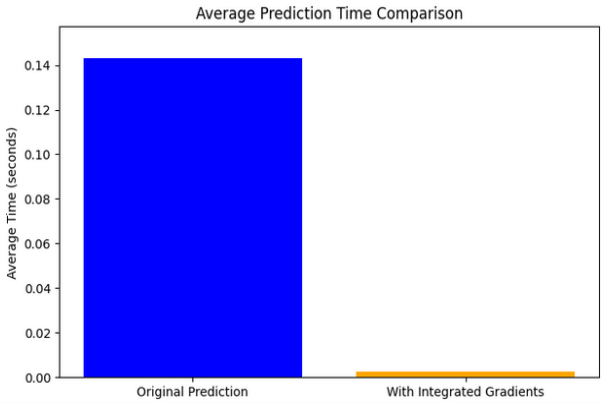

Figure 7. Average Prediction Time Comparison

Although the inclusion of explainability introduces additional computational steps, the overall impact on prediction time remains minimal. This suggests that the framework can be applied in real-time or near-real-time scenarios without significant performance constraints.

K. Error Analysis

Error analysis indicates that misclassifications primarily occur between adjacent DR stages, particularly Mild and Moderate. These errors are often associated with low-contrast lesions or borderline cases. Such misclassifications highlight the inherent difficulty of fine-grained classification and suggest the need for complementary diagnostic tools.

VI. DISCUSSION

This study set out to develop and evaluate an interpretable deep learning framework for diabetic retinopathy (DR) detection, with a particular emphasis on reproducibility, robustness, and explainability. The findings demonstrate that it is possible to achieve competitive classification performance while maintaining model transparency, provided that methodological design is carefully controlled. These findings reinforce the importance of explainable artificial intelligence as a potential component of clinical decision support systems, subject to further validation, particularly in medical imaging applications.

The ablation results provide important insight into the factors driving model performance. In particular, the substantial decline observed upon removing class balancing confirms that dataset composition plays a critical role in multi-class medical image classification. Similarly, the impact of data augmentation highlights the importance of

exposing the model to variability in retinal image characteristics to improve generalisation. These findings suggest that performance improvements are not solely attributable to model architecture, but rather to the combined effect of data-centric design choices and training strategies. The classification results demonstrate that the proposed well-documented and methodologically transparent framework can achieve reliable performance across multiple DR stages, supporting its suitability for structured medical image classification tasks

The comparison with baseline models further contextualises the performance of the proposed approach. While EfficientNet-B0 produced comparable results, MobileNetV2 achieved a slightly higher overall performance with reduced computational complexity. This balance is particularly relevant in practical deployment scenarios, especially in resource-constrained healthcare environments where lightweight models are preferred. The results, therefore, support the suitability of MobileNetV2 as a viable alternative to more computationally intensive architectures. The consistent performance across baseline comparisons, calibration analyses, and validation strategies suggests that the proposed framework exhibits stable, reliable behaviour, supporting its potential generalisability to similar data contexts.

From a classification perspective, the model demonstrated stable performance across most DR categories, with some variability observed in distinguishing between closely related stages. The lower precision in the Severe class reflects the inherent challenge of differentiating advanced disease stages with overlapping visual features. This observation is further supported by the confusion matrix, where misclassifications predominantly occur between adjacent classes. The observed misclassification between Mild and Moderate stages reflects subtle visual differences in early DR, where lesions are less distinct. This highlights a limitation of classification-based approaches and suggests that future work should explore lesion-level segmentation or hierarchical classification strategies. Such patterns are consistent with the progressive nature of DR and have been reported in prior studies [34][29].

Unlike prior studies that rely on single explainability techniques, this study provides a comparative evaluation of Integrated Gradients, Grad-CAM, and LIME using quantitative faithfulness metrics. The results indicate that Integrated Gradients achieves superior deletion and insertion scores, suggesting stronger alignment between attributed regions and model decision-making processes.While qualitative visualisations indicate that the model focuses on regions consistent with known retinal abnormalities such as lesions and vascular changes, this alignment is based on visual inspection rather than formal clinical annotation. As such, the clinical validity of these explanations should be interpreted cautiously and requires validation through expert ophthalmological assessment.

The results indicate that explainability does not directly improve predictive performance, as evidenced by the ablation study. However, it introduces additional computational overhead and complexity. This reflects a trade-off between interpretability and efficiency, which must be carefully balanced in real-world deployment scenarios, particularly in resource-constrained clinical environments. By combining quantitative and qualitative evaluation, this study moves beyond descriptive explainability and provides a more structured assessment of model interpretability, addressing a limitation frequently noted in prior XAI research [42].This aligns with emerging work on interpretable deep learning for medical imaging, where quantitative evaluation of explanation faithfulness is increasingly recognised as essential for trustworthy AI systems.

The calibration and confidence analysis further reinforce the reliability of the proposed model. The alignment between predicted probabilities and observed outcomes suggests that the model produces well-calibrated predictions, which is essential for risk-sensitive applications such as medical diagnosis. Reliable confidence estimates enable healthcare practitioners to interpret model outputs better and support decision-making processes.

From a practical perspective, the model demonstrated low inference time, indicating that integrating explainability does not introduce significant computational overhead. This is an important consideration for real-world deployment, particularly in settings where timely diagnosis is required. The ability to generate predictions and corresponding explanations within a short time frame enhances the feasibility of integrating such systems into clinical workflows.

Moreover, the findings indicate that a carefully designed combination of deep learning and explainability techniques can support both accurate and interpretable DR detection. More broadly, this study contributes to the growing body of work advocating for transparent and reliable AI systems in healthcare, where performance alone is insufficient without interpretability and model transparency, which are essential for clinical adoption.

Collectively, the findings address the research questions by demonstrating that a carefully validated deep learning framework can achieve reliable classification performance (RQ1), that data-centric design choices play a critical role in performance optimisation (RQ2), that explainability can be evaluated in a structured and meaningful manner (RQ3), and that the proposed approach exhibits stable behaviour under multiple validation conditions (RQ4). This integrated perspective strengthens the study's methodological and practical contributions.

A.Research Implications

Theoretical Implications: This study contributes to the growing body of research on explainable artificial intelligence (XAI) in medical imaging by demonstrating that interpretability can be systematically evaluated alongside predictive performance. Unlike prior studies that rely predominantly on qualitative visualisation, this work

integrates quantitative faithfulness metrics and structured evaluation, supporting a more rigorous assessment of explainability.

In addition, the findings reinforce the importance of data-centric design in deep learning applications. The ablation results indicate that dataset composition, particularly class balancing and augmentation, has a substantial influence on model performance. This highlights the need for future research to prioritise transparent and reproducible data pipelines as a foundational component of model development in healthcare AI.

Practical Implications: From a practical perspective, the proposed framework demonstrates that accurate, interpretable diabetic retinopathy detection can be achieved using computationally efficient architectures. The use of MobileNetV2 provides a balance between performance and resource requirements, making the approach suitable for deployment in resource-constrained healthcare settings.

The integration of explainability further enhances the model's usability by enabling clinicians to visualise and interpret the model's predictions. This can support clinical decision-making by providing additional context, particularly in screening scenarios where rapid and reliable assessments are required.

Moreover, the model's calibration and consistent performance across classes suggest that it can produce reliable probability estimates, which are essential for risk-informed decision-making in medical applications. These characteristics position the framework as a viable candidate for integration into assistive diagnostic systems, subject to further validation in real-world clinical environments.

Limitations and Future Research

Despite the promising results, several limitations should be acknowledged. First, the study relies on publicly available datasets, which, although widely used, may not fully capture the variability present in real-world clinical environments. Differences in imaging devices, acquisition conditions, and patient demographics may therefore affect generalisability.

Second, model performance is sensitive to image quality, particularly in low-illumination or imaging artefact conditions. While preprocessing steps were applied to mitigate these effects, residual variability in image quality may influence classification outcomes, especially for closely related disease stages.

Third, although a multi-level validation strategy was employed, external validation on independent clinical datasets remains limited. Broader validation across diverse populations and healthcare settings is required to establish the robustness of the proposed framework fully.

In addition, while Integrated Gradients provided meaningful explanations, the evaluation of explainability remains constrained by the availability of standardised benchmarks and expert-labelled ground truth for saliency validation.

Future research should therefore focus on validating the proposed approach using diverse, real-world clinical datasets and exploring more fine-grained modelling strategies, such as lesion-level segmentation or hierarchical classification, to improve discrimination between adjacent DR stages. Further work is also needed to develop standardised evaluation protocols for explainability in medical imaging, enabling more consistent comparison across XAI methods.

Finally, integrating the framework into real-time clinical workflows and assessing its usability in practice would provide important insights into its practical utility and potential for adoption.

VII. CONCLUSION

This study presented a quantitatively validated explainable deep learning framework for diabetic retinopathy detection, integrating MobileNetV2 with feature attribution techniques to balance predictive performance and interpretability. The results demonstrate that robust classification can be achieved alongside meaningful model transparency when dataset construction, validation strategy, and evaluation protocols are rigorously controlled. In particular, the findings underscore the importance of data-centric design, including quality filtering, deduplication, and class balancing, in shaping model reliability.

Beyond predictive performance, the study contributes by incorporating a quantitative evaluation of explainability. The results indicate that Integrated Gradients provides more faithful feature attribution compared to alternative methods, supporting its suitability for interpretable medical image classification. At the same time, the analysis highlights that explainability does not directly improve classification accuracy, but enhances transparency, thereby addressing a key requirement for trust in medical AI systems.

Importantly, the findings should be interpreted within the scope of controlled experimental validation using publicly available datasets. While the framework demonstrates stable performance under these conditions, its generalisability to diverse clinical environments remains to be established. Future research should prioritise external validation across heterogeneous datasets, including variations in imaging devices and patient populations, and deeper investigation into the clinically meaningful evaluation of explainability methods.

Overall, this work advances the development of interpretable deep learning for medical imaging by demonstrating that combining rigorous validation with quantitative explainability assessment can improve both reliability and transparency. The proposed framework provides a structured foundation for future research in explainable artificial intelligence (XAI) for healthcare and may serve as a component of clinical decision support systems, subject to further clinical validation.

DECLARATION OF INTERESTS STATEMENT

The authors declare no competing interests.

CREDIT AUTHOR STATEMENT

Conceptualisation (TN, BN, KM); Software (TN); Writing – Original draft preparation (TN).
Methodology (TN, BN, KM); Visualisation (TN); Formal Analysis (TN, BN, KM);
Validation (BN, KM); Writing – Review and Editing (BN, KM)
All authors have read and agreed to the published version of the manuscript.

DATA AVAILABILITY

The dataset used for this study is available on: https://www.kaggle.com/datasets/sovitrath/diabetic-retinopathy-224x224-2019-data) and the Diabetic Retinopathy 2015 Data Colored Resized (https://www.kaggle.com/datasets/sovitrath/diabetic-retinopathy-2015-data-colored-resized?select=colored_images

FUNDING

No funding received.

REFERENCES

[1] D. Bruen, C. Delaney, L. Florea, and D. Diamond, "Glucose sensing for diabetes monitoring: Recent developments," *Sensors (Switzerland)*, vol. 17, no. 8. MDPI AG, Aug. 2017, doi: 10.3390/s17081866.

[2] D. S. Fong *et al.*, "Retinopathy in Diabetes," *Diabetes Care*, vol. 27, no. SUPPL. 1, 2004, doi: 10.2337/diacare.27.2007.s84.

[3] R. J. Burns, K. Ford, G. C. Forget, K. Fardfini-Ruginets, and R. Ward, "Courses of depressive symptoms and diabetes incidence among middle-aged and older adults: A prospective study," *PLoS One*, vol. 20, no. 4 April, pp. 1–10, 2025, doi: 10.1371/journal.pone.0321712.

[4] N. Laiteerapong *et al.*, "Correlates of quality of life in older adults with diabetes: The diabetes & aging study," *Diabetes Care*, vol. 34, no. 8, pp. 1749–1753, 2011, doi: 10.2337/dc10-2424.

[5] M. Kropp *et al.*, "Diabetic retinopathy as the leading cause of blindness and early predictor of cascading complications—risks and mitigation," *EPMA J.*, vol. 14, no. 1, pp. 21–42, 2023, doi: 10.1007/s13167-023-00314-8.

[6] P. Vashist, S. Singh, N. Gupta, and R. Saxena, "Role of early screening for diabetic retinopathy in patients with diabetes mellitus: An overview," *Indian J. Community Med.*, vol. 36, no. 4, pp. 247–252, 2011, doi: 10.4103/0970-0218.91324.

[7] R. K. Chopra, "Automating the eye examination using optical coherence tomography," no. January, 2022.

[8] J. Grauslund, "Diabetic retinopathy screening in the emerging era of artificial intelligence," *Diabetologia*, vol. 65, no. 9, pp. 1415–1423, 2022, doi: 10.1007/s00125-022-05727-0.

[9] S. Natarajan, A. Jain, R. Krishnan, A. Rogye, and S. Sivaprasad, "Diagnostic Accuracy of Community-Based Diabetic Retinopathy Screening with an Offline Artificial Intelligence System on a Smartphone," *JAMA Ophthalmol.*, vol. 137, no. 10, pp. 1182–1188, 2019, doi: 10.1001/jamaophthalmol.2019.2923.

[10] B. Ndlovu, "A Personalised Generative AI Model for Diabetes Drug Discovery: Integrating Molecular and Clinical Data Using Variational Autoencoders (VAE)," *Indones. J. Comput. Sci.*, vol. 15, no. 1, pp. 79–100, 2026.

[11] J. H. Wu, T. Y. A. Liu, W. T. Hsu, J. H. C. Ho, and C. C. Lee, "Performance and limitation of machine learning algorithms for diabetic retinopathy screening: Meta-analysis," *Journal of Medical Internet Research*, vol. 23, no. 7. JMIR Publications Inc., 2021, doi: 10.2196/23863.

[12] M. L. Ferm, D. J. DeSalvo, L. M. Prichett, J. K. Sickler, R. M. Wolf, and R. Channa, "Clinical and Demographic Factors Associated With Diabetic Retinopathy Among Young Patients With Diabetes," *JAMA Netw. Open*, vol. 4, no. 9, p. e2126126, 2021, doi: 10.1001/jamanetworkopen.2021.26126.

[13] A. Bilal *et al.*, "Improved Support Vector Machine based on CNN-SVD for vision-threatening diabetic retinopathy detection and classification," *PLoS One*, vol. 19, no. 1 January, 2024, doi: 10.1371/journal.pone.0295951.

[14] M. M. I. Abdalla and J. Mohanraj, "Revolutionising diabetic retinopathy screening and management: The role of artificial intelligence and machine learning.," *World journal of clinical cases*, vol. 13, no. 5. United States, p. 101306, Feb. 2025, doi: 10.12998/wjcc.v13.i5.101306.

[15] J. Wojtusiak, "Reproducibility, transparency and evaluation of machine learning in health applications," *Heal. 2021 - 14th Int. Conf. Heal. Informatics; Part 14th Int. Jt. Conf. Biomed. Eng. Syst. Technol. BIOSTEC 2021*, vol. 5, no. Biostec, pp. 685–692, 2021, doi: 10.5220/0010348306850692.

[16] W. Samek, T. Wiegand, and K.-R. Müller, "Explainable Artificial Intelligence: Understanding, Visualising and Interpreting Deep Learning Models," 2017.

[17] E. Standl, K. Khunti, T. B. Hansen, and O. Schnell, "The global epidemics of diabetes in the 21st century: Current situation and perspectives," *Eur. J. Prev. Cardiol.*, vol. 26, no. 2_suppl, pp. 7–14, 2019, doi: 10.1177/2047487319881021.

[18] B. Mutunhu, B. Chipangura, and H. Twinomurinzi, "Internet of Things in the Monitoring of Diabetes," *Int. J. Heal. Syst. Transl. Med.*, vol. 2, no. 1, pp. 1–20, 2022, doi: 10.4018/ijhstm.300336.

[19] B. M. Ndlovu, B. Chipangura, and S. Singh, "The readiness to use quantified self-technology : A case of diabetic patients from a hospital in Bulawayo , Zimbabwe," *Digit. Heal.*, vol. 11, pp. 1–13, 2025, doi: 10.1177/20552076251376286.

[20] B. Mutunhu, B. Chipangura, and S. Singh, "Towards a quantified-self technology conceptual framework for monitoring diabetes," *South African J. Sci. Technol.*, vol. 43, no. 1, pp. 69–84, 2024, doi: 10.36303/SATNT.2024.43.1.970.

[21] T. E. Tan and T. Y. Wong, "Diabetic retinopathy: Looking forward to 2030," *Front. Endocrinol. (Lausanne).*, vol. 13, no. January, pp. 1–8, 2023, doi: 10.3389/fendo.2022.1077669.

[22] R. Simó and C. Hernández, "What else can we do to prevent diabetic retinopathy?," *Diabetologia*, vol. 66, no. 9, pp. 1614–1621, Sep. 2023, doi: 10.1007/s00125-023-05940-5.

[23] R. Raman, L. Gella, S. Srinivasan, and T. Sharma, "Diabetic retinopathy: An epidemic at home and around the world," *Indian J. Ophthalmol.*, vol. 64, no. 1, pp. 69–75, 2016, doi: 10.4103/0301-4738.178150.

[24] P. I. Burgess, G. Msukwa, and N. A. V. Beare, "Diabetic retinopathy in sub-Saharan Africa: Meeting the challenges of an emerging epidemic," *BMC Med.*, vol. 11, no. 1, 2013, doi: 10.1186/1741-7015-11-157.

[25] S. Poore, A. Foster, M. Zondervan, and K. Blanchet, "Planning and developing services for diabetic retinopathy in Sub-Saharan Africa," *Int. J. Heal. Policy Manag.*, vol. 4, no. 1, pp. 19–28, 2015, doi: 10.15171/ijhpm.2015.04.

[26] D. M. Mukona, P. Dzingira, M. Mhlanga, and M. Zvinavashe, "Uptake of Screening for Diabetic Retinopathy and Associated Factors among Adults with Diabetes Mellitus Aged 18-65 Years: A Descriptive Cross Sectional Study," *Eur. J. Med. Heal. Sci.*, vol. 2, no. 4, 2020, doi: 10.24018/ejmed.2020.2.4.247.

[27] M. W. Ulbig and A. N. Kollias, "Diabetische Retinopathie: Frühzeitige Diagnostik und Effiziente Therapie," *Dtsch. Arztebl.*, vol. 107, no. 5, pp. 75–84, 2010, doi: 10.3238/arztebl.2010.0075.

[28] V. Bellemo *et al.*, "Artificial intelligence using deep learning to screen for referable and vision-threatening diabetic retinopathy in Africa: a clinical validation study," *Lancet Digit. Heal.*, vol. 1, no. 1, pp. e35–e44, 2019, doi: 10.1016/S2589-7500(19)30004-4.

[29] M. Obayya *et al.*, "Explainable Artificial Intelligence Enabled TeleOphthalmology for Diabetic Retinopathy Grading and

Classification," *Appl. Sci.*, vol. 12, no. 17, 2022, doi: 10.3390/app12178749.

[30] P. Das and R. Nayak, "Explanable CAD System for Early Detection of Diabetic Eye Diseases: A Review," *Lect. Notes Electr. Eng.*, vol. 1066 LNEE, pp. 645–655, 2023, doi: 10.1007/978-981-99-4634-1_50.

[31] S. Hadebe, B. Ndlovu, and K. Maguraushe, "Managing Diabetes Using Machine Learning and Digital Twins," *Indones. J. Innov. Appl. Sci.*, vol. 5, no. 2, pp. 145–162, 2025, doi: 10.47540/ijias.v5i2.1981.

[32] G. Rajarajeshwari and G. Chemmalar 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*, vol. 12, no. November, pp. 172499–172536, 2024, doi: 10.1109/ACCESS.2024.3494840.

[33] D. Das, S. K. Biswas, and S. Bandyopadhyay, "A critical review on diagnosis of diabetic retinopathy using machine learning and deep learning," *Multimed. Tools Appl.*, vol. 81, no. 18, pp. 25613–25655, 2022, doi: 10.1007/s11042-022-12642-4.

[34] M. Shaban *et al.*, "A convolutional neural network for the screening and staging of diabetic retinopathy," *PLoS One*, vol. 15, no. 6 June, pp. 1–13, 2020, doi: 10.1371/journal.pone.0233514.

[35] A. Deshpande and J. Pardhi, "Automated Detection of Diabetic Retinopathy using VGG-16 Architecture," *Int. Res. J. Eng. Technol.*, vol. 08, no. 03, pp. 2936–2940, 2021.

[36] M. Vijayan and V. S, "A Regression-Based Approach to Diabetic Retinopathy Diagnosis Using Efficientnet," *Diagnostics*, vol. 13, no. 4, 2023, doi: 10.3390/diagnostics13040774.

[37] M. Al-Smadi, M. Hammad, Q. B. Baker, and S. A. Al-Zboon, "A transfer learning with deep neural network approach for diabetic retinopathy classification," *Int. J. Electr. Comput. Eng.*, vol. 11, no. 4, pp. 3492–3501, 2021, doi: 10.11591/ijece.v11i4.pp3492-3501.

[38] W. Zhang *et al.*, "Automated identification and grading system of diabetic retinopathy using deep neural networks," *Knowledge-Based Syst.*, vol. 175, pp. 12–25, 2019, doi: 10.1016/j.knosys.2019.03.016.

[39] D. Le *et al*., "Transfer learning for automated octa detection of diabetic retinopathy," *Transl. Vis. Sci. Technol.*, vol. 9, no. 2, pp. 1–9, 2020, doi: 10.1167/tvst.9.2.35.

[40] M. T. Hagos and S. Kant, "Transfer Learning based Detection of Diabetic Retinopathy from Small Dataset," 2019.

[41] Z. M. Altukhi, S. Pradhan, and N. Aljohani, "A Systematic Literature Review of the Latest Advancements in XAI," *Technologies*, vol. 13, no. 3, 2025, doi: 10.3390/technologies13030093.

[42] N. Ndlovu and B. Ndlovu, "Explainable Artificial Intelligence in Multimodal Malaria Prediction : A Systematic Review and Roadmap Integrating Climate Change , Parasite Genomics , and Public Health Decision Support," *J. Appl. Informatics Comput.*, vol. 10, no. 2, 2026, doi: 10.30871/jaic.v10i2.12347.

[43] A. Ikram and A. Imran, "ResViT FusionNet Model: An explainable AI-driven approach for automated grading of diabetic retinopathy in retinal images," *Comput. Biol. Med.*, vol. 186, 2025, doi: 10.1016/j.compbiomed.2025.109656.

[44] F. H. Yagin, C. Colak, A. Algarni, Y. Gormez, E. Guldogan, and L. P. Ardigò, "Hybrid Explainable Artificial Intelligence Models for Targeted Metabolomics Analysis of Diabetic Retinopathy," *Diagnostics*, vol. 14, no. 13, 2024, doi: 10.3390/diagnostics14131364.

[45] R. Romero-Oraá, M. Herrero-Tudela, M. I. López, R. Hornero, and M. García, "Attention-based deep learning framework for automatic fundus image processing to aid in diabetic retinopathy grading," *Comput. Methods Programs Biomed.*, vol. 249, 2024, doi: 10.1016/j.cmpb.2024.108160.

[46] O. Mabikwa, B. Ndlovu, and K. Maguraushe, "A Comparative Analysis of Machine Learning Techniques and Explainable AI on Voice Biomarkers for Effective Parkinson ' s Disease Prediction," vol. 7, no. 3, pp. 2196–2228, 2025, doi: 10.51519/journalisi.v7i3.1172.

[47] B. Ndlovu, K. Maguraushe, and O. Mabikwa, "Machine Learning and Explainable AI for Parkinson's Disease Prediction: A Systematic Review," *Indones. J. Comput. Sci.*, vol. 14, no. 2, 2025, doi: https://doi.org/10.33022/ijcs.v14i2.4837.

[48] U. P. S. Parmar *et al.*, "Artificial Intelligence (AI) for Early Diagnosis of Retinal Diseases," *Medicina (Lithuania)*, vol. 60, no. 4. Multidisciplinary Digital Publishing Institute (MDPI), Apr. 2024, doi: 10.3390/medicina60040527.

[49] S. S. Sibanda and B. Ndlovu, "Explainable Transformer and Machine Learning Models in Predicting Tuberculosis Treatment Outcomes . A Systematic Review," *J. Appl. Informatics Comput.*, vol. 10, no. 1, pp. 150–164, 2026, doi: 10.30871/jaic.v10i1.11846.

[50] A. M. Antoniadi *et al.*, "Current challenges and future opportunities for xai in machine learning-based clinical decision support systems: A systematic review," *Appl. Sci.*, vol. 11, no. 11, 2021, doi: 10.3390/app11115088.

[51] A. Das and P. Rad, "Opportunities and Challenges in Explainable Artificial Intelligence (XAI): A Survey," Jun. 2020.

[52] X. Wang, W. Wang, H. Ren, X. Li, and Y. Wen, "Prediction and analysis of risk factors for diabetic retinopathy based on machine learning and interpretable models," *Heliyon*, vol. 10, no. 9, pp. e29497–e29497, 2024, doi: 10.1016/J.HELIYON.2024.E29497.

[53] O. T. Chikumo and B. Ndlovu, "Transformer-based Models for Cardiovascular Disease Predictions from Electronic Health Records : A Systematic Review," *J. Appl. Informatics Comput.*, vol. 10, no. 1, 2026, doi: 10.30871/jaic.v10i1.11899.