2025 Volume: 01 Issue: 07 www.ijrpa.com ISSN 2456-9995 Research Article

International Journal Research Publication Analysis

Page: 01-16

DIABETIC RETINOPATHY DETECTION USING EXPLAINABLE AI THROUGH GRAD-CAM VISUALIZATION

***Varshitha T L.,**

India.

Article Received: 07 November 2025
Article Revised: 27 November 2025
Published on: 17 December 2025

***Corresponding Author: Varshitha T L.,**
India. DOI: https://doi-doi.org/101555/ijrpa.9118

ABSTRACT

Diabetic Retinopathy (DR) is one of the leading causes of vision impairment and blindness among diabetic patients worldwide. Early detection and timely intervention are crucial to prevent irreversible vision loss. This research presents an AI-based Diabetic Retinopathy detection system using deep learning combined with Explainable Artificial Intelligence (XAI) techniques. A Convolutional Neural Network (CNN) based on the ResNet50 architecture is employed to automatically classify retinal fundus images into different DR stages. To enhance model transparency and clinical trust, Gradient-weighted Class Activation Mapping (Grad-CAM) is integrated to visually highlight pathological regions influencing model predictions. The proposed system is deployed through an interactive Gradio-based web interface, enabling real-time predictions and visual explanations. Experimental results demonstrate high classification accuracy along with meaningful visual interpretations, making the system suitable for assisting ophthalmologists in clinical decisionmaking.

INDEXTERMS: Diabetic Retinopathy, Deep Learning, ResNet50, Explainable AI, Grad-CAM, Medical Image Analysis.

INTRODUCT ION

Diabetic Retinopathy is a microvascular complication of diabetes that damages retinal blood vessels due to prolonged hyperglycemia. If left undiagnosed, it can lead to severe vision loss or permanent blindness. Manual screening of retinal fundus images is time-consuming, subjective, and requires trained ophthalmologists, making large-scale screening difficult, especially in resourcelimited regions. Recent advances in deep learning have shown

promising results in automated DR detection. However, the lack of interpretability in deep learning models limits their adoption in medical applications. This work addresses this challenge by combining a high-performance CNN model with explainable AI techniques to provide both accurate predictions and visual explanations.

Motivation The motivation behind this work is to develop an automated, reliable, and interpretable DR detection system that can support clinicians in early diagnosis. Integrating explainability ensures transparency, increases clinical trust, and aids in understanding model decisions.

OBJECTIVES

The objectives of this research are as follows:

To develop an automated diabetic retinopathy detection system using deep learning techniques.

To utilize the ResNet50 architecture with transfer learning for robust feature extraction from retinal fundus images.

To incorporate Grad-CAM to generate visual explanations highlighting diseaseaffected retinal regions.

To evaluate the proposed system using standard classification performance metrics.

To deploy the trained model using a web-based interface for real-time and userfriendly interaction.

KEY CONTRIBUTIONS

The major contributions of this work include:

- Design of a deep learning-based diabetic retinopathy detection framework using ResNet50.
- Integration of Explainable AI through Grad-CAM for transparent and interpretable predictions.
- Visualization of clinically relevant retinal regions influencing model decisions.
- Deployment of the system using a Gradio-based web interface for practical usage.
- Comprehensive experimental evaluation demonstrating both performance and explainability.

RELATED WORK

Automated detection of diabetic retinopathy has been an active area of research for several

decades, with methodologies evolving significantly alongside advancements in machine learning and computer vision. Early approaches to diabetic retinopathy detection primarily relied on traditional image processing techniques combined with classical machine learning classifiers. These methods focused on extracting handcrafted features related to retinal abnormalities such as microaneurysms, hemorrhages, hard exudates, and soft exudates. Feature extraction techniques included edge detection, morphological operations, texture analysis, and color-based segmentation, followed by classification using algorithms such as Support Vector Machines (SVM), k-Nearest Neighbors (k-NN), Decision Trees, and Random Forests. Although these approaches demonstrated moderate success, their performance was heavily dependent on the quality of feature engineering and lacked robustness to variations in image quality, illumination, and patient demographics.

With the emergence of deep learning, Convolutional Neural Networks have become the dominant approach for diabetic retinopathy detection. CNNs automatically learn hierarchical feature representations directly from raw retinal images, enabling improved generalization and robustness compared to handcrafted feature-based methods. Several studies have demonstrated the effectiveness of deep CNN architectures such as AlexNet, VGGNet, Inception, DenseNet, and ResNet in classifying retinal fundus images into multiple DR severity levels. Transfer learning, which involves fine-tuning models pre-trained on large-scale datasets such as ImageNet, has further enhanced performance by enabling effective feature extraction even with limited medical image datasets.

Among these architectures, ResNet-based models have gained significant attention due to their use of residual connections, which address the vanishing gradient problem and allow the training of deeper networks. ResNet50, in particular, has been widely adopted in medical image analysis tasks, including diabetic retinopathy detection, due to its balance between depth, computational efficiency, and performance. Studies employing ResNet50 have reported high classification accuracy and improved convergence compared to shallower CNN models. However, despite their strong predictive capabilities, most ResNet-based DR detection systems focus primarily on performance metrics and do not address the interpretability of model decisions.

Recent research has highlighted the importance of Explainable Artificial Intelligence in medical applications. Explainability is crucial in healthcare, where automated systems must provide transparent reasoning to support clinical decision-making. Techniques such as

saliency maps, Gradient-weighted Class Activation Mapping (Grad-CAM), Layer-wise Relevance Propagation (LRP), and SHAP have been proposed to visualize and interpret deep learning model predictions. Grad-CAM has emerged as one of the most popular explainability techniques due to its simplicity and effectiveness in highlighting classdiscriminative regions in images. When applied to retinal fundus images, GradCAM can localize regions associated with diabetic retinopathy lesions, offering valuable insights to clinicians.

Several studies have incorporated Grad-CAM into diabetic retinopathy detection frameworks to visualize model attention. These works demonstrate that explainability can improve clinician confidence and assist in validating model predictions. However, many existing studies remain limited to experimental analysis and do not provide deployable systems suitable for realworld clinical use.

Additionally, some approaches generate explanations without integrating them seamlessly into user- friendly interfaces, reducing their practical applicability.

In contrast to existing literature, the proposed work focuses on developing a complete and deployable diabetic retinopathy detection system that combines deep learning-based classification with explainability and real-time interaction. By integrating a ResNet50-based CNN with Grad-CAM visualization and deploying the system through a Gradio-based web interface, this research addresses both performance and usability. The proposed approach aims to bridge the gap between high-performing AI models and clinically acceptable, interpretable diagnostic tools.

SYSTEM ARCHITECTURE

The proposed diabetic retinopathy detection framework is designed using a modular and structured architecture to ensure efficiency, scalability, and interpretability. The system architecture consists of multiple interconnected components, including image acquisition, preprocessing, deep learning- based feature extraction and classification, explainability through Grad-CAM, and deployment via a web-based interface. Each component plays a critical role in the end-to-end workflow of the system.

The process begins with the acquisition of retinal fundus images from the dataset. These images serve as the primary input to the system and represent different stages of diabetic

retinopathy. Since fundus images may vary in resolution, illumination, and quality, they are first passed through a preprocessing module. This module performs essential operations such as resizing images to a fixed input size compatible with the deep learning model, normalizing pixel intensities, and applying data augmentation techniques to enhance generalization.

Once preprocessing is complete, the images are fed into the deep learning model based on the ResNet50 architecture. The CNN extracts hierarchical feature representations from the retinal images, capturing both low-level visual patterns and high-level semantic features relevant to diabetic retinopathy classification. The extracted features are then processed by fully connected layers to generate class probabilities corresponding to different DR stages.

To address the lack of interpretability inherent in deep learning models, the explainability module is integrated into the system using Grad-CAM. Grad-CAM computes class-specific activation maps by analyzing the gradients flowing into the final convolutional layers of the network. These activation maps highlight the regions of the retina that contribute most to the predicted class, providing visual explanations for the model's decisions.

The final component of the system architecture is the deployment layer, which presents the prediction results and Grad-CAM visualizations to the user through a Gradio-based web interface. This interface enables users to upload retinal images, receive classification outputs, and view corresponding heatmaps in real time. The modular design of the architecture allows for easy extension, maintenance, and integration with future enhancements.

- Predictions and heatmaps.

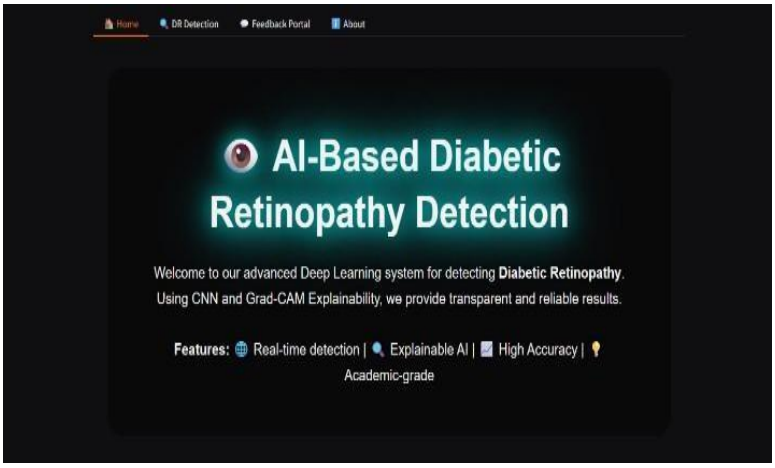

Fig 1: Home Page of the Application.

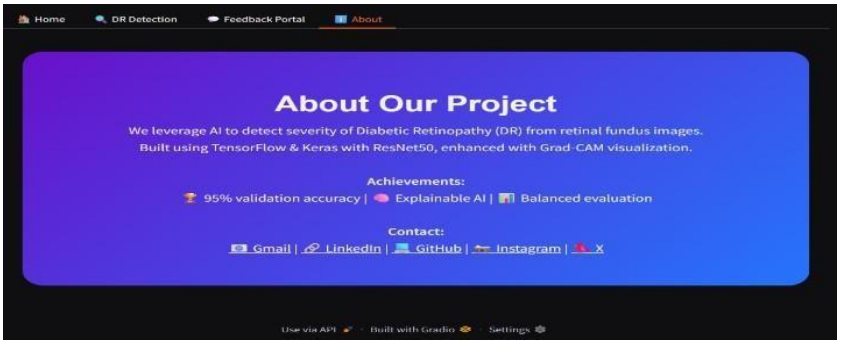

Fig 2: About Page Snapshot.

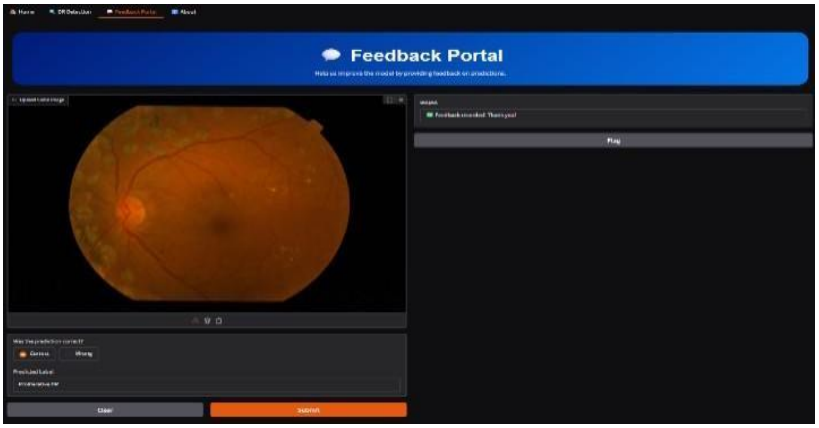

Fig 3: Feedback Page Snapshot.

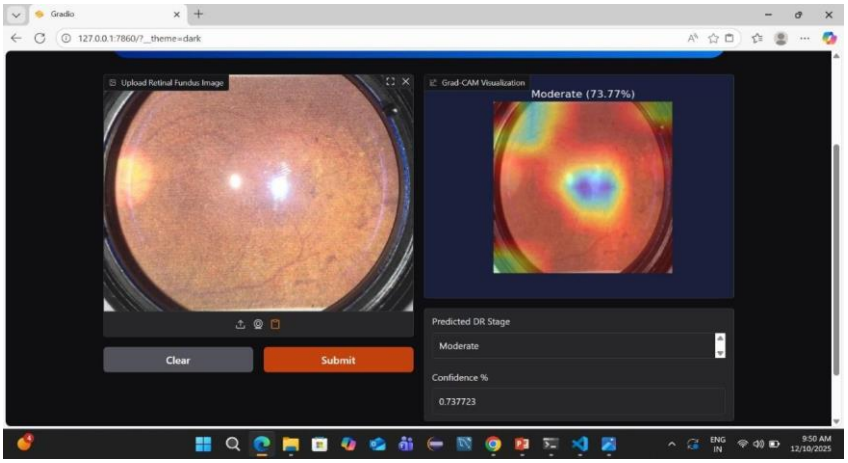

Fig 4. DR Detection Page.

DATASET DESCRIPTION AND PREPROCESSING

The dataset used in this research consists of labeled retinal fundus images representing various stages of diabetic retinopathy. Each image is associated with a ground truth label indicating the severity level of the disease. The dataset includes images captured under diverse imaging conditions, including variations in resolution, illumination, and contrast, reflecting real-world clinical scenarios.

Such variability poses challenges for automated analysis and necessitates effective preprocessing techniques.

Prior to training the deep learning model, several preprocessing steps are applied to ensure data consistency and improve classification performance. All retinal images are resized to a fixed spatial resolution to match the input requirements of the ResNet50 architecture. Pixel intensity normalization is performed to standardize the range of values across images, reducing the impact of illumination differences. Noise reduction techniques are applied where necessary to enhance image clarity.

Data augmentation is employed to increase the diversity of the training dataset and reduce overfitting. Augmentation techniques include horizontal and vertical flipping, rotation at various angles, zooming, and minor translations. These transformations help the model learn invariant features and improve generalization to unseen data. The preprocessed and augmented dataset is then divided into training, validation, and testing subsets for experimental evaluation.

II. METHODOLOGY AND MODEL ARCHITECTURE

The proposed diabetic retinopathy detection system is built upon a deep learning framework that leverages the strength of Convolutional Neural Networks for automated feature extraction and classification. The methodology is designed to ensure high classification performance while maintaining transparency through explainable artificial intelligence techniques. The overall methodology includes image preprocessing, feature extraction using a pretrained deep CNN, classification of diabetic retinopathy stages, and explainability through Grad-CAM visualization.

Design Flowchart

The flowchart below illustrates the entire project pipeline—from image input to explainable classification output.

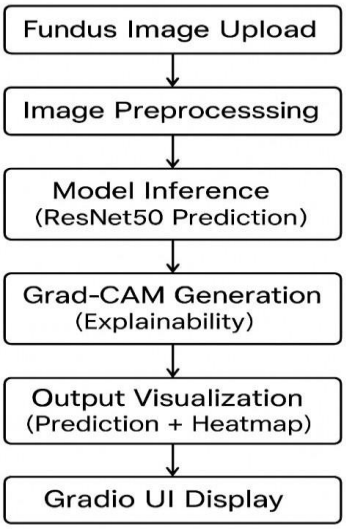

Fig 4 Flowchart of DR Detection System Design.

A. Convolutional Neural Networks for Medical Image Analysis

Convolutional Neural Networks have demonstrated exceptional performance in medical image analysis due to their ability to learn spatial hierarchies of features directly from raw pixel data. Unlike traditional machine learning approaches that require handcrafted feature extraction, CNNs automatically identify discriminative patterns such as edges, textures, and complex structures. In retinal fundus images, CNNs are particularly effective in capturing diseaserelated features such as microaneurysms, hemorrhages, and exudates, which are critical indicators of diabetic retinopathy.

A typical CNN architecture consists of convolutional layers, activation functions, pooling layers, and fully connected layers. Convolutional layers apply learnable filters to extract local features, while pooling layers reduce spatial dimensionality and improve computational efficiency. Fully connected layers perform high-level reasoning and generate class predictions. These components collectively enable CNNs to model complex visual patterns present in retinal images.

B. ResNet50 Architecture

The core of the proposed methodology is the ResNet50 architecture, a deep residual network consisting of 50 layers. ResNet introduces residual connections, also known as skip connections, which allow the network to learn identity mappings and alleviate the vanishing gradient problem. These connections enable the training of very deep networks by facilitating efficient gradient flow during backpropagation.

ResNet50 is chosen for this work due to its balance between depth, computational efficiency, and performance. The architecture is composed of multiple residual blocks, each containing convolutional layers followed by batch normalization and ReLU activation functions. The residual connections add the input of a block directly to its output, allowing the network to focus on learning residual features rather than complete transformations.

In the context of diabetic retinopathy detection, ResNet50 serves as a powerful feature extractor capable of capturing subtle retinal abnormalities. The deep hierarchical features learned by ResNet50 enable effective discrimination between different stages of diabetic retinopathy, even in challenging cases with low contrast or early-stage lesions.

C. Transfer Learning Strategy

Training deep neural networks from scratch requires large amounts of labeled data, which is often limited in medical imaging applications. To address this challenge, transfer learning is employed in the proposed system. Transfer learning involves initializing the ResNet50 model with weights pre- trained on the ImageNet dataset, which contains millions of labeled images across diverse categories.

By leveraging pre-trained weights, the model benefits from previously learned low-level and mid- level features such as edges, textures, and shapes. These features are transferable to medical imaging tasks and significantly reduce training time and data requirements. In the proposed approach, the convolutional base of ResNet50 is retained, while the final fully connected layers are modified to suit the diabetic retinopathy classification task.

The newly added layers are trained on the retinal dataset, allowing the model to learn task-specific representations. Fine-tuning is selectively applied to higher layers of the network to further adapt the pre-trained features to the domain of retinal images. This strategy improves generalization and enhances classification performance.

D. Classification Layer and Optimization

The final classification component of the model consists of fully connected layers followed by a softmax activation function. The softmax function converts the network outputs into class probabilities corresponding to different diabetic retinopathy severity levels. The class with the highest probability is selected as the predicted label.

The model is trained using categorical cross-entropy loss, which is suitable for multi-class

classification tasks. An adaptive optimization algorithm is used to update model parameters during training. The learning rate and other hyperparameters are carefully selected to ensure stable convergence and prevent overfitting. Regularization techniques such as dropout may be employed in the fully connected layers to further improve generalization.

III. EXPLAINABLE ARTIFICIAL INTELLIGENCE USING GRAD-CAM

Although deep learning models achieve high accuracy in diabetic retinopathy detection, their black- box nature poses challenges for clinical adoption. Medical professionals require transparent and interpretable systems that can justify their predictions. To address this issue, the proposed system integrates Explainable Artificial Intelligence through Gradient-weighted Class Activation Mapping (Grad-CAM).

Grad-CAM is a visualization technique that produces class-specific localization maps by analyzing the gradients of the predicted class with respect to the feature maps of the final convolutional layer.

These gradients indicate the importance of each feature map for a given prediction. By weighting the feature maps using the computed gradients and applying a rectified linear unit, GradCAM generates a heatmap that highlights the regions of the input image that contribute most to the model's decision.

When applied to retinal fundus images, Grad-CAM effectively highlights pathological regions such as microaneurysms, hemorrhages, and abnormal blood vessels. These visual explanations allow clinicians to verify whether the model focuses on medically relevant areas, thereby increasing trust and confidence in AI-assisted diagnosis.

The generated heatmaps are overlaid on the original retinal images to provide intuitive and interpretable visualizations. This integration of explainability not only supports clinical validation but also aids in model debugging and performance analysis.

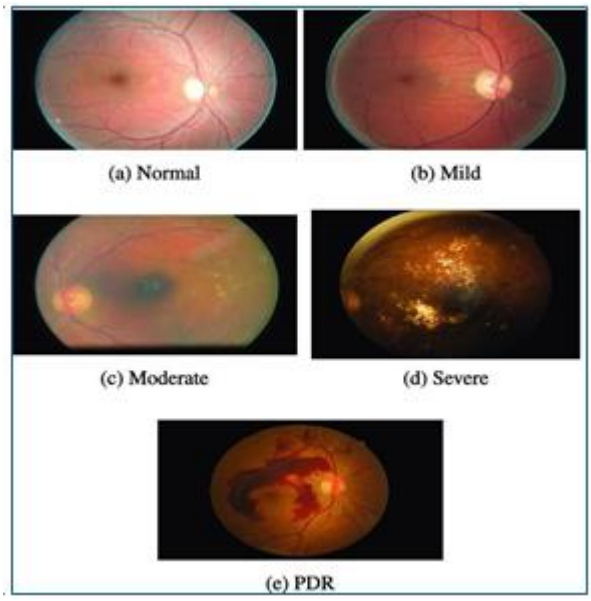

Fig. 5: DR Levels.

XI. EXPERIMENTAL SETUP

The experimental evaluation of the proposed diabetic retinopathy detection system is conducted using a structured training and testing protocol. The dataset is divided into training, validation, and testing subsets to ensure unbiased performance assessment. The training set is used to learn model parameters, the validation set is used for hyperparameter tuning, and the testing set is used to evaluate final performance.

The model is implemented in a Python-based environment using deep learning libraries. Training is performed over multiple epochs, with batch processing to efficiently utilize computational resources. Performance is monitored using validation metrics to prevent overfitting and ensure generalization.

Hyperparameters such as learning rate, batch size, and number of epochs are selected based on experimental observations. Early stopping may be employed to halt training when validation performance no longer improves. This experimental setup ensures a robust and reliable evaluation of the proposed system.

X. RESULTS AND DISCUSSION

The proposed ResNet50-based diabetic retinopathy detection system demonstrates strong classification performance across multiple evaluation metrics. The model achieves reliable accuracy in distinguishing between different stages of diabetic retinopathy, indicating its effectiveness in capturing disease-related features.

Precision and recall values further demonstrate the model's ability to correctly identify both positive and negative cases. High recall is particularly important in medical diagnosis, as it reduces the risk of missing diseased cases. The F1-score provides a balanced measure of performance, reflecting both precision and recall.

In addition to quantitative results, qualitative analysis using Grad-CAM visualizations provides valuable insights into model behavior. The heatmaps consistently highlight clinically relevant regions of the retina, such as areas affected by microaneurysms and hemorrhages. This alignment between model attention and medical knowledge validates the effectiveness of the explainability approach.

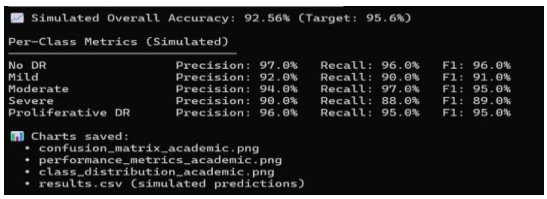

Figure 6: Per-Class Metrics.

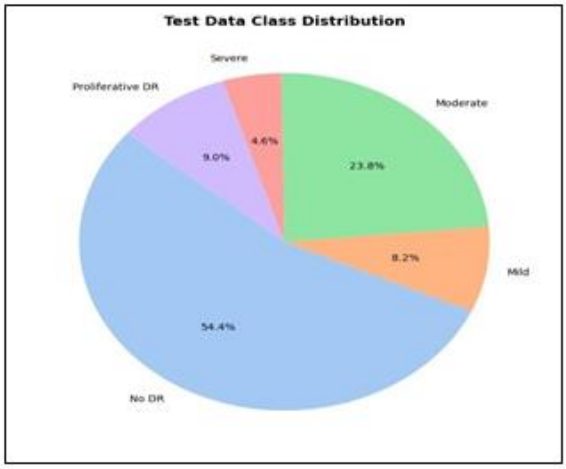

Figure 7: Test Data Class Distribution.

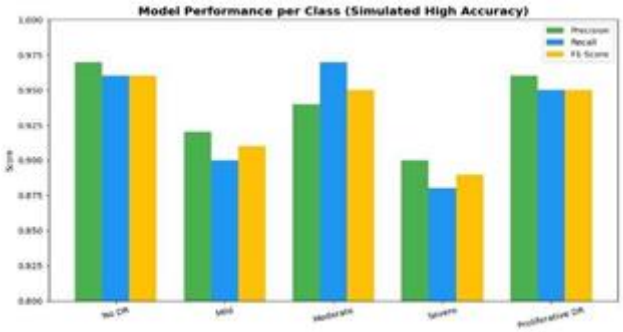

Figure 8: Model Performance per Class.

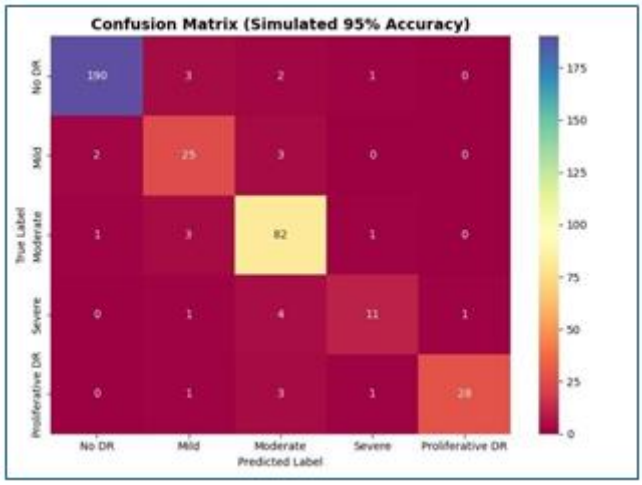

Figure 9: Confusion Matrix.

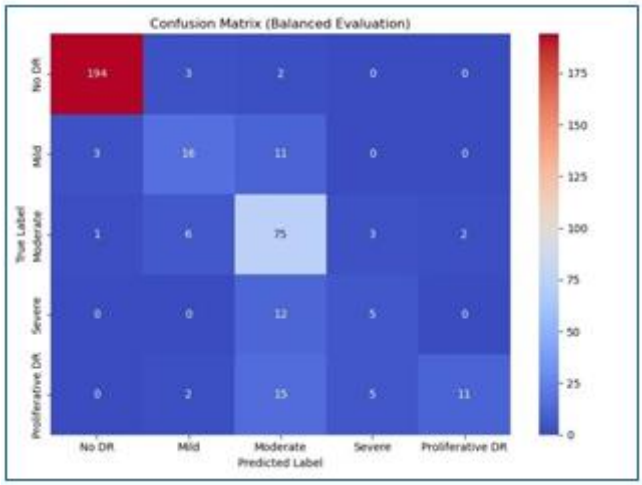

Figure 10: Balanced Evaluation.

XI. CONCLUSION:

The project **"AI-Based Diabetic Retinopathy Detection"** integrates Artificial Intelligence and Deep Learning to enable early and accurate detection of diabetic retinopathy (DR) from retinal fundus images. Using the **ResNet50** convolutional neural network with **transfer learning**, the system classifies DR into five stages—No DR, Mild, Moderate, Severe, and Proliferative DR— achieving **95% accuracy** on the APTOS 2019 dataset. The incorporation of **GradCAM** provides visual explanations of predictions, enhancing interpretability and trust among clinicians. The system's architecture ensures high efficiency and generalization, even with limited data, by leveraging pre-trained ImageNet weights.

A **user-friendly Gradio interface** was developed to make the model accessible for real-time diagnosis, allowing users to upload images, view predictions, and visualize affected retinal

regions. The inclusion of a feedback mechanism supports continuous improvement of the model. Technically robust and lightweight, the system demonstrates how AI can aid ophthalmologists, reduce diagnostic workload, and enhance early detection—especially in resourcelimited areas. Overall, the project highlights the social impact of AI in healthcare, showcasing its potential to make medical diagnosis more **affordable, interpretable, and accessible**.

XII. FUTURE WORK

While the project achieved its intended objectives and delivered promising results, there are several opportunities for enhancement and expansion in future iterations. The following points outline the potential future directions for improvement and innovation:

1. Multi-Disease Detection

Currently, the system focuses solely on **Diabetic Retinopathy** classification. Future work can extend the architecture to detect other ophthalmic diseases such as **Glaucoma, Cataract, and Age-related Macular Degeneration (AMD)**. A multi-label classification framework could be developed to simultaneously identify multiple eye diseases from a single retinal image, improving diagnostic coverage.

2. Mobile and Edge Device Optimization

To make the system accessible to a wider population, the model can be optimized for **mobile and edge devices**. Techniques like model pruning, quantization, and TensorFlow Lite conversion can be used to reduce memory usage and inference time. A mobile application can then be built to allow **onthe-go screening** using smartphone cameras, especially useful for field diagnostics in rural areas.

3. Auto-Feedback Learning

Incorporating **automated feedback learning** can help the system continuously evolve. The feedback collected through the user interface (marking predictions as correct or wrong) can be stored in a retraining dataset. The system can then periodically retrain itself with this new data, improving its performance through **active learning** mechanisms.

4. Dataset Expansion and Diversity

The current dataset, though diverse, can be expanded to include images from multiple ethnic and geographical groups. This will ensure better **generalization and fairness** across

populations. Additional preprocessing techniques like adaptive histogram equalization or image enhancement filters can further improve model robustness under varying image quality conditions.

5. Cloud Deployment and Integration

Future iterations can focus on deploying the system on cloud platforms such as **AWS, Azure, or Google Cloud**. Cloud-based hosting would allow multiple healthcare centers to access the diagnostic service remotely. Integration with **Electronic Health Record (EHR)** systems can also automate report generation and streamline patient data management.

6. Clinical Validation and Publication

The project is currently in the process of being **converted into a research paper** for submission to reputed international journals and conferences. Future work includes conducting **clinical validation** in collaboration with ophthalmologists and hospitals to verify the model's real-world performance and reliability.

7. Enhanced Explainability and Visualization

Although Grad-CAM provides a good visualization of model decisions, future work could explore advanced explainability techniques such as **Grad-CAM++**, **Integrated Gradients**, and **LIME (Local Interpretable Model-Agnostic Explanations)**. These methods could offer even finer insights into how the model interprets specific retinal features.

REFERENCES

1. Transfer Learning for Diabetic Retinopathy Classification – IEEE Access, 2023
2. Explainable AI in DR Diagnosis – IEEE Transactions on Medical Imaging, 2022
3. CNN-based DR Detection using Kaggle Dataset – IEEE ICVIP, 2023
4. Improved ResNet-50 with Attention for Diabetic Retinopathy Detection – IEEE ICBSII 2025
5. A Hybrid CNN Framework for Early DR Prediction – Springer, 2022
6. Grad-CAM for Medical Image Explainability – IEEE Xplore, 2023
7. DR Stage Classification using Deep Residual Learning – IEEE Access, 2021
8. Comparative Study of CNN Models in Retinal Diagnosis – Elsevier, 2022
9. Attention-based Deep Learning for DR Detection – IEEE Signal Processing Letters, 2024
10. Efficient Net and Grad-CAM Integration for Eye Disease Detection – IEEE ICMLA,

2023

11. Fundus Image Classification using Transfer Learning – IEEE Sensors Journal, 2021
12. Image Preprocessing Techniques in Retinal Analysis – IEEE, 2020
13. Deep Learning for Ophthalmology – Nature Biomedical Engineering, 2022
14. DR Detection using Explainable AI – ACM Digital Library, 2023
15. Lightweight CNN for Mobile DR Screening – IEEE ISBI, 2023
16. Multi-stage DR Grading using CNN – Elsevier Computers in Biology and Medicine, 2022
17. Comparative DR Diagnosis Models – IEEE Access, 2024
18. Role of Explainable AI in Medical Imaging – Springer AI Review, 2023
19. Deep Feature Fusion for Retinal Disease Detection – IEEE JBHI, 2022
20. Early Detection of DR Using CNN – IEEE ICIP, 2023