IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁙ Peer-reviewed & Refereed journal ⁙ Vol. 14, Issue 12, December 2025

DOI: 10.17148/IJARCCE.2025.141206

A REVIEW ON EXPLAINABLE CNN FOR EARLY DETECTION OF DIABETIC RETINOPATHY DIAGNOSIS

NIMISHA PS[1], AYSWARIYA VJ[2]

Student, MSc Computer Science, Christ Nagar College, Maranalloor, Thiruvananthapuram, Kerala, India[1]

Assistant Professor, PG Department of Computer Science, Christ Nagar College, Maranalloor, Thiruvananthapuram, Kerala, India[2]

Abstract: Diabetic retinopathy (DR) is a leading cause of vision impairment globally, emphasizing the urgent need for early and accurate detection methods. Recent advancements in deep learning (DL) have demonstrated significant potential in automating DR diagnosis from retinal fundus images, thereby aiding clinicians in timely intervention. Nevertheless, the opacity of DL models remains a barrier to their widespread clinical adoption, necessitating transparent and explainable solutions. This paper proposes an integrated framework that combines state of the art deep learning architectures with explainable artificial intelligence (XAI) techniques, specifically Grad-CAM, to improve the interpretability of the diagnosis process. The methodology involves training multiple DL models, including a novel customized convolutional neural network (CNN), on high-resolution fundus image datasets, complemented by extensive data augmentation and preprocessing strategies to address class imbalance and image variability. The incorporation of XAI enables visualization of model decisions, fostering trust and facilitating clinical validation. Experimental results demonstrate that the proposed approach achieves high classification accuracy, superior early-stage detection capabilities, and meaningful interpretability insights, potentially enhancing clinical decision support systems for diabetic retinopathy.

Keywords*:* Diabetic Retinopathy, Retinal Fundus Images, Grad-CAM, Customized Convolutional Neural Network.

I. INTRODUCTION

Diabetic retinopathy (DR) is a major cause of blindness globally, with early detection being critical to prevent irreversible vision loss. Traditional diagnosis relies on manual examination of fundus images by specialists, which can be time-consuming and subject to variability. The rapid advancements in artificial intelligence (AI), particularly deep learning, have shown promise in automating DR detection with high accuracy. Various models, including convolutional neural networks (CNNs), have demonstrated the capability to analyze retinal images efficiently.

Recent efforts have focused not only on improving diagnostic accuracy but also on enhancing the transparency and interpretability of AI systems. Integrating explainable artificial intelligence (XAI) techniques, such as Grad-CAM, has been effective in visualizing model decision processes. This enables healthcare professionals to understand and trust AI outputs, facilitating clinical acceptance.

The approach developed combines deep learning architectures with visualization methods to provide accurate and interpretable assessments of retinal images for early DR detection. Such integration aims to advance automation in ophthalmology by offering reliable tools that support timely diagnosis and treatment planning, ultimately reducing the risk of vision loss from diabetic retinopathy.

II. BACKGROUND AND CONTEXT

Diabetic Retinopathy (DR) is a serious eye-related complication that develops in individuals living with diabetes. It occurs when prolonged high blood sugar levels damage the tiny blood vessels in the retina, leading to vision problems and, if untreated, permanent blindness. With diabetes cases rising rapidly around the world, DR has become a major public health concern. According to the World Health Organization, more than 422 million people are affected by diabetes, and this number continues to grow every year. As a result, the population at risk of developing vision-threatening eye diseases is steadily increasing. Early diagnosis and timely treatment can prevent up to 90% of vision loss cases, making early detection extremely important.

Traditionally, DR screening is performed manually by ophthalmologists through fundus image examination. However, this process is time-consuming, labor-intensive, and requires trained specialists. In many low- and middle-income regions, access to eye care services is limited, and screening programs are difficult to operate due to high costs and a shortage of experts. Furthermore, subtle early stage symptoms, such as microaneurysms and small hemorrhages, are often difficult to identify, even for experienced specialists.

To address these challenges, researchers have turned to Artificial Intelligence (AI) and deep learning techniques to build automated systems capable of detecting DR from retinal images. Models such as CNNs, SVMs, and transfer learning architectures like VGG, ResNet, and Xception have shown promising results in accurately identifying disease severity. These AI-based systems can analyze thousands of retinal images quickly and consistently, enabling large-scale screening and assisting medical professionals with efficient decision-making.

However, several limitations still remain. Many AI models rely on imbalanced datasets, which affects performance when predicting less common but critical disease stages, such as severe or proliferative DR. Another challenge is the lack of interpretability—clinicians require transparent explanations for AI predictions, rather than just numerical outputs. This has encouraged researchers to incorporate Explainable Artificial Intelligence (XAI) techniques such as Grad-CAM, which highlight the exact regions of the retina used by the model in making its decisions.

The present study advances current research in automated diabetic retinopathy detection by integrating deep learning models with explainable AI techniques. A customized CNN architecture is proposed, achieving improved accuracy and offering interpretable visual insights through Grad-CAM visualization, making the system more reliable.

III. RELATED WORKS

Several studies have explored deep learning approaches to improve diabetic retinopathy (DR) detection and classification. Researchers are continuously trying to build models that can accurately identify the disease at an early stage, as early diagnosis plays a major role in preventing vision loss. These advancements not only help doctors make faster decisions but also support screening in remote areas where specialist availability is limited.

‘Enhancing Early Detection of Diabetic Retinopathy Through the Integration of Deep Learning Models and Explainable Artificial Intelligence’[1] proposed an advanced deep learning model combined with Explainable AI (XAI) to improve early detection of diabetic retinopathy. Their approach not only enhances classification accuracy but also helps doctors understand which regions of the image influenced the model’s decision, increasing trust and transparency in clinical use.

‘Development and Validation of a Deep Learning Algorithm for Detection of Diabetic Retinopathy in Retinal Fundus Photographs’ [2] developed a highly successful deep learning model for DR screening that achieved about 94% accuracy, proving that AI can perform at a level comparable to expert ophthalmologists.

‘Convolutional Neural Networks for Diabetic Retinopathy’[3] proposed a CNN-based method with around 75% accuracy, which, although lower, played an important role by showing how deep learning can automatically extract features from retinal images.

‘Artificial Intelligence Approach for Diabetic Retinopathy Using GSCNN’[4] improved DR detection using transfer learning and data augmentation, reaching around 89% accuracy. Their work emphasized how pretrained models can enhance performance.

Similarly, ‘General Deep Learning Model for Detecting Diabetic Retinopathy Using NASNet-Large’[5] achieved 92.5% accuracy by focusing on stronger featureextraction techniques to reduce misclassification

‘Graph-Based Deep Learning Framework for Diabetic Retinopathy Classification’ [6]presented an optimized deep learning framework with 89.9% accuracy, highlighting the benefits of improved training strategies and fine-tuning.

‘Deep Learning–Based System for Diabetic Retinopathy Detection in Real Clinical Settings’ [7] developed a robust model attaining 93.72% accuracy, showing good capability in identifying multiple stages of DR.

‘Lightweight Transformer–CNN Hybrid Model for Diabetic Retinopathy Detection’ [8] obtained about 90.17% accuracy using a hybrid deep learning approach to improve classification reliability across various severity levels

‘Classification of Diabetic Retinopathy Severity Using DenseNet121 and ResNet50’[9] achieved 85% accuracy, demonstrating the effectiveness of preprocessing techniques even when datasets are limited

‘E-DenseNet Hybrid Deep Model for Diabetic Retinopathy Severity Classification’[10] further enhanced performance with 91.2% accuracy, integrating data balancing and image enhancement strategies

Meanwhile, ‘Accuracy of Diabetic Retinopathy Staging Using Deep CNN with UltraWide-Field Fundus and OCTA Images’[11] reported results in the range of 88–89%, proving that deep learning methods can still perform strongly even under challenging real-world image variation.

Overall, the progression of research shows a consistent improvement in accuracy and system reliability. The shift from simple CNN models to more advanced, optimized, and hybrid deep learning techniques demonstrates the growing potential of AI-based solutions to support early diagnosis and assist ophthalmologists in real clinical applications.

DOI: 10.17148/IJARCCE.2025.141206

IV. SYSTEMATIC ANALYSIS

A systematic analysis of existing work on automated diabetic retinopathy detection reveals that although numerous deep learning, transfer learning, and hybrid CNN-based models have been introduced, each presents trade-offs in accuracy, computational requirements, dataset dependency, and real-world adaptability. Many high-performing models achieve strong results on public datasets, yet lack external validation and struggle with class imbalance, especially for early and severe DR stages. Furthermore, interpretability remains limited in most studies, restricting clinical acceptance, while only a few approaches address practical deployment challenges such as lightweight architectures, privacy, and cross-center generalization. This analysis highlights the need for a more efficient, explainable, and generalizable DR classification framework supported by real-world testing. In this section, we critically evaluate the performance, strengths, and limitations discussed in the related works.

Reference no.	Methodology	Dataset(s)	Accuracy	Merits	Demerits
Alavee K.A. et al. [1]	Transfer Learning models, SVM , RNN, Proposed Custom CNN with Explainable AI	APTOS Kaggle dataset	95.27%	High accuracy,integr ates XAI for transparency, supports early diagnosis,effec tive multi model comparison	Dataset details not clearly mentioned, high computational cost,limited generalizability due to unclear dataset diversity
Gulshan V. et al. [2]	Deep CNN	EyePACS	94%	Very high sensitivity, performs well across devices, near-clinical performance	Requires huge dataset,interpre tability limited.
Pratt H. et al [3]	CNN	Kaggle DR	75%	Strong retinal feature extraction,han dles moderate variations.	Struggles on low-quality images, tends to overfit.
Rajamani S. et al. [4]	GSCNN (Grid Search tuned CNN), augmentation , multi-class DR classification	Kaggle DR dataset	89%	Hyperparamet er tuning (GSCV), better than baseline CNN, good early-stage detection.	Only Kaggle used (no external testing), risk of overfitting
Chen P.N. et al. [5]	NASNetLarge transfer learning, SMOTE, two-module DR detection	EyePACS, DIARETDB0, DIARETDB1, eOphtha, MESSIDOR, DRIVE	92.5%	Two-stage detection, strong transfer learning, tested on many datasets	High compute cost (NASNet), synthetic oversampling may bias results.
G. Zhang et al. [6]	CNN backbone, Graph ,GCN, KNN graph, three custom loss functions (graph-center, pseudocontrastive,	EyePACS-1, Messidor-2	89.9%	No manual annotations required, uses graph correlation learning, good accuracy on 2 datasets	Graph construction is complex, computationall y heavy, lower sensitivity than some supervised CNNs

	transformation-invariant)				annotations required, uses graph correlation learning, good accuracy on 2 datasets
A. Bajwa et al. [7]	Modified CNN, image quality filter, clinician validation	Test conducted on 398 patients at + their fundus images	93.72%	Real-world testing in hospital , quality-assessment of images, clinically validated	Relatively small / private dataset, binary classification (DR-positive vs DR-negative)
Khan I.U. et al. [8]	Compact Convolutional Transformer (CCT) + finetuning +	APTOS, Messidor-2, IDRiD, DDR, Kaggle DR dataset	90.17%	Lightweight model (transformer-CNN), computationall	Accuracy isn't very high, transformer-CNN complexity,
	transfer learning			y efficient, uses multiple public datasetsl	might underperform on more varied data
Zhang J. et al. [9]	DenseNet121 + ResNet50, image modality comparison, CNN classification	Public fundus image datasets (Kaggle/EyePACS-like sources)	85%	Compares multiple architectures, tests RGB/green/contrast images, interpretable results	Medium accuracy, no major performance gain from preprocessin
Mohanty C. et al. [10]	DenseNetbased hybrid CNN, preprocessing , DR multi-class classification	Public fundus datasets (color fundus images)	91.2%	High accuracy, DenseNet feature reuse, strong sensitivity	High computational cost, potential overfitting, class imbalance issues
Nagasawa T. et al. [11]	VGG16 CNN, transfer learning, crossvalidation	UWF fundus OCTA images from clinical subjects	88–89%	Uses advanced imaging ,VGG 16 transfer learning, good staging results	small dataset, limited generalizability

V. CONCLUSION AND FUTURE WORK

Extensive exploration of existing research on automated detection of diabetic retinopathy shows clear progress in the application of deep learning and explainable artificial intelligence for medical imaging. Early methods relying on handcrafted features demonstrated limited capability in identifying subtle clinical indicators, while modern convolutional and transformer-based architectures achieved far higher accuracy and robustness.Despite these advancements, many approaches continue to face challenges related to interpretability, data diversity, and real-world usability. The integration of explainability tools has proven essential in addressing clinician concerns by highlighting the visual evidence behind model predictions. Through systematic examination of multiple techniques and methodologies, the most effective strategies appear to be those that combine strong predictive performance with meaningful, clinically aligned explanations. This balanced approach offers a practical path toward enhancing early detection and strengthening trust in AI-assisted diagnostic support.

Going forward, research in this field should place greater emphasis on developing models that perform consistently across different datasets, patient groups, and real clinical settings. Expanding training and evaluation using diverse, multi-center data will help ensure that models remain reliable outside controlled research environments. It is also important to improve explainability tools so that clinicians can clearly understand and trust the system's decisions rather than treating the model as a "black box."

Future work should explore practical deployment aspects such as real-time processing, user-friendly interfaces, and smooth integration with existing hospital workflows. There is also strong potential in combining retinal images with patient medical records or longterm monitoring data, which could lead to better prediction of disease progression and more personalized risk assessment. Developing lightweight and efficient models that run on mobile or edge devices can improve accessibility, especially in rural and low-resource settings where specialist care is limited.

REFERENCES

[1]. K. A. Alavee, M. Hasan, A. H. Zillanee, M. Mostakim, J. Uddin, E. S. Alvarado, I. de la Torre Diez, I. Ashraf and M. A. Samad, "Enhancing Early Detection of Diabetic Retinopathy Through the Integration of Deep Learning Models and Explainable Artificial Intelligence," IEEE Access, vol. 12, pp. 73950–73969, 2024, doi: 10.1109/ACCESS.2024.3405570.

[2]. V. Gulshan, L. Peng, M. Coram, M. C. Stumpe, D. Wu, A. Narayanaswamy, S. Venugopalan, K. Widner, T. Madams, J. Cuadros, R. Kim, R. Raman, P. C. Nelson, J. L. Mega and D. R. Webster, "Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs," JAMA, vol. 316, no. 22, pp. 2402–2410, 2016, doi: 10.1001/jama.2016.17216.

[3]. H. Pratt, F. Coenen, D. M. Broadbent, S. P. Harding and Y. Zheng, "Convolutional Neural Networks for Diabetic Retinopathy," Procedia Computer Science, vol. 90, pp. 200–205, 2016, doi: 10.1016/j.procs.2016.07.014.

[4]. S. Rajamani and S. Sasikala, "Artificial Intelligence Approach for Diabetic Retinopathy Severity Detection," Informatica, vol. 46, no. 8, pp. 463-470, 2022, doi: 10.31449/inf.v46i8.4425.

[5]. P.-N. Chen, C.-C. Lee, C.-M. Liang, S.-I. Pao, K.-H. Huang and K.-F. Lin, "General deep learning model for detecting diabetic retinopathy," BMC Bioinformatics, vol. 22, Art. 84, 2021, doi: 10.1186/s12859-021-04005-x.

[6]. G. Zhang, B. Sun, Z. Chen, Y. Gao, Z. Zhang, K. Li and W. Yang, "Diabetic Retinopathy Grading by Deep Graph Correlation Network on Retinal Images Without Manual Annotations," Frontiers in Medicine, vol. 9, Art. 872214, 2022, doi:10.3389/fmed.2022.872214.

[7]. A. Bajwa, N. Nosheen, K. I. Talpur and S. Akram, "A Prospective Study on Diabetic Retinopathy Detection Based on Modified Convolutional Neural Network Using Fundus Images," Diagnostics, vol. 13, no. 3, Art. 393, 2023, doi: 10.3390/diagnostics13030393.

[8]. I. U. Khan, M. A. K. Raiaan, K. Fatema, S. Azam, R. U. Rashid, S. H. Mukta, M. Jonkman and F. De Boer, "A Computer-Aided Diagnostic System to Identify Diabetic Retinopathy, Utilizing a Modified Compact Convolutional Transformer and LowResolution Images to Reduce Computation Time," Biomedicines, vol. 11, no. 6, Art. 1566, 2023, doi: 10.3390/biomedicines11061566.

[9]. J. Zhang, "Classification of Diabetic Retinopathy Severity in Fundus Images with DenseNet121 and ResNet50," arXiv preprint, arXiv:2108.08473, 2021.

[10]. C. Mohanty, S. Mahapatra, B. Acharya, F. Kokkoras, V. C. Gerogiannis, I.Karamitsos and A. Kanavos, "Using Deep Learning Architectures for Detection and Classification of Diabetic Retinopathy," Sensors, vol. 23, no. 12, Art. 5726, 2023, doi: 10.3390/s23125726.

[11]. T. Nagasawa, H. Tabuchi, H. Masumoto, H. Enno, M. Niki et al., "Accuracy of ultrawide-field fundus ophthalmoscopy-assisted deep learning for detecting treatmentnaïve proliferative diabetic retinopathy," International Ophthalmology, vol. 39, pp. 2153–2159, 2019, doi: 10.1007/s10792-019-01074