Academic Journal of Science and Technology,
Aug, 2025, vol.6, No.1, p283-288

Deep Learning-Based Approach for Diabetic Retinopathy Detection with Explainable AI

Abdelkader Alrabai[1*]

1. Physics Department, Faculty of Education, Wadi Alshatti University, Alshatti – Libya

ABSTRACT

Diabetic Retinopathy (DR) is one of the most common complications of diabetes and a leading cause of vision loss among working-age adults worldwide. Therefore, early and accurate detection is crucial to preventing vision impairment and blindness. Automated deep learning-based diagnostic tools can play a transformative role in large-scale screening programs by enabling rapid, consistent, and cost-effective diagnosis of DR. This study explored the use of Convolutional Neural Networks (CNNs), specifically the VGG16 model, for detecting DR from retinal images and evaluating its diagnostic performance on a labeled dataset. Experimental results show that VGG16 performed strongly across all metrics, achieving an accuracy of 98.19%, precision of 98,21%, recall of 98.17%, and an F1-score of 98.18%, indicating robust and reliable performance in DR detection. In addition, this study applies Explainable AI (XAI) method—Occlusion —to improve the transparency and interpretability of deep learning models for DR detection. The findings highlight the importance of both accuracy and interpretability in building trust in automated diagnostics. By enabling early detection and supporting clinical workflows, the integration of high-performing models with XAI techniques offers a promising direction for reliable, AI-powered eye care solutions.

KEYWORDS: *CNNs, DR, Occlusion, VGG16, XAI.*

* *Corresponding author:* a.alrabai@wau.edu.ly

1. INTRODUCTION

Diabetic retinopathy (DR), a primary vision-related consequence of diabetes mellitus, affects roughly 30–40% of those with the condition. Over 100 million people worldwide currently live with DR, making it a prominent contributor to vision loss, particularly among adults of working age. The number of affected individuals is anticipated to grow markedly—from 103 million in 2020 to 130 million by 2030, and reaching 161 million by 2045. This upward trend—exceeding a 25% rise within a decade—is linked to the growing incidence of diabetes, shifts in lifestyle habits, and increasing longevity. The expanding burden is expected to put additional pressure on healthcare infrastructure and result in considerable financial implications [1]. DR progresses through two main clinical stages: non-proliferative (NPDR) and proliferative (PDR). NPDR, the initial phase, involves damage to retinal blood vessels, leading to signs such as microaneurysms, hemorrhages, and hard exudates, often without noticeable symptoms. PDR is the more severe stage, characterized by abnormal new vessel growth (neovascularization) that can result in complications like vitreous hemorrhage and retinal detachment. In addition, Diabetic Macular Edema (DME), caused by fluid leakage due to blood-retinal barrier breakdown, can occur at any stage and contributes significantly to vision impairment [2]. NPDR itself is subdivided into mild, moderate, and severe forms based on the severity of retinal damage. Mild cases primarily show microaneurysms; moderate stages present additional lesions like hemorrhages, hard exudates, and cotton wool spots; severe NPDR is marked by extensive ischemia with features such as venous beading and intra-retinal microvascular abnormalities. Since early identification is key to preventing vision loss, regular ophthalmic evaluations are recommended to detect these vascular changes promptly [3].

Routine retinal screening is essential for the early detection of DR, particularly because the condition often shows no symptoms during its initial phases. Early identification allows for timely intervention, which can prevent or slow vision loss. Traditionally, ophthalmologists depend on manual grading methods—carefully examining retinal images by eye—to diagnose and assess the severity of DR. This process, while effective, can be time-consuming and requires significant expertise, underscoring the need for consistent and thorough screening protocols [4].

Traditionally, DR screening relies on imaging the eye's fundus to identify signs of disease. In recent years, the evaluation of these retinal images has been significantly enhanced through the use of machine learning techniques. These advanced algorithms can accurately detect subtle abnormalities, including microaneurysms, which represent the earliest detectable indicators of retinal damage. This integration of machine learning into screening protocols improves diagnostic accuracy, and additionally helps streamline the analysis process, enabling earlier and more reliable identification of DR [5]. Retinal imaging, especially fundus photography, plays a key role in DR detection by identifying signs like microaneurysms and hemorrhages. However, traditional methods are limited by high costs, reliance on experts, and difficulty in detecting depth-related conditions like DME [6]. Deep learning, a subset of machine learning, uses deep neural networks to automatically learn complex features from data, making it highly effective for DR detection in fundus images. Unlike traditional image analysis methods that often miss subtle patterns, Deep learning —especially through deep CNNs—can capture detailed hierarchical information directly from raw pixels. With end-to-end training, these models efficiently classify disease stages or recognize healthy retinas. Their performance improves with large annotated datasets, allowing for early DR detection and continuous model refinement over time [7]. Although effective, deep learning techniques are often criticized for their black box nature, providing little insight into how predictions are made or which features contribute to the output. This lack of interpretability poses a barrier to clinical adoption, as ophthalmologists and other end-users may struggle to trust systems they cannot fully understand. Ethical, safety, and legal concerns also arise due to the absence of transparency and human oversight. While achieving high classification accuracy is important, the ability to understand the reasoning behind model decisions is increasingly valued. Various interpretability methods are now being explored to visualize and explain CNN behavior in DR applications [8].

Recently, deep learning models—particularly CNNs—have shown impressive performance in various medical imaging tasks, including DR detection. Numerous investigations [9–12] have explored the application of deep learning techniques for identifying and categorizing DR, employing a variety of models and datasets. These efforts have generally yielded promising performance, highlighting the potential of such approaches in automated interpretation of diagnostic imagery.

However, the black-box nature of deep learning models poses a challenge in clinical settings, where transparency and trust are crucial. The lack of interpretability raises concerns about safety, accountability, and clinical acceptance. This study evaluates a deep learning framework based on the VGG16 model for binary classification of DR, incorporating interpretability techniques to improve transparency. Occlusion method are used to visually highlight regions influencing the model's predictions. By combining strong classification performance with visual explanations, the study aims to support accurate and clinically meaningful AI-assisted diagnosis, particularly in identifying key pathological features like microaneurysms, hemorrhages, and exudates.

2. METHODOLOGY

This study introduces an explainable deep learning approach for detecting DR, using the VGG16 model to classify retinal fundus images into DR and No DR categories. To understand and interpret the model's predictions, explainability technique was applied. The methodology involves several key stages, including the careful selection and preparation of the dataset, model training and performance evaluation, followed by the application of an appropriate interpretability technique to explain the model's predictions.

2.1. Dataset and preprocessing:

The dataset used in this study comprises retinal fundus images aimed at detecting DR. The problem was formulated as a binary classification task, distinguishing between DR and No DR cases to prioritize early-stage detection. The dataset utilized is the AISOP2019 [13] collection, which categorizes images into five distinct classes: No DR, mild, moderate, severe, and proliferative DR. The distribution of images across these classes is as follows: 1805 images in the No DR class, 370 in mild, 999 in moderate, 193 in severe, and 295 in proliferative DR. Examples representing each class from the dataset are presented in Figure 1. To simplify the classification task into a binary prediction model, the dataset was reorganized into two groups: No DR and DR. The No DR group contains all 1805 images labeled as free from DR, while the DR group merges all images from the other four categories (mild, moderate, severe, and proliferative DR), resulting in a combined total of 1857 images exhibiting varying degrees of DR. This grouping enables the model to focus on distinguishing between the presence and absence of DR, which is critical for timely intervention and treatment.

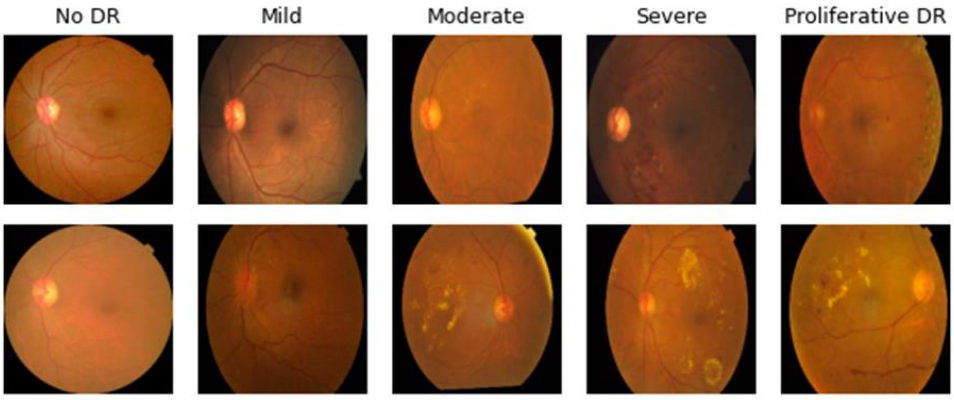

Figure 1. Sample dataset images

To ensure compatibility with the input specifications of the model utilized, all retinal images were uniformly resized to 224×224 pixels. Following resizing, the images underwent normalization to standardize the pixel intensity values, which helps improve the models' convergence and overall performance. The dataset was then divided into three subsets for effective model training and evaluation: 70% of the images were allocated for training, 15% for validation to fine-tune model parameters, and the remaining 15% were reserved for testing to assess the final model's predictive accuracy on unseen data.

2.2. Employed model:

To detect diabetic retinopathy, the pre-trained VGG16 model was utilized. Through transfer learning, the model was fine-tuned on a retinal image dataset to evaluate its effectiveness in early diagnosis, contributing to the development of automated ophthalmic screening systems.

VGG16 [14] is a well-established CNN architecture, frequently applied in medical image analysis due to its reliable performance. Its design is straightforward, utilizing repeated 3×3 convolutional filters and organizing layers into sequential blocks, each followed by max-pooling. The network comprises 16 weight layers, ending with fully connected layers. Its popularity stems from a strong balance between architectural depth and computational efficiency. Leveraging its pre-trained weights from ImageNet, VGG16 can be fine-tuned for specialized tasks using smaller datasets via transfer learning. In this study, the model was adapted for binary classification to detect DR by replacing its original classification layers with custom fully connected layers, enabling it to differentiate between DR and non-DR images effectively.

2.3. Training and evaluation:

To ensure robust and consistent model training, the VGG16 architecture was initialized with weights pre-trained on the ImageNet dataset. Training was conducted using the Adam optimizer with a learning rate of 0.0001, a batch size of 32, and for 25 epochs. The performance of VGG16 was evaluated using widely accepted classification metrics to provide a comprehensive understanding of the model's ability to accurately detect DR. In addition, a confusion

matrix was generated to offer more analysis of classification errors. All evaluations were performed exclusively on the held-out test set to ensure an unbiased assessment of the model's ability to generalize to new, unseen data. This rigorous evaluation framework supports the validation of VGG16's robustness and reliability in real-world diagnostic scenarios.

2.4. Explainable AI technique:

To improve the interpretability of the model's predictions and foster greater clinical trust, explainable AI technique—occlusion—was employed. This method helps visualize and highlight the regions of retinal images that most influenced the model's decisions, offering valuable insights into its diagnostic reasoning. The Captum [15] library was used to implement this explainability technique in a consistent and efficient manner. Using Captum, occlusion sensitivity [16] analysis identified how masking different parts of the image affected the prediction, indicating which regions were most critical to the model's output. By applying this method, a direct qualitative understanding of how the model attends to relevant pathological features was achieved. This interpretability framework not only helps validate the model's predictions but also supports clinicians in understanding, trusting, and potentially integrating AI-driven tools into real-world diagnostic workflows.

3. RESULTS AND DISCUSSIONS

The results of the model's performance evaluation are presented in Table 1, which summarizes the key evaluation metrics used to assess the model. These metrics were calculated based on the model's predictions on the test dataset and are used to quantify its effectiveness in terms of classification performance. The values reported offer a clear and concise representation of the model's behavior and serve as the basis for further analysis and comparison with other approaches.

Table 1. Summary of model evaluation results

Metrics	Accuracy	Precision	Recall	F1-score
VGG16	0.9819	0.9821	0.9817	0.9818

As shown in Table 1, the VGG16 model achieves an accuracy of 0.9819, precision of 0.9821, recall of 0.9817, and an F1-score of 0.9818. These consistently high values indicate the model's strong and balanced performance in classifying DR images, demonstrating its reliability across all key evaluation metric. In addition, the confusion matrix, which provides a detailed breakdown of the model's classification results, was calculated and is shown in Figure 2.

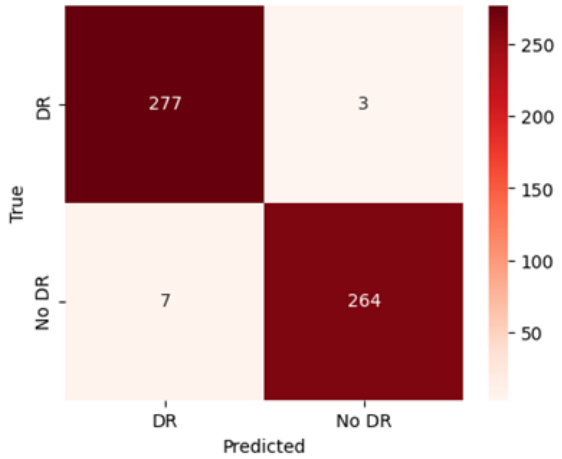

Figure 2. Confusion matrix

The confusion matrix for the VGG16 model shows its performance in classifying DR and No DR cases. The model correctly identified 277 true DR cases and 264 true No DR cases. There were 3 instances where the model incorrectly predicted No DR when it was actually DR, and 7 instances where it predicted DR when it was actually No DR. In total, the dataset used for evaluation consisted of 551 samples. Specifically, there were 280 true DR cases and 271 true No DR cases in the set. The presented confusion matrix indicates that the VGG16 model is performing well, demonstrating a strong ability to accurately distinguish between DR and No DR. These results collectively suggest that the model is reliable and effective for this specific classification task.

Figure 3 provides the model's prediction on a sample from the test set, displaying the original image along with the true label, predicted class, and the model's confidence probability. The model predicts the sample as class (predicted class) with a confidence (confidence percentage), which matches the true class label (true class), indicating accurate classification on this example.

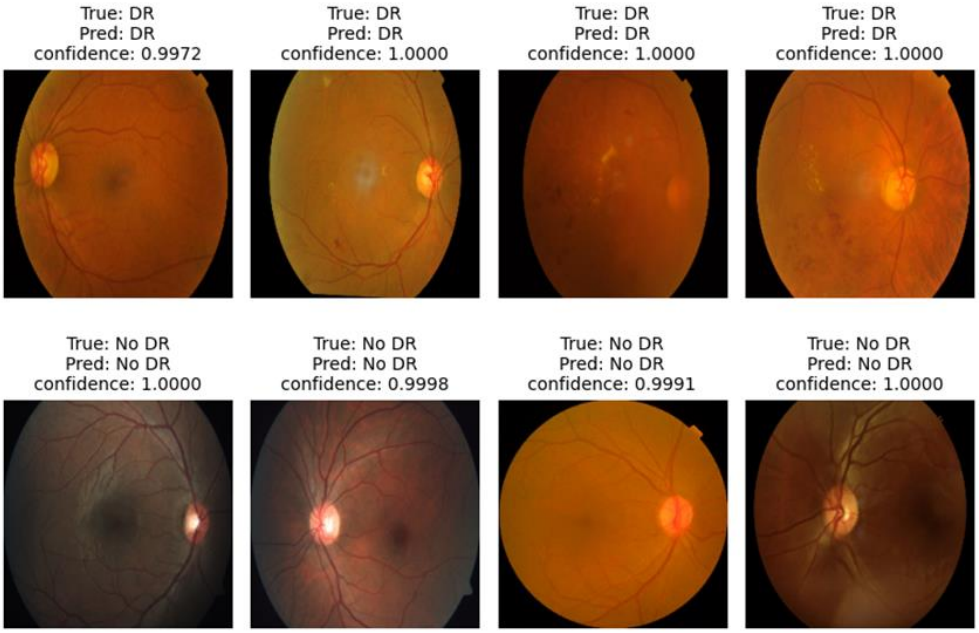

Figure 3. Model prediction sample

As shown in Figure 3, the model demonstrates strong classification performance on individual retinal images. The top row highlights four samples with true labels of DR, all correctly predicted by the model with very high confidence scores ranging from 0.9972 to 1.0000. The bottom row shows four No DR cases, each also accurately classified with confidence scores between 0.9991 and 1.0000. These examples visually confirm the model's high accuracy and confidence in distinguishing between DR and No DR, supporting the positive results observed in the confusion matrix.

Figure 4 presents the results occlusion method applied to a test set image to highlight the regions most influential for the model's prediction.

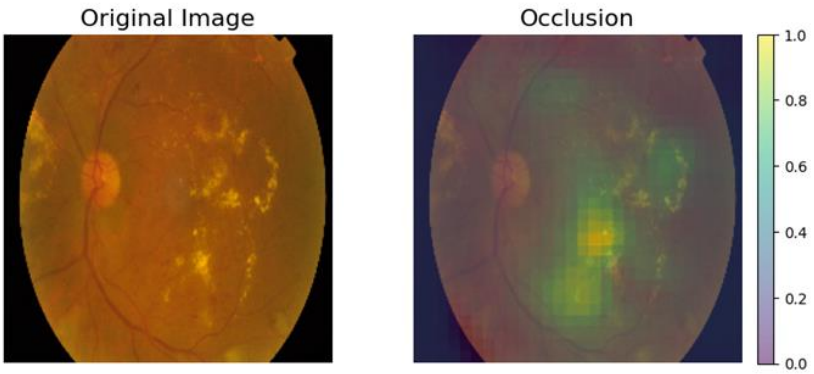

Figure 4. Occlusion heatmaps on a test image

Occlusion perturbs patches of the input to measure their impact on the output. This attribution method provides complementary insights into the model's decision-making process by identifying key areas in the image that contribute to the classification. As shown in Figure 4, the occlusion map analysis provides insights into the regions of the retinal image that the model focuses on when making predictions. The left panel displays the original retinal image, which shows signs of DR, such as exudates and microaneurysms. The right panel presents the occlusion map, highlighting areas where occlusion leads to the greatest decrease in the model's confidence. Brighter, warmer colors (yellows and greens) correspond to regions that significantly influence the model's decision. In this example, the occlusion map clearly emphasizes the characteristic lesions of DR, indicating that the model attends to clinically relevant features rather than irrelevant image regions. This visualization supports the

model's interpretability and reliability in identifying key pathological signs.

4. CONCLUSION

This study explored the use of deep learning and explainable AI technique—specifically occlusion—for early detection of DR from retinal fundus images. The VGG16 model demonstrated high diagnostic accuracy and clinical relevance, effectively highlighting key features like microaneurysms and hemorrhages. By producing interpretable visualizations, explainable AI improved model transparency and clinician trust, addressing a key barrier to clinical adoption. The findings support VGG16's potential for integration into real-world DR screening, especially in resource-limited settings. Automated, interpretable systems can aid early diagnosis, reduce workload, and increase diagnostic consistency. The study emphasizes the importance of combining accurate models with interpretability tools for safe, trustworthy AI deployment in healthcare. Future work will expand to include multi-class grading of DR, incorporate relevant clinical metadata, and examine a broader range of interpretability methods. Additional efforts will involve testing the model on larger datasets, utilizing alternative architectures, and assessing performance in practical, real-world clinical environments.

REFERENCE

1. Tan, T. E., & Wong, T. Y. (2023). Diabetic retinopathy: Looking forward to 2030. Frontiers in Endocrinology, 13, 1077669.
2. Wang, W., & Lo, A. C. (2018). Diabetic retinopathy: pathophysiology and treatments. International journal of molecular sciences, 19(6), 1816.
3. Kusuhara, S., Fukushima, Y., Ogura, S., Inoue, N., & Uemura, A. (2018). Pathophysiology of diabetic retinopathy: the old and the new. Diabetes & metabolism journal, 42(5), 364.
4. Mohanty, C., Mahapatra, S., Acharya, B., Kokkoras, F., Gerogiannis, V. C., Karamitsos, I., & Kanavos, A. (2023). Using deep learning architectures for detection and classification of diabetic retinopathy. Sensors, 23(12), 5726.
5. Faura, G., Boix-Lemonche, G., Holmeide, A. K., Verkauskiene, R., Volke, V., Sokolovska, J., & Petrovski, G. (2022). Colorimetric and electrochemical screening for early detection of diabetes mellitus and diabetic retinopathy—application of sensor arrays and machine learning. Sensors, 22(3), 718.
6. Asif, M., ur Rehman, F., Rashid, Z., Hussain, A., Mirza, A., & Qureshi, W. S. (2025). An Insight on the Timely Diagnosis of Diabetic Retinopathy using Traditional and AI-Driven Approaches. IEEE Access.
7. Bhulakshmi, D., & Rajput, D. S. (2024). A systematic review on diabetic retinopathy detection and classification based on deep learning techniques using fundus images. PeerJ Computer Science, 10, e1947.
8. Lim, W. X., Chen, Z., & Ahmed, A. (2022). The adoption of deep learning interpretability techniques on diabetic retinopathy analysis: a review. Medical & biological engineering & computing, 60(3), 633-642.
9. Khudaier, A. H., & Radhi, A. M. (2024). Binary classification of diabetic retinopathy using CNN architecture. Iraqi Journal of Science, 963-978.
10. Aldughayfiq, B., Ashfaq, F., Jhanjhi, N. Z., & Humayun, M. (2023). Explainable AI for retinoblastoma diagnosis: interpreting deep learning models with LIME and SHAP. Diagnostics, 13(11), 1932.
11. Kolla, M., & Venugopal, T. (2021, March). Efficient classification of diabetic retinopathy using binary cnn. In 2021 International conference on computational intelligence and knowledge economy (ICCIKE) (pp. 244-247). IEEE.
12. Saproo, D., Mahajan, A. N., & Narwal, S. (2024). Deep learning based binary classification of diabetic retinopathy images using transfer learning approach. Journal of Diabetes & Metabolic Disorders, 23(2), 2289-2314.
13. APTOS 2019 Blindness Detection, https://www.kaggle.com/c/aptos2019-blindness-detection/data
14. Simonyan, K., & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556.
15. Kokhlikyan, N., Miglani, V., Martin, M., Wang, E., Alsallakh, B., Reynolds, J., ... & Reblitz-Richardson, O. (2020). Captum: A unified and generic model interpretability library for pytorch. arXiv preprint arXiv:2009.07896.
16. Zeiler, M. D., & Fergus, R. (2014, September). Visualizing and understanding convolutional networks. In European conference on computer vision (pp. 818-833). Cham: Springer International Publishing.