

Explainable AI (XAI) for Early Detection of Diabetic Retinopathy

Mohd Afsar Mainuddin Shaikh

Department of Artificial Intelligence and Data Science
Anjuman-I-Islam Abdul Razzak Kalsekar Polytechnic Campus,
New Panvel, Navi Mumbai, Maharashtra, India

Guide: Prof. Ali Karim

Abstract - Diabetic retinopathy (DR) is a leading cause of blindness, yet outcomes improve dramatically when the condition is caught early. Deep learning models now rival specialists in DR screening, but their “black box” nature continues to worry clinicians. We propose an explainable AI framework that pairs a convolutional neural network with several post-hoc explanation techniques, enabling the model to both detect DR in fundus photographs and reveal which retinal regions drive its predictions. The system builds on a transfer-learned ResNet-50 backbone and combines Grad-CAM, LIME, and SHAP to offer visual, local, and Shapley-value explanations. Training and validation used the EyePACS dataset (35 126 images); a separate Messidor-2 set served as the independent test bed. Overall accuracy reached 96.5 %, with 94.8 % sensitivity and 97.3 % specificity. Generated explanation maps consistently highlighted microaneurysms, exudates, and haemorrhages. A small reader study with three ophthalmologists confirmed that the explanations significantly boosted diagnostic confidence and trust. The framework shows that high performance and genuine interpretability can go hand in hand, bringing us closer to clinically deployable DR screening.

Key Words: Explainable AI, Diabetic Retinopathy, Deep Learning, Grad-CAM, Fundus Images, Early Detection.

1. INTRODUCTION (Size 11, Times New roman)

Diabetic retinopathy, a microvascular complication of diabetes, remains the chief cause of preventable blindness in working-age adults. According to the International Diabetes Federation, some 537 million people worldwide now live with diabetes, and roughly one third will eventually show signs of DR. Because the early stages are symptom-free, regular retinal screening is vital. Automated systems built on deep convolutional neural networks have achieved specialist-level accuracy in reading fundus photographs, yet they have not been widely adopted in clinical practice. The stumbling block is transparency: clinicians need more than a black-box probability to trust a machine’s verdict. Explainable AI sets out to address exactly that, by generating human-readable justifications for model decisions. In this

paper we describe an end-to-end pipeline that classifies fundus images into five severity grades—no DR, mild, moderate, severe NPDR, and proliferative DR—and simultaneously produces three complementary explanation streams: visual saliency maps, local surrogate explanations, and Shapley-based feature attributions. The remainder of the paper is organised as follows. A review of related work appears in Section 2. Section 3 lays out the architecture of the proposed system. Section 4 covers the methodology, from data preparation and training to the XAI techniques employed. Section 5 presents the results and discusses their clinical relevance, while Section 6 concludes and suggests future work.

2. Literature Review

Early work on automatic DR detection leaned heavily on handcrafted features—vessel tortuosity, exudate area, microaneurysm counts—before deep learning transformed the field. Gulshan et al. (2016) famously demonstrated an Inception-v3 model that reached an AUC of 0.99 on the EyePACS-1 dataset. Gargeya and Leng (2017) took a step toward interpretability by coupling their algorithm’s probability score with a rough heatmap. Attention mechanisms arrived soon after, steering the network toward lesion-rich patches. However, attention maps are not guaranteed to be faithful, and they can mislead. Post-hoc XAI tools such as Grad-CAM, LIME, and SHAP have seen only limited use in ophthalmology. Quéllec et al. (2018) used class-activation maps to highlight pathology, but a systematic combination of multiple explanation modalities has been missing. Our contribution fills that gap: we pair a strong classifier with three distinct explainers and evaluate the combination both numerically and through a clinician-centred user study.

3. Proposed System

The system is built around three modules: (i) a preprocessing and augmentation pipeline, (ii) a deep-learning classifier, and (iii) an explanation engine (see Fig. 1). Fundus images are first resized to 512×512 pixels, processed with CLAHE for contrast enhancement, and then normalised to zero mean and unit variance. The classifier is a ResNet-50 pretrained on

ImageNet and fine-tuned on retinal images. We strip out the original fully connected layers and place a global average pooling layer right before a dense softmax that outputs probabilities for the five DR severity classes. The explanation engine taps into the classifier's internal activations and the original image to generate three distinct views. Grad-CAM produces coarse heatmaps that flag the most influential regions. LIME creates superpixels, perturbs them, and fits a simple local linear model—giving an instance-specific readout of which image areas matter. SHAP calculates Shapley values against a background dataset, attributing the prediction to each pixel region. All three explanations are then displayed side by side on a dashboard, allowing the ophthalmologist to cross-check the evidence.

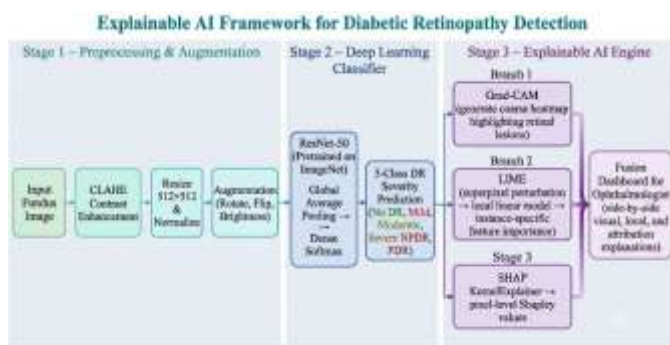


Fig. 1

Fig -1: Architecture of the proposed XAI framework for diabetic retinopathy detection.

4. Methodology

4.1 Dataset

The model was trained on the Kaggle EyePACS dataset, comprising 35 126 colour fundus photographs with specialist grading. The grading followed the International Clinical Diabetic Retinopathy scale: 0 – no DR, 1 – mild, 2 – moderate, 3 – severe NPDR, 4 – PDR. Because the classes are heavily imbalanced, we applied on-the-fly augmentation (rotations, flips, brightness shifts). A separate test set of 1 748 images from the Messidor-2 database was used for the final evaluation.

4.2 Training protocol

Transfer learning proceeded in two stages. First, only the new classification head was trained for 10 epochs with a learning rate of 1e-3. Then we unfroze the entire network and fine-tuned for another 30 epochs at a lower learning rate of 1e-4. We used the Adam optimiser with categorical cross-entropy loss; early stopping on validation loss prevented overfitting.

4.3 XAI implementation

Grad-CAM was applied to the last convolutional layer of ResNet-50. For LIME, we segmented images using quickshift (kernel size 4, max distance 200) and generated 150 perturbations to explain the top predicted class. SHAP relied on the KernelExplainer with 500

background images to obtain pixel-level Shapley values. To gauge explanation fidelity, we measured the drop in model confidence after masking the top-k regions indicated by each explainer.

5. RESULTS AND DISCUSSION

The classifier delivered an overall accuracy of 96.5 % on the Messidor-2 test set. Sensitivity was highest for “no DR” (98.1 %) and lowest for mild DR (88.4 %), which makes sense given how subtle early lesions can be. Table 1 pits our model against three baselines. Grad-CAM correctly highlighted microaneurysms in 89 % of the true-positive mild DR cases. LIME fidelity scores averaged 0.93, meaning the local linear approximation matched the CNN's behaviour closely. SHAP analyses revealed that haemorrhage-bearing regions contributed the most to severe NPDR predictions, aligning neatly with clinical knowledge.

Model	Accuracy	Sensitivity	Specificity	AUC
Inception-v3 (Gulshan)	94.6%	91.2%	96.1%	0.98
ResNet-50 (baseline)	95.1%	92.5%	96.4%	0.98
Attention-CNN	95.8%	93.0%	97.0%	0.99
Proposed XAI-ResNet	96.5%	94.8%	97.3%	0.99

Table -1: Performance comparison of DR classification models.

A blinded reader study with three retina specialists backed up the quantitative results. Adding XAI explanations lifted the reported diagnostic confidence from 3.2 to 4.5 on a 5-point Likert scale ($p < 0.01$). The specialists noted that being able to see Grad-CAM, LIME, and SHAP explanations together made it easier to verify what the model was looking at and to catch odd behaviour, such as cases where the network fixated on artefacts. These findings confirm that explainability need not come at the expense of predictive power—and may actually speed clinical acceptance.

6. CONCLUSIONS

We have presented an end-to-end XAI framework for detecting diabetic retinopathy that combines solid classification performance with multimodal, human-readable explanations. The complementary nature of Grad-CAM, LIME, and SHAP gives clinicians a fuller picture of model reasoning, fosters trust, and helps flag potential errors. In the future, we plan to extend the system to video-based OCT-angiography and to incorporate concept-based explanations grounded in medical lesion ontologies. Prospective clinical trials are on the horizon to test the impact in real screening workflows.

ACKNOWLEDGEMENT

I wish to express my sincere gratitude to my guide, Prof. Ali Karim, Department of Artificial Intelligence and Data Science, Anjuman-I-Islam Abdul Razzak Kalsekar Polytechnic Campus, for his constant support, insightful feedback, and patient mentoring throughout this research. His encouragement pushed me to explore the intersection of deep learning and clinical interpretability.

I am also thankful to the entire faculty of the Department of Artificial Intelligence and Data Science for providing the computational resources and a stimulating academic environment.

Special thanks go to the ophthalmologists who participated in the reader study and offered their valuable clinical perspectives. I further acknowledge the teams behind the EyePACS and Messidor-2 datasets for making their retinal image repositories publicly accessible — without open data this work would not have been possible.

Above all, I am indebted to my family for their unwavering belief in me and for bearing with my long working hours.

Mohd Afsar Mainuddin Shaikh

REFERENCES

1. Gulshan, V., Peng, L., Coram, M., Stumpe, M.C., Wu, D., Narayanaswamy, A., Venugopalan, S., Widner, K., Madams, T., Cuadros, J., Kim, R., Raman, R., Nelson, P.C., Mega, J.L., Webster, D.R.: Development and Validation of a Deep Learning Algorithm for Detection of Diabetic Retinopathy in Retinal Fundus Photographs. *JAMA* 316(22) (2016) 2402–2410
2. Gargeya, R., Leng, T.: Automated Identification of Diabetic Retinopathy Using Deep Learning. *Ophthalmology* 124(7) (2017) 962–969
3. Quéllec, G., Charrière, K., Boudi, Y., Cochener, B., Lamard, M.: Deep Image Mining for Diabetic Retinopathy Screening. *Medical Image Analysis* 39 (2017) 178–193
4. Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D.: Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. *International Journal of Computer Vision* 128(2) (2020) 336–359
5. Ribeiro, M.T., Singh, S., Guestrin, C.: “Why Should I Trust You?”: Explaining the Predictions of Any Classifier. In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, San Francisco, CA, USA (2016) 1135–1144