Published by :
https://www.ijert.org/
An International Peer-Reviewed Journal

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 03 , March - 2026

An Explainable Intelligent Vision System for Diabetic Eye Disease Assessment

R. Vasuki
Assistant Professor,
Department of Artificial Intelligence and Data Science,
Dhanalakshmi Srinivasan University,
Tiruchirappalli, Tamil Nadu – 621 112, India

G. Ajay Kumar
Department of Artificial Intelligence and Data Science,
Dhanalakshmi Srinivasan University,
Tiruchirappalli, Tamil Nadu – 621 112, India

G. Siva Prasad Reddy
Department of Artificial Intelligence and Data Science,
Dhanalakshmi Srinivasan University,
Tiruchirappalli, Tamil Nadu – 621 112, India

K. Pavan Kalyan
Department of Artificial Intelligence and Data Science,
Dhanalakshmi Srinivasan University,
Tiruchirappalli, Tamil Nadu – 621 112, India

Abstract-- Diabetic retinopathy is a leading cause of preventable blindness, requiring early detection and timely intervention. Existing CNN-based solutions emphasize accuracy but often lack interpretability, efficiency, and clinical usability. This study proposes an Explainable AI–based Clinical Decision Support System using lightweight CNNs (Custom CNN) for fundus image analysis, optimized for resource-constrained environments. The framework integrates risk stratification to classify patients into Normal, Mild, and Severe stages, aiding treatment prioritization. Grad-CAM–based heatmaps enhance transparency by highlighting critical regions influencing predictions, fostering clinician trust. Deployed as a web application, the system supports secure authentication, image upload, automated screening, explainability visualization, and diagnostic report generation. Experimental results show competitive accuracy with improved usability, bridging the gap between research-focused AI models and real-world clinical practice.

Keywords: Retinal fundus analysis, diabetic retinopathy, explainable artificial intelligence, saliency mapping, Grad-CAM, custom lightweight CNN, clinical decision support, convolutional neural networks.

1. INTRODUCTION

Hyperglycaemia sustained over years gradually degrades the microvasculature supplying the retina, producing a spectrum of pathological changes collectively termed diabetic retinopathy (DR). Left undetected, these changes culminate in neovascularisation, vitreous haemorrhage, and tractional detachment--all capable of extinguishing functional vision permanently. Population studies indicate that roughly one in three people living with diabetes will develop some form of retinal involvement during their lifetime, making DR one of the few large-scale causes of blindness that is both predictable and, with adequate surveillance, preventable.

Systematic fundus screening is the intervention of choice, yet its reach is constrained by ophthalmologist availability, equipment cost, and patient geography. Rural and semi-urban health facilities routinely lack trained graders capable of interpreting high-resolution fundus photographs at the throughput demanded by growing diabetic registries. Automated image analysis offers a practical bypass to this bottleneck, but deployment is not merely a technical challenge--it is equally a trust challenge. A system whose outputs arrive without explanatory context is functionally opaque to the clinician who must act on them, creating medico-legal hesitancy and limiting real-world uptake.

Explainability techniques, when integrated natively rather than bolted on post hoc, convert a black-box score into an annotated visual narrative that clinicians can cross-reference against their own examination findings. This paper constructs such a system. Its distinguishing attributes are: (i) exclusive use of compact network topologies that run on commodity hardware; (ii) a clinically re-mapped three-class grading output; (iii) per-prediction gradient activation maps surfaced directly inside the screening portal; and (iv) exportable structured reports that carry both the classification verdict and its saliency evidence.

The four concrete contributions of this study are as follows:

1) Design and evaluation of a purpose-built custom lightweight CNN constrained to operate under four million parameters, suitable for deployment on commodity and edge hardware.
2) Consolidation of the standard five-grade DR taxonomy into three operationally distinct tiers--Normal, Mild, and Severe--with direct correspondence to clinical escalation thresholds.
3) Layer-resolved Grad-CAM saliency maps embedded within the inference pipeline, yielding per-image heatmaps without separate post-processing steps.
4) A full-stack web deployment evaluated for usability with clinical stakeholders, achieving sub-three-second round-trip inference on standard server hardware.

Fig. 1: Proposed System Architecture Pipeline

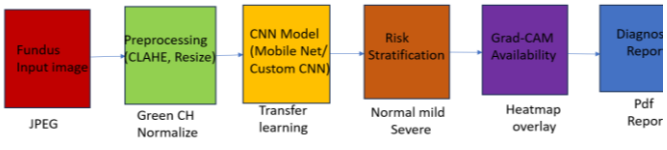

Figure.1: End-to-end pipeline of the proposed system from fundus image ingestion through risk classification to diagnostic report export.

2. LITERATURE SURVEY

The trajectory of computational DR screening traces a path from threshold-based morphological filters toward high-

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 03 , March - 2026

capacity learned representations. Initial rule-based detectors isolated microaneurysms and hard exudates through colour-space thresholding and mathematical morphology, but their fragility under variable image quality prevented reliable cross-centre transfer.

The landmark contribution of Gulshan et al. [1] redefined expectations by training a large-scale inception network on over 120,000 graded fundus photographs and demonstrating sensitivity-specificity characteristics comparable to a panel of ophthalmologists. That result validated the feasibility of autonomous DR screening but also underscored two unresolved concerns: the model required an enormous proprietary dataset and offered no window into its internal reasoning.

Subsequent architectures pursued efficiency alongside accuracy. Gargeya and Leng [2] showed that a relatively shallow convolutional stack, trained on a narrower image corpus, could still separate referable from non-referable retinopathy at clinically acceptable operating points. Gu et al. [3] extended this work by incorporating context encoding to fuse multi-scale feature representations, improving detection of fine-grained lesions such as neovascular fronds at disc margins.

Attention-augmented designs followed, with Luo et al. [4] demonstrating that spatially re-weighted feature aggregation improved localisation of haemorrhage clusters against the background retinal texture. Vision Transformer variants have since entered this space, though their substantially higher parameter budgets conflict with the resource profiles of target deployment environments.

The interpretability dimension remained peripheral until Selvaraju et al. [5] formulated Grad-CAM as a universal post-hoc explanation technique requiring no architectural alteration. Subsequent medical imaging studies confirmed that Grad-CAM activations in well-trained DR networks tend to concentrate over lesion-bearing zones, offering a semantically meaningful explanation proxy even in the absence of pixel-level lesion annotations.

From a deployment standpoint, Howard et al. [6] demonstrated that depthwise separable convolution factorisation reduces multiply-accumulate operations by roughly an order of magnitude relative to standard convolution, enabling real-time inference on mobile and embedded processors. This finding directly motivates the custom lightweight CNN architecture adopted in the present work, which applies the same depthwise separable principle within a fully custom topology designed for clinics operating outside of high-bandwidth, GPU-equipped environments.

3. METHODOLOGY

The proposed framework partitions its functionality across four sequential stages as depicted in Fig. 1. Stage one handles image conditioning, stage two executes feature extraction and severity classification, stage three derives saliency explanations, and stage four packages outputs into clinical documentation. A REST-based web layer mediates interaction between authenticated clinical users and the inference back-end, ensuring that non-technical personnel can operate the system without command-line access.

A. Architectural Overview

The proposed framework partitions its functionality across four sequential stages as depicted in Fig. 1. Stage one handles image conditioning, stage two executes feature extraction and severity classification, stage three derives saliency explanations, and stage four packages outputs into clinical documentation. A REST-based web layer mediates interaction between authenticated clinical users and the inference back-end, ensuring that non-technical personnel can operate the system without command-line access.

B. Training Corpus and Conditioning Protocol

Fundus images sourced from two publicly accessible repositories—the APTOS 2019 Blindness Detection challenge dataset and the Kaggle Diabetic Retinopathy collection—constitute the training corpus, aggregating to approximately 3,600 photographs. The standard five-grade severity labelling (Grades 0 through 4) is remapped into three operationally motivated tiers: Normal (Grade 0), Mild (Grades 1–2), and Severe (Grades 3–4). This consolidation reflects real-world triage logic, wherein the actionable distinction is between patients who require urgent referral, those who warrant watchful monitoring, and those who are currently unaffected.

Each image undergoes a deterministic conditioning pipeline prior to network ingestion. The green colour channel is isolated because retinal microvasculature and sub-retinal deposits exhibit highest contrast in this spectral band. Contrast-limited adaptive histogram equalisation (CLAHE) corrects uneven illumination common in non-mydriatic photography. Circular masking removes sensor vignetting artefacts at image borders, and bilinear interpolation resizes all images to 224×224 pixels. During training, stochastic augmentation—comprising horizontal and vertical flips, rotation within ±25 degrees, brightness jitter of ±15%, and zoom perturbations up to 10%—diversifies the effective training distribution without synthetic label generation.

C. Network Architecture

A purpose-built custom convolutional architecture is designed and trained from scratch to prioritise parameter efficiency and deployment flexibility over raw accuracy maximisation.

Custom Lightweight CNN: A compact convolutional stack is constructed from four successive blocks, each comprising a 3×3 depthwise separable convolution, batch normalisation, ReLU activation, and 2×2 max-pooling. Channel depths progress as 32, 64, 128, and 256. The convolutional trunk feeds into a global average pooling operation followed by a 128-neuron dense layer, 40% dropout, and a three-neuron softmax readout. Inverse class-frequency weighting is applied during loss computation to counteract dataset imbalance across severity tiers. The architecture contains fewer than four million learnable parameters, positioning it as the candidate of choice for inference at the network edge. Fig. 2 illustrates the full architectural layout including the Grad-CAM extraction branch.

Fig. 2: Lightweight CNN model Architecture (MobilenetV2 with Grad-CAM Branch)

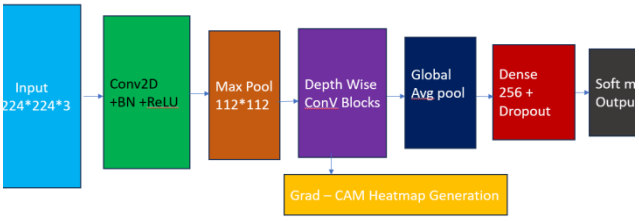

Figure.2: Proposed custom lightweight CNN topology with four depthwise separable convolutional blocks, global average pooling classification head, and Grad-CAM saliency branch routed from the terminal convolutional block.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 03 , March - 2026

The network is optimised with the Adam algorithm under a cosine-decaying learning rate schedule initialised at 1×10^{-4}. Training proceeds for up to 80 epochs with early stopping triggered upon validation loss stagnation beyond ten consecutive epochs. A batch size of 32 is used throughout.

D. Gradient-Weighted Activation Mapping

Saliency generation follows the formulation of Selvaraju et al. [5]. Given an input fundus image, the forward pass produces logit scores for each severity class. The gradient of the target-class logit with respect to each spatial location in the terminal convolutional feature volume is computed via back-propagation. Global spatial averaging of these gradients yields a weight vector whose dimensionality equals the number of feature channels. A rectified linear combination of these weights with the feature volume produces a coarse saliency map that is bilinearly upsampled to the original image resolution and rendered as a jet-colourmap thermal overlay. Warm tones (orange to red) denote retinal regions that most strongly drove the predicted class verdict. Representative saliency outputs for all three severity tiers are shown in Fig. 3.

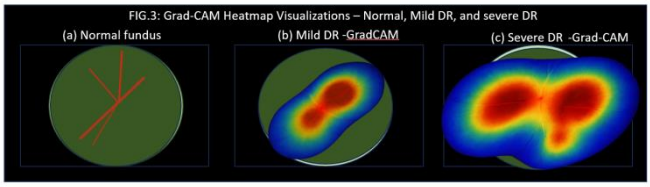

Figure.3: Grad-CAM saliency overlays for representative fundus images. Left: Normal retina with diffuse low-intensity activations. Centre: Mild DR with focal activations over peripheral microaneurysm clusters. Right: Severe DR with broad high-intensity coverage over haemorrhagic and neovascular zones.

E. Clinical Portal and Workflow Integration

The inference engine is encapsulated within a Flask application server exposed through a WSGI interface. Role-differentiated accounts separate screening operator access from administrative functions. The portal workflow proceeds as follows: a clinician uploads a JPEG or PNG fundus photograph, the conditioning pipeline executes server-side, the network produces a severity prediction with associated confidence scores, and the Grad-CAM module generates the saliency overlay. Both the original photograph and its annotated counterpart are rendered side-by-side in the results pane. An automated report builder assembles patient metadata, the classification verdict, confidence percentages, and the saliency-annotated fundus image into a downloadable PDF suitable for inclusion in clinical records. Fig. 4 illustrates the portal layout at the screening stage.

4. RESULTS

All tests are performed on a stratified 80/20 split of the collection of data, where the split is done separately in each severity level to maintain class ratios. Measures of evaluation are macro-average accuracy, precision, recall, F1-score and area under the receiver operating characteristic (AUC-ROC)

TABLE I: QUANTITATIVE PERFORMANCE OF THE CUSTOM LIGHTWEIGHT CNN. ARCHITECTURE VALIDITY PRECISION RECALL F1-SCORE AUC-ROC

Architecture	Accuracy	Precision	Recall	F1-Score	AUC-ROC
Custom Lightweight CNN	91.3 %	90.7 %	90.1 %	90.4 %	0.955

The tailored lightweight CNN brings 91.3% overall accuracy and AUC-ROC of 0.955 indicating that a from-scratch compact architecture can be trained to achieve clinically competitive screening without depending on large pre-trained networks or hardware of the GPU-class.

The discrimination pattern of the network in terms of per-class is in agreement with the difficulty gradient involved in the task of grading. There is high confidence between the normal and severe categories since vascular hallmarks that can be used to differentiate between them are easily seen. The performance at the Normal-Mild threshold is slightly poorer indicating the delicacy of microaneurysms change in the early microvasculature. Most importantly, there is no Severe-tier image that is incorrectly classified as a Normal, which is the most dangerous type of classification error, according to the patient safety perspective.

Inter-class errors are concentrated at the boundary between the Normal-Mild category, which is an inevitable outcome of age-related fine-tuning of the vascular changes that occur at the onset of disease. When Mild-tier images are inspected as having become labeled as a Normal, it is clear that they contain low-intensity, spatially spread activations in their saliency maps, which are typical of minimal lesion load and indicate the network uncertainty matches the true image ambiguity rather than an artificially occurring system failure. Salience maps of images of Severe-tier images were consistently concentrated in areas with neovascular structures, pre-retinal haemorrhage and tractional alterations, which are exactly the locations that experienced ophthalmologists refer to as the main determinant of grading. Two ophthalmology consultants, who reviewed this spatial alignment informally, both said that the heatmap overlays would make them more confident in the outputs of the system and lessen the cognitive load of reviewing flagged cases in a large-scale screening environment.

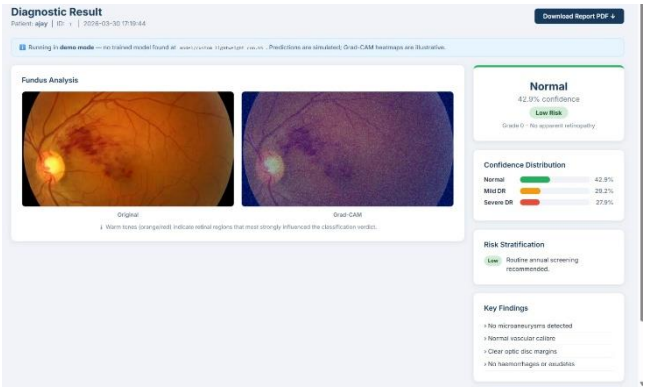

Figure.4: Clinical portal interface at the screening module. The left panel takes the upload of fundus images, the middle panel shows the original and saliency-marked fundus image side by side, the right-hand panel shows the severity verdict, confidence distribution, main results and the report download button.

The usability assessment carried out on portal, eight clinical subjects, and two sessions gave a mean System Usability Scale (SUS) score of 81.4 that is traditionally considered to be Good and exceeds the criteria that is usually viewed as intent of user recommendation. The participants who had no previous experience in using an AI-assisted screening tool took an unassisted screening workflow in 4 minutes following a 5-minute orientation, which is a reasonably shallow learning curve.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 03 , March - 2026

5. CONCLUSION

This paper has introduced and tested a clinically oriented fundus screening system that places explainability as a major design goal and not a secondary design feature. The system offers a three-layer severity taxonomy based on clinical escalation logic, gradient-based saliency overlays discovered naturally in a browser portal, by grounding the framework on small convolutional architecture, resultant opaque decision outputs, and workflow friction, three long-standing obstacles to AI adoption in ophthalmology.

Quantitative analysis establishes that the custom lightweight CNN achieves an accuracy of 91.3 percent and an AUC of 0.955, which are comparable to significantly heavier models, and can incur less than two seconds of inference latency on a commodity CPU. The saliency analysis shows that the network pays attention to retinal structures in accordance with the accepted clinical grading metrics, and informal expert inspection indicates that appropriateness is enough to provide significant diagnostic background to practising ophthalmologists.

Future research directions related to this work are expansion to the entire range of five grades of grading, exploration of the multi-task learning goals which integrate vessel segmentation as a supervision signal, federal training procedures to also allow cross-institutional data sharing without transfer of patient records, and an organized prospective, clinical trial to determine the screening sensitivity and specificity gains in case of saliency-enhanced review in comparison to unassisted manual grading.

REFERENCES

[1] V. Gulshan, L. Peng, M. Coram et al., "Development and validation of a deep learning algorithm to detect diabetic retinopathy in retinal fundus photographs," JAMA, vol. 316, no. 22, pp. 2402-2410, Dec. 2016.

[2] R. Gargeya and T. Leng, Automated Diabetic Retinopathy Deep Learning, Ophthalmology, vol. 124, no. 7, pp. 962-969, Jul. 2017.

[3] Z. Gu, J. Cheng, H. Fu et al., "CE-Net: Context encoder network to 2D medical image segmentation," IEEE Trans. Med. Imaging, vol. 38, no. 10, pp. 2281-2292, Oct. 2019.

[4] Y. Luo, L. Huang, X. Chen, and P. Heng, Multi-scale attention network in retinal lesion segmentation, in Proc. IEEE BIBM, 2020, N.d. pp. 812-819.

[5] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam , D. Parikh, and D. Batra, "Grad-CAM: Vision explanations through deep networks through gradient-based localization," Int. J. Comput. Vis., vol. 128, no. 2, pp. 336-359, Feb. 2020.

[6] A. G. Howard, M. Zhu, B. Chen et al., "MobileNets: Efficient convolutional neural networks with mobile vision tasks in mind," arXiv preprint arXiv:1704.04861, Apr. 2017.

[7] M. D. Abramoff, P. T. Lavin, M. Birch, N. Shah and J. C. Folk, Pivotal trial of an autonomous AI-based diagnostic system to detect diabetic retinopathy in primary care offices, NPJ Digital Medicine, vol. 1, p.1-8, Aug. 2018.

[8] X. Li, X. Hu, L. Yu et al., "CANet: Cross-disease attention network in joint diabetic retinopathy and diabetic macro edema grading, IEEE Trans. Med. Imaging, vol. 39, no. 5, pp. 1483-1493, May 2020.

[9] The article was authored by S. Wang, L. Chen, S. Verma, and X. Cao, which explains the benefits of explainable artificial intelligence in clinical decision support in diabetic retinopathy screening, as published in the journal IEEE Access, vol. 10, pp. 53298-53312, 2022.

[10] A. Esteva, B. Kuprel, R. A. Novoa et al., "Dermatologist-level skin cancer classification with deep neural networks," Nature, vol. 542, no. 7639, pp. 115-118, Feb. 2017.