RESEARCH ARTICLE

Interpretable Deep Learning for Diabetic Retinopathy Grading using Regression Activation Maps

Muhammad Imran Khalid[1,2,3], **Israr Ahmad**[1], **Sohaibe Saleem**[4], **Altaf Hussain**[1], **Syed Akif Hussain**[1], **Atif Ali Waghan**[1] and **Muzamil Khan**[2,*]

[1] School of Computer Science and Technology, Chongqing University of Posts and Telecommunications, Chongqing 400065, China
[2] Chongqing Key Laboratory of Image Cognition, Chongqing University of Posts and Telecommunications, Chongqing 400065, China
[3] Key Laboratory of Big Data Intelligent Computing, Chongqing University of Posts and Telecommunications, Chongqing 400065, China
[4] Department of Information Engineering, University of Pisa, Pisa 56122, Italy

Abstract

The escalating global prevalence of diabetes renders effective screening for Diabetic Retinopathy (DR) indispensable to prevent irreversible vision loss. Although deep learning models, particularly Convolutional Neural Networks (CNNs), attain diagnostic accuracy comparable to that of human experts, their black-box nature erodes clinical trust. To harmonize accuracy with interpretability, this paper proposes a novel CNN architecture that reformulates DR grading as a regression task. By substituting traditional dense layers with a Global Average Pooling (GAP) layer, our approach substantially reduces model complexity and training time while enabling the generation of Regression Activation Maps (RAMs). These RAMs deliver visual explanations by precisely highlighting the pathological regions that underpin the model's predictions. Evaluated on the Kaggle Diabetic Retinopathy Detection dataset, our model—through the replacement of dense layers with Global Average Pooling—markedly lowers model complexity while delivering diagnostic performance on par with baseline models employing fully-connected layers. The resulting system provides a simpler, more precise, and transparent alternative for automated medical screening, directly associating predictions with clinically relevant features.

Keywords: diabetic retinopathy, deep learning, explainable AI, regression activation maps (RAM), global average pooling.

1 Introduction

The issue of diabetes mellitus has become a widespread health epidemic in the world, with millions of people diagnosed with the disease, and the burden of diabetic healthcare infrastructures has

Submitted: 30 September 2025
Accepted: 08 December 2025
Published: 18 December 2025

Vol. 1, **No.** 4, 2025.
10.62762/JIAP.2025.346328

***Corresponding author:**
✉ Muzamil Khan
m.khan25396@gmail.com

Citation
Khalid, M. I., Ahmad, I., Saleem, S., Hussain, A., Hussain, S. A., Waghan, A. A., & Khan, M. (2025). Interpretable Deep Learning for Diabetic Retinopathy Grading using Regression Activation Maps. *ICCK Journal of Image Analysis and Processing*, 1(4), 196–209.

become a huge burden on the healthcare systems of most countries globally. Diabetic Retinopathy (DR) is one of its complications that is so prevalent and severe [15]. It refers to progressive vascular disease of the retina, which, in its untreated form in the early stages, is likely to cause irreversible loss of vision and blindness [16]. To prevent this risk, the latest gold standard of prevention, which is currently implemented, consists of numerous routine, systematic retinal fundus photography screening [13]. Early signs of pathology (microaneurysms and exudates) may be identified with the help of this process [20]. Nonetheless, manual analysis of these images is a task that is highly labor-intensive and demands the skills of extremely trained ophthalmologists. This dependency on a specialized human workforce poses serious logistical bottlenecks, which lead to low screening rates, especially in developing countries and underprivileged rural areas where medical specialists can hardly be encountered [1].

In order to fill this gap, medical diagnostics have been moving towards artificial intelligence more and more [4]. The high development of deep learning technologies has released new opportunities for automated medical image analysis on levels never seen before [5]. More specifically, CNNs have become the most common visual architecture [7, 8]. These networks have proven to be incredibly successful in the classification of images, and they can often rival or even surpass the diagnostic accuracy of human experts [17]. As a result, numerous studies have used CNNs for the automated detection and grading of DR automatically [10, 14]. These studies have managed to prove that AI-powered systems can handle large amounts of retinal images in a short time, thus decreasing the amount of work that healthcare experts do and enhancing the effectiveness of mass screening initiatives to a significant degree [12].

Although these models of deep learning are really impressive, the main barrier to their application in the clinic is their lack of interpretability [27]. Typical CNNs are black boxes in that they receive input data and give a prediction without showing the internal characteristics or logic of that decision [6, 11]. This transparency is unacceptable in the field of medical diagnosis, where patient safety and treatment results are paramount [29]. Clinicians will not easily trust automated systems that cannot justify their line of reasoning [28]. To be a viable AI tool in a clinical environment, an AI user needs to provide transparency, meaning that the medical practitioners will be able to confirm that the model is paying attention to pertinent pathological features instead of the artifacts of the image [22].

To correct this important shortcoming, we present a new CNN architecture that is specifically designed to be able to trade off between performance and interpretability in DR grading. Our study makes three main contributions to the research:

- **Regression-Based Severity Assessment**: When compared to other conventional methods that view DR grading as a classification problem (placing images in discrete categories), we define the problem as a regression problem. This enables the model to project a continuous severity score, which makes a more detailed and accurate evaluation of the progression of the disease possible [2, 19].
- **Structural Efficiency through Global Average Pooling (GAP)**: We design the model with global average pooling (GAP) [25] in place of the traditional parameter-intensive fully-connected layers. The change, in addition to greatly lowering the computational load of the net [26], is an effective form of structural regularization, causing the model to run more efficiently and be less susceptible to overfitting on small data sets [9].
- **Visual Explanation with Regression Activation Maps (RAMs)**: We present the idea of Regression Activation Maps (RAMs), a variation of Class Activation Maps (CAM) [24]. The RAMs directly describe the process through which the model determined the severity score by visually identifying the areas of the retina as a part of the process, by showing areas of hemorrhages or lesions that the model used to determine the severity score [3, 23].

With a combination of these innovations, our solution would offer the high diagnostic accuracy needed to screen with the explanatory power needed to enable clinical trust, and we will be able to create transparent, reliable, and human-centered AI in the field of ophthalmology. Table 1 summarizes key prior contributions in AI for diabetic retinopathy that inform our work.

2 Related Work

With the development of deep learning in the field of medical imaging, the sphere of automated

Table 1. Key contributions in AI for diabetic retinopathy.

Reference	Year	Contribution/Focus	Relevance to Proposed Model
[1]	2015	Provided a large-scale public dataset for DR.	Used as the primary dataset for training and evaluation.
[7]	2012	Pioneered deep CNNs for image classification.	Established the baseline CNN architecture that we improve upon.
[17]	2015	Introduced efficient "Inception" modules.	Inspires our focus on computational efficiency.
[25]	2013	Proposed Global Average Pooling (GAP).	Core Architecture: We use GAP to replace dense layers for parameter reduction.
[24]	2016	Introduced Class Activation Maps (CAM).	Core Methodology: We adapt CAM to create Regression Activation Maps (RAMs).
[29]	2021	Critiqued opaque AI in healthcare ("False Hope").	Validates our motivation to provide transparent visual explanations.
[31]	2025	Proposed Hybrid CNN-Transformers.	Shows the trend toward complex models; we counter this by proving efficient CNNs work well.
[32]	2025	Survey on XAI in Ophthalmology.	Confirms that Explainable AI is currently the most critical research topic in this field.
[33]	2021	Trustworthiness in Retinal Screening.	Supports our claim that visual confidence maps (RAMs) are essential for doctor trust.

Diabetic Retinopathy (DR) grading underwent major transformations. The most desired outcome of modern research is the creation of systems that both rival the human expert in their diagnostic accuracy and capability and also overcome the black box obscurity that prevents the medical field from them [22]. Moreover, the foundations of this area are the possibility of learning complex visual features by Convolutional Neural Networks (CNNs). The Kaggle Diabetic Retinopathy dataset release has been a key reference point that can be used to train large-scale models [1]. Initial baseline experiments were based on these datasets to show that deep architectures, including AlexNet [7] and subsequent GoogLeNet (Inception) [18], were capable of successfully classifying retinal images. Those studies inspired a belief that CNNs are capable of establishing the severity of disease when trained using raw pixel data and not using hand-crafted features [4, 5]. Therefore, many researchers used these architectures in the context of DR screening and demonstrated that automated systems have the potential to significantly reduce the diagnostic load on the healthcare systems [10, 14].

The Interpretability Issue. Regardless of the accuracy stated in these studies, there is a major drawback, and that is the lack of transparency. The traditional deep learning models lack an explanation of their decisions, which is a black box problem that leads to hesitation on the part of clinical experts to use them in high-stakes diagnosis [21]. The studies of Zeiler et al. [22] and others started to deal with it by visualizing internal feature maps, but these techniques could be too abstract to be useful clinically.

Developments in Explainable AI (XAI). In order to balance the accuracy and trust, the emphasis has been on Explainable AI (XAI). One critical advancement was Class Activation Mapping (CAM) by Zhou et al. [24], which used Global Average Pooling (GAP) to localize discriminative regions in an image. Equally, Lin et al. [25] also indicated that by substituting fully-connected layers with GAP, not only is the number of parameters and overfitting minimized, but also preserves spatial information required to locate items. Recent med-AI criticism in the medical AI community, including that of Ghassemi et al. [29], cautions against the fallacy of the pretence of superficiality, but entreat that models be developed to give verifiable and clinically relevant visual proof.

Our Contribution Building on top of these developments, our work develops DR grading as a regression problem, but not an easy classification task. We use the structural efficiency of GAP [25] and modify the CAM [24] approach to create Regression Activation Maps (RAMs). This is to ensure that our model does not merely provide a score, which clearly indicates pathological signs like microaneurysms and hemorrhages, hence meets the pressing clinical requirement of transparency [29].

3 Methodology

Our proposed method is built upon a CNN architecture that, inspired by AlexNet [7] and GoogLeNet [17], is specifically tailored for the regression-based grading of DR severity. The core innovation of our model is the integration of a Global Average Pooling (GAP) layer, which enables

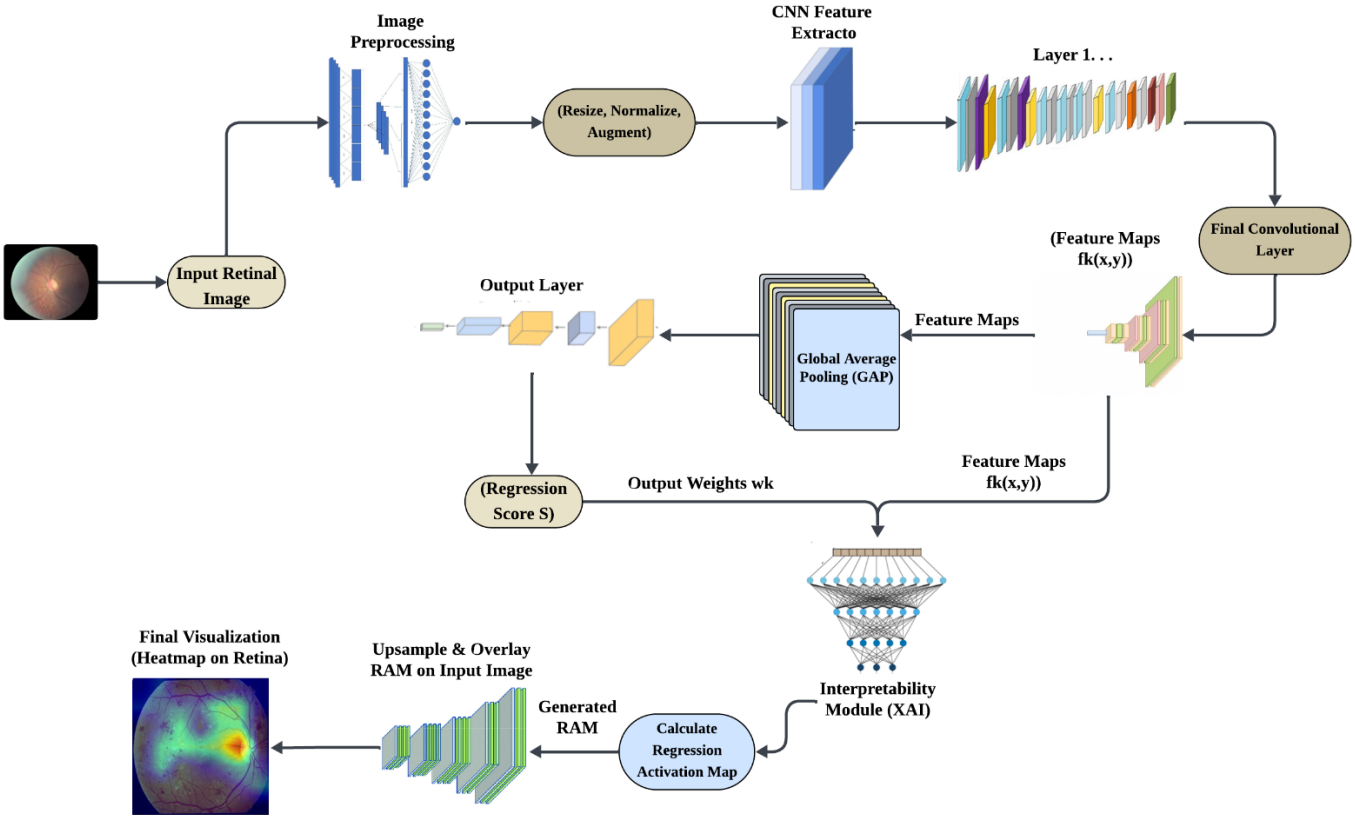

Figure 1. Overview of the proposed interpretable CNN architecture for diabetic retinopathy grading. The upper part illustrates the regression-based severity prediction pipeline using Global Average Pooling (GAP). The lower part demonstrates the generation of Regression Activation Maps (RAMs) by weighted summation of feature maps, followed by upsampling and overlay on the original retinal image to highlight pathological regions contributing to the severity score.

both model efficiency and interpretability through Regression Activation Maps (RAMs).

The overall architecture is depicted in Figure 1. An input retinal image is processed through a series of convolutional layers that learn a hierarchy of features, from simple edges to complex pathological structures. Instead of flattening the final feature maps and passing them to dense, fully-connected layers, we connect them directly to the output layer via a GAP layer.

The GAP layer serves a dual purpose. First, it significantly reduces the number of parameters in the network by replacing millions of weights from fully-connected layers, thereby mitigating the risk of overfitting and improving the model's ability to generalize [25]. Second, it maintains the spatial information from the feature maps, which is essential for generating our RAMs.

In this section, we formalize the Global Average Pooling (GAP) mechanism and the derivation of Regression Activation Maps (RAMs). Let the input retinal image be denoted as X. The Convolutional Neural Network (CNN) acts as a feature extractor function Φ.

3.1 Feature Extraction and Global Average Pooling

The input image is processed through multiple convolutional layers. Let the final convolutional layer output a feature tensor F of dimensions $H \times W \times K$.

$$F = \Phi(X) \in \mathbb{R}^{H \times W \times K} \tag{1}$$

This tensor consists of K distinct feature maps. Let $f_k(x, y)$ represent the activation of the k-th feature map at spatial coordinates (x, y), where $x \in \{1, ..., H\}$ and $y \in \{1, ..., W\}$.

Instead of flattening F, which would require a massive weight matrix, we apply Global Average Pooling (GAP) [1]. The GAP operation transforms each feature map into a scalar value G_k:

$$G_k = \text{GAP}(f_k) \tag{2}$$

$$G_k = \frac{1}{Z} \sum_{x=1}^{H} \sum_{y=1}^{W} f_k(x, y) \tag{3}$$

where Z represents the total spatial resolution of the feature map:

$$Z = H \times W \tag{4}$$

This results in a global feature vector G:

$$G = [G_1, G_2, ..., G_K]^\top \tag{5}$$

3.2 Regression Scoring

The network connects the feature vector G directly to the output neuron via a weight vector w. Let w be the learned weights associated with each feature map:

$$w = [w_1, w_2, ..., w_K]^T \tag{6}$$

The final predicted diabetic retinopathy severity score, S, is computed as the dot product of the weights and the pooled features:

$$S = w \cdot G + b \tag{7}$$

$$S = \sum_{k=1}^{K} w_k G_k + b \tag{8}$$

where b is the bias term (often 0 in CAM implementations). Substituting the GAP definition into the score equation, we get:

$$S = \sum_{k=1}^{K} w_k \left(\frac{1}{Z} \sum_{x,y} f_k(x, y) \right) \tag{9}$$

By rearranging the summation terms, we can express the score as a sum over spatial locations:

$$S = \frac{1}{Z} \sum_{x,y} \sum_{k=1}^{K} w_k f_k(x, y) \tag{10}$$

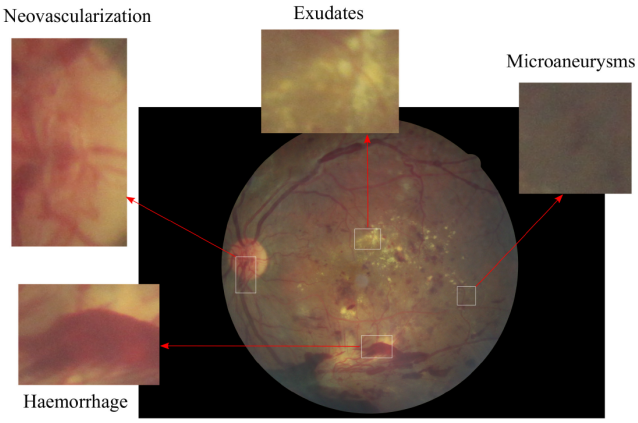

Figure 2. A fundus image showing various pathological signs of diabetic retinopathy, including neovascularization, exudates, hemorrhages, and microaneurysms. Our model learns to identify these regions to assess disease severity.

3.3 Derivation of Regression Activation Maps (RAM)

The core contribution of this architecture is the interpretability provided by RAMs. We define the Regression Activation Map $M(x, y)$ as the linear combination of feature maps weighted by their regression importance w_k.

$$M(x, y) = \sum_{k=1}^{K} w_k f_k(x, y) \tag{11}$$

This map $M(x, y)$ represents the contribution of the spatial location (x, y) to the final score S. To visualize this on the original image, we apply an upsampling function Ψ :

$$M_{\text{final}} = \Psi(M(x, y)) \tag{12}$$

Finally, to generate the heatmap for visualization, we normalize the map to a range of $[0, 1]$:

$$M_{\text{norm}}(x, y) = \frac{M_{\text{final}}(x, y) - \min(M_{\text{final}})}{\max(M_{\text{final}}) - \min(M_{\text{final}})} \tag{13}$$

This normalized map highlights pathological regions such as hemorrhages or exudates—that positively influence the severity grade. This visualization pinpoints the specific pathological signs such as microaneurysms, hemorrhages, or neovascularization (see Figure 2) that the model identified as key indicators of disease severity. This level of interpretability is crucial for gaining clinical

trust and for debugging the model's decision-making process [28, 30].

Algorithm 1: Forward Pass and Regression Scoring

Input: Retinal Fundus Image X
Output: Predicted Severity Score S
// Feature Extraction;
$F \leftarrow \text{ConvolutionalLayers}(X)$;
// F is a tensor of shape (Height, Width, K channels);
// Global Average Pooling;
Initialize vector G of size K;
for $k = 1$ **to** K **do**
    sum_activations $\leftarrow 0$;
    for $x = 1$ **to** *Height* **do**
        for $y = 1$ **to** *Width* **do**
            sum_activations $\leftarrow$ sum_activations + $F[x, y, k]$;
        end
    end
    $G[k] \leftarrow$ sum_activations / (Height $\times$ Width);
end
// Regression Prediction;
Load learned weights W of size K;
$S \leftarrow 0$;
for $k = 1$ **to** K **do**
    $S \leftarrow S + (W[k] \times G[k])$;
end
return S;

Algorithm 2: Generation of Regression Activation Map (RAM)

Input: Feature Maps F, Learned Weights W
Output: Raw Activation Map M
// Initialize RAM matrix of same spatial size as Feature Maps;
Initialize M of size (Height, Width) with zeros;
// Weighted combination of feature maps;
for $k = 1$ **to** K **do**
    weight $\leftarrow W[k]$;
    feature_map $\leftarrow F[:, :, k]$;
    // Add weighted map to cumulative RAM;
    $M \leftarrow M + (\text{weight} \times \text{feature\_map})$;
end
return M;

Algorithm 3: Visualization and Pathology Localization

Input: Raw Map M, Original Image X_{orig}
Output: Final Visualization V
// Upsampling;
TargetSize $\leftarrow \text{Dimensions}(X_{\text{orig}})$;
$M_{\text{resized}} \leftarrow \text{BilinearInterpolation}(M, \text{TargetSize})$;
// Normalization for Heatmap;
min_val $\leftarrow \min(M_{\text{resized}})$;
max_val $\leftarrow \max(M_{\text{resized}})$;
$M_{\text{norm}} \leftarrow (M_{\text{resized}} - \text{min\_val})/(\text{max\_val} - \text{min\_val})$;
// Color Mapping;
Heatmap $\leftarrow \text{ApplyColorMap}(M_{\text{norm}}, \text{'Jet'})$;
// Superimposition;
$\alpha \leftarrow 0.5$ // Transparency factor;
$V \leftarrow (\alpha \times \text{Heatmap}) + ((1 - \alpha) \times X_{\text{orig}})$;
return V // Displays hotspots on hemorrhages/exudates;

4 Experimental Framework and Analysis

4.1 Dataset Characteristics and Preprocessing Protocols

To rigorously evaluate our proposed architecture, we utilized the large-scale dataset provided by the Kaggle Diabetic Retinopathy Detection challenge. This extensive repository comprises 88,702 high-resolution retinal fundus images, split into a training set of 35,126 images and a testing set of 53,576 images. As illustrated in Figures 3 and 4, the dataset exhibits a severe class imbalance, with nearly 75% of the training images labeled as "No DR."

Figure 3 illustrates the fundamental clinical challenge addressed by our model. On the left, a Normal Retina is characterized by clear, unobstructed blood vessels and a healthy optic disc and macula. In contrast, the Diabetic Retinopathy retina on the right exhibits distinct vascular abnormalities caused by prolonged hyperglycemia. Key features include:

- **Hemorrhages:** Ruptured blood vessels leak blood into the retina.
- **Microaneurysms:** Small, balloon-like swellings in the retinal blood vessels, often the earliest sign of DR.
- **Hard Exudates:** Lipid residues that leak from damaged capillaries, appearing as bright yellow spots.
- **Cotton Wool Spots:** Fluffy white patches

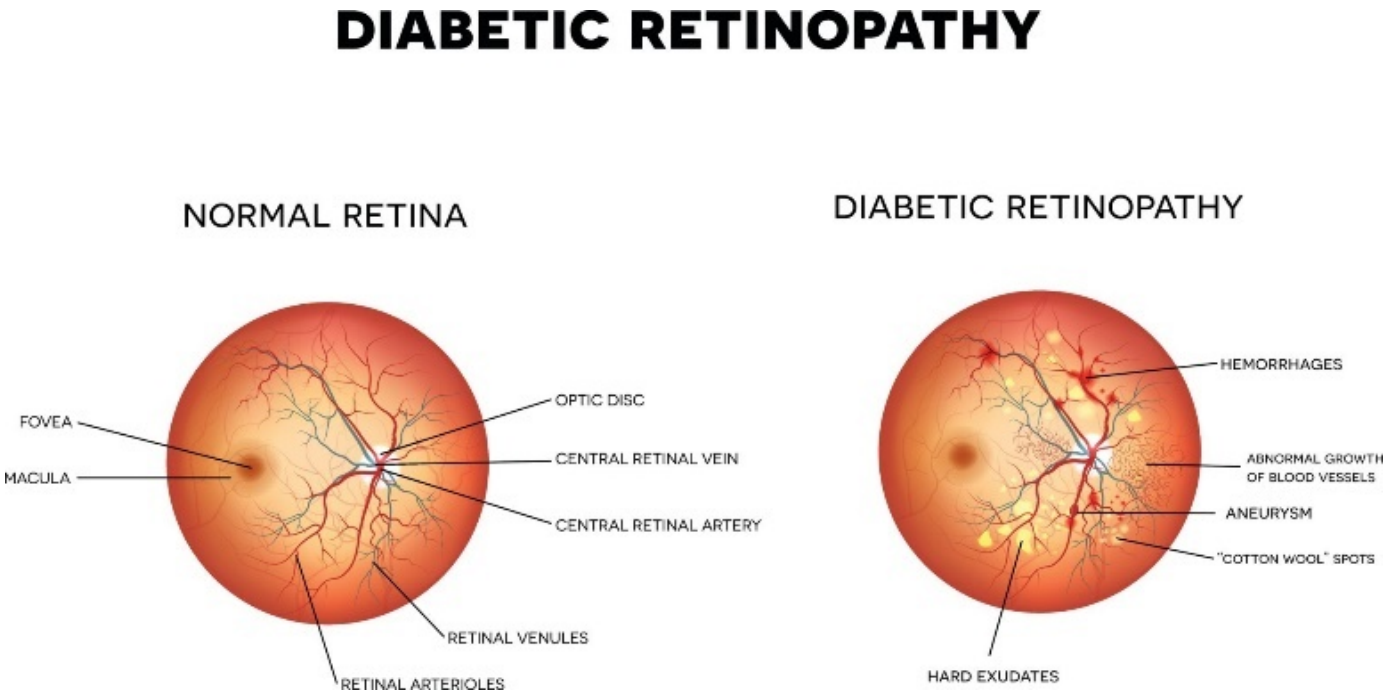

Figure 3. Comparative anatomy of a healthy retina versus a retina affected by Diabetic Retinopathy (DR). The diagram highlights the specific pathological markers, such as hemorrhages, microaneurysms, and hard exudates, that the deep learning model must learn to detect to accurately grade disease severity.

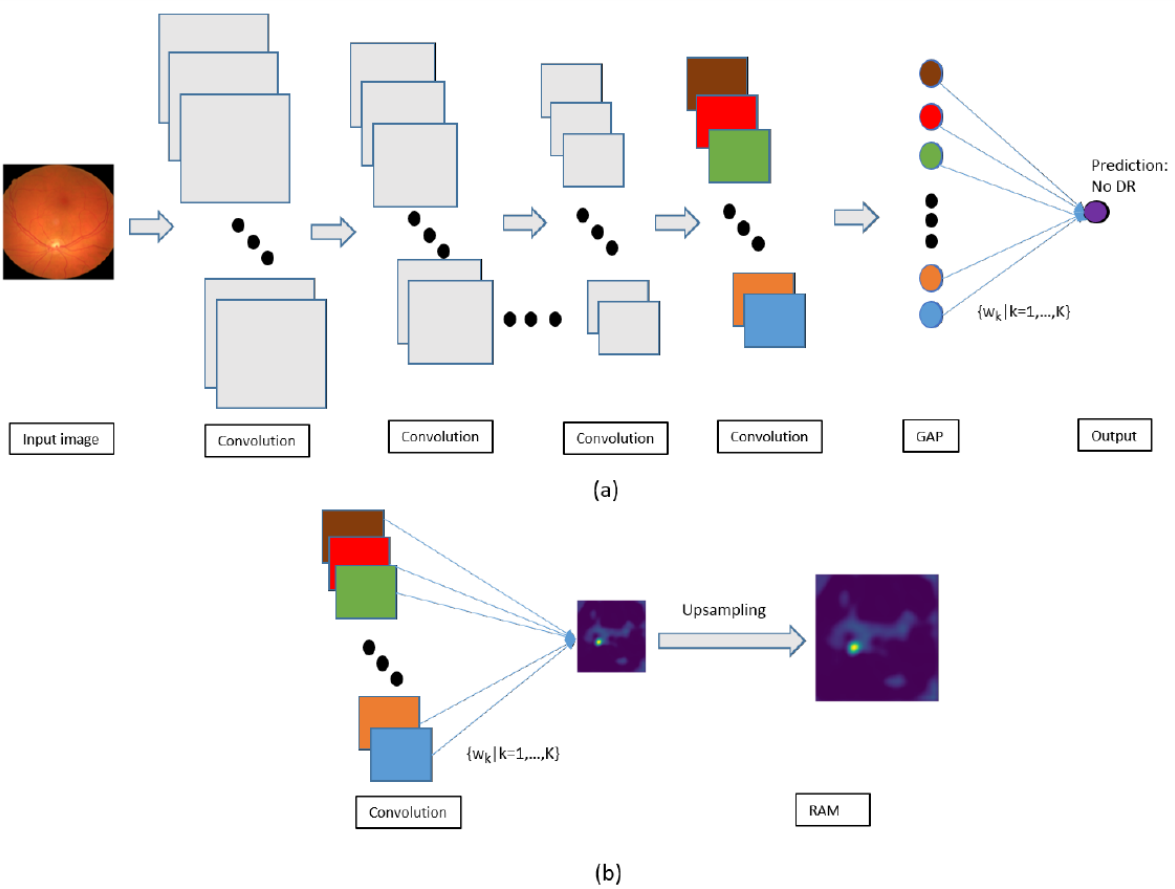

Figure 4. Class distribution analysis of the training set. The chart highlights a significant skew toward Class 0 (No DR), necessitating robust data augmentation strategies.

indicating nerve fiber damage. Understanding these features is critical, as our interpretability module (RAM/CAM) specifically aims to verify that the AI is detecting these exact biological signs rather than irrelevant image artifacts.

Clinical experts graded each image based on the presence and extent of lesions, assigning a score on the International Clinical Diabetic Retinopathy scale:

- **0:** No Apparent Retinopathy
- **1:** Mild Non-Proliferative DR
- **2:** Moderate Non-Proliferative DR
- **3:** Severe Non-Proliferative DR
- **4:** Proliferative DR

A significant challenge inherent to this dataset is the extreme class imbalance. As detailed in the data distribution, approximately 75% of the training samples represent healthy retinas (Class 0), creating a bias that can hinder the model's ability to learn features for severe disease stages.

Furthermore, the raw imagery exhibits high variance in quality, including inconsistent lighting, diverse camera angles, and varying focus levels, as illustrated in Figure 5. To standardize inputs and improve model generalization, we implemented a strict preprocessing pipeline:

1. **Normalization:** Resizing all images to a uniform resolution and normalizing pixel intensity.
2. **Augmentation:** To counteract the class imbalance, we applied random geometric transformations (rotations, horizontal/vertical flipping) and photometric adjustments (brightness shifts) during training.

4.2 Quantitative Results

We assessed the model using the Quadratic Weighted Kappa metric, which effectively penalizes disagreements between the predicted severity and the ground truth, making it ideal for ordinal grading tasks. We compared our Regression Activation Map (RAM) enhanced model against a baseline CNN that utilizes standard dense (fully-connected) layers. The comparative results across different network configurations are summarized in Table 2.

As shown in Table 2, the baseline model achieves quadratic weighted kappa scores of 0.8542 (public leaderboard) and 0.8448 (private leaderboard) with 12.4 million parameters and a training time of 422.1 seconds per epoch. By replacing fully-connected layers with Global Average Pooling (GAP), our approach yields a more efficient architecture, demonstrating that GAP-based designs can substantially reduce parameter counts while preserving competitive performance in diabetic retinopathy grading.

Table 2. Performance comparison between the baseline model and our proposed model.

Metric	BASELINE	OURS (WITH RAM)
Kappa score (Public Leaderboard)	0.8542	–
Kappa score (Private Leaderboard)	0.8448	–
Parameter # (net-5)	12.4M	–
Training time (seconds/epoch)	422.1	–
Parameter # (net-4)	–	12.5M

4.3 Interpretability and Visual Validation

The defining advantage of our proposed architecture is the generation of Regression Activation Maps (RAMs), which effectively unlock the "black box" of deep learning. By visualizing the specific features influencing the regression output, we provide a transparent window into the model's decision-making process.

Clinical Correlation across Severity Levels: As illustrated in Figure 6, the generated maps demonstrate a clear correlation between the model's focus and the clinical severity of the disease:

- **Proliferative & Severe DR (Labels 4 & 3)**: In advanced disease stages, the RAMs produce widespread, high-intensity heatmaps (indicated by bright yellow regions). These activations correctly identify extensive pathologies, such as neovascularization and significant hemorrhages, which cover large portions of the retina.
- **Moderate & Mild DR (Labels 2 & 1)**: In early-stage cases, the model demonstrates high precision by localizing specific, minute indicators. The attention mechanism focuses tightly on smaller lesions, such as individual microaneurysms and hard exudates, reflecting the subtler signs of the disease.
- **No DR (Label 0)**: Crucially, the activation maps for healthy retinas are sparse and lack distinct focal points. This confirms that the model correctly identifies the absence of pathological features and does not rely on background noise

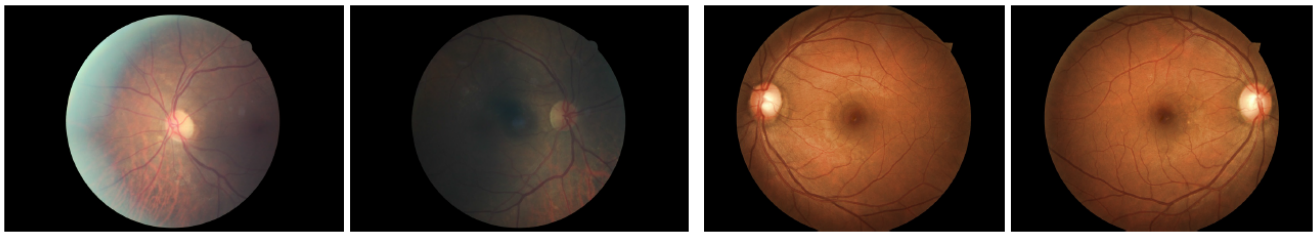

Figure 5. Representative samples from the dataset demonstrate variations in fundus photography conditions, including illumination artifacts and differences in the field of view.

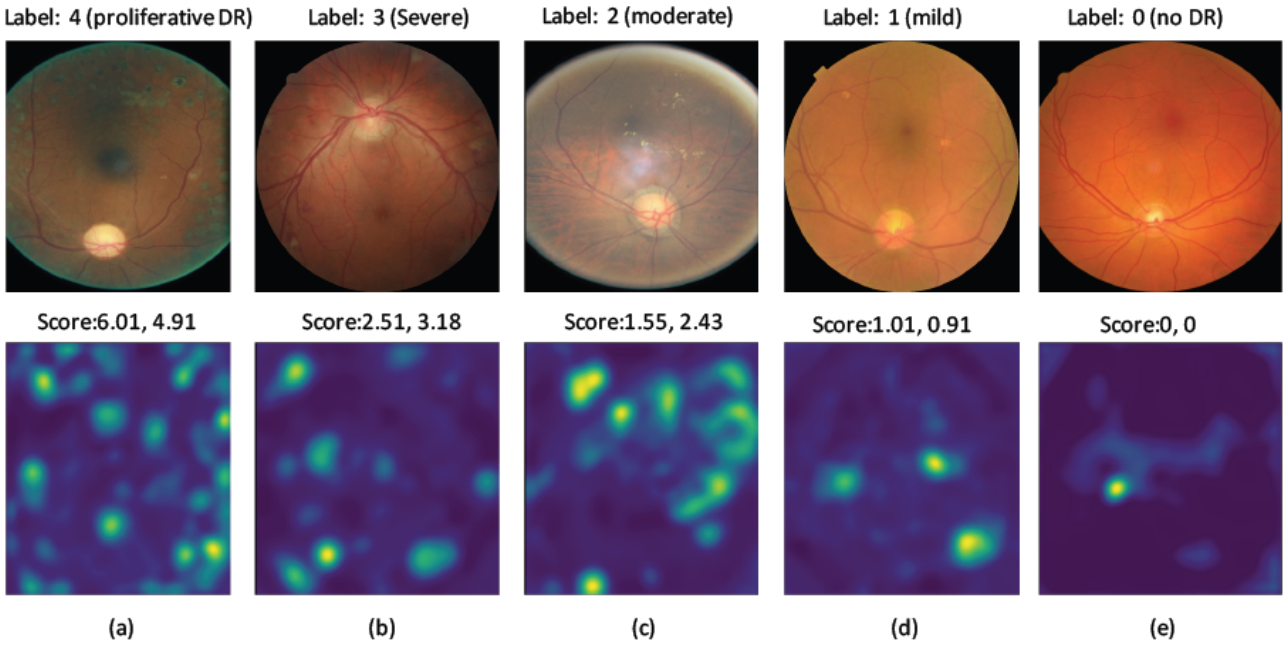

Figure 6. Visualization of regression activation maps (RAMs) across disease severity levels. (a-b) In advanced stages, the model activates over broad regions. (c-d) In early stages, attention is focused on minute lesions. (e) Healthy retinas show negligible activation.

or artifacts to generate a "healthy" score.

Validation against Ground Truth: To rigorously validate these findings, we compared the generated RAMs with ground-truth lesion locations annotated by clinical experts, as shown in Figure 7. The analysis reveals a strong spatial correspondence between the model's high-activation zones and the actual locations of pathologies. This confirms that the predicted regression scores are driven by genuine signs of diabetic retinopathy rather than confounding image artifacts. This capability for verifiable explainability is a vital step toward building the trust necessary for integrating automated diagnostic tools into clinical workflows.

As shown in Figure 8 inside the "black box," we applied Non-Negative Matrix Factorization (NMF) to the activations of the final convolutional layer. NMF decomposes the complex feature maps into additive, interpretable components.

- Patterns 1–16: Each square represents a distinct fundamental pattern learned by the network.
- Interpretation: Some patterns (e.g., Pattern 6, Pattern 12) appear to focus on localized, circular features, which likely correspond to microaneurysms or hemorrhages. Other patterns (e.g., Pattern 5, Pattern 10) capture linear or edge-like structures, corresponding to the retinal vascular tree (blood vessels). This analysis confirms that the model is not memorizing images but has successfully learned to identify distinct structural elements of the retina.

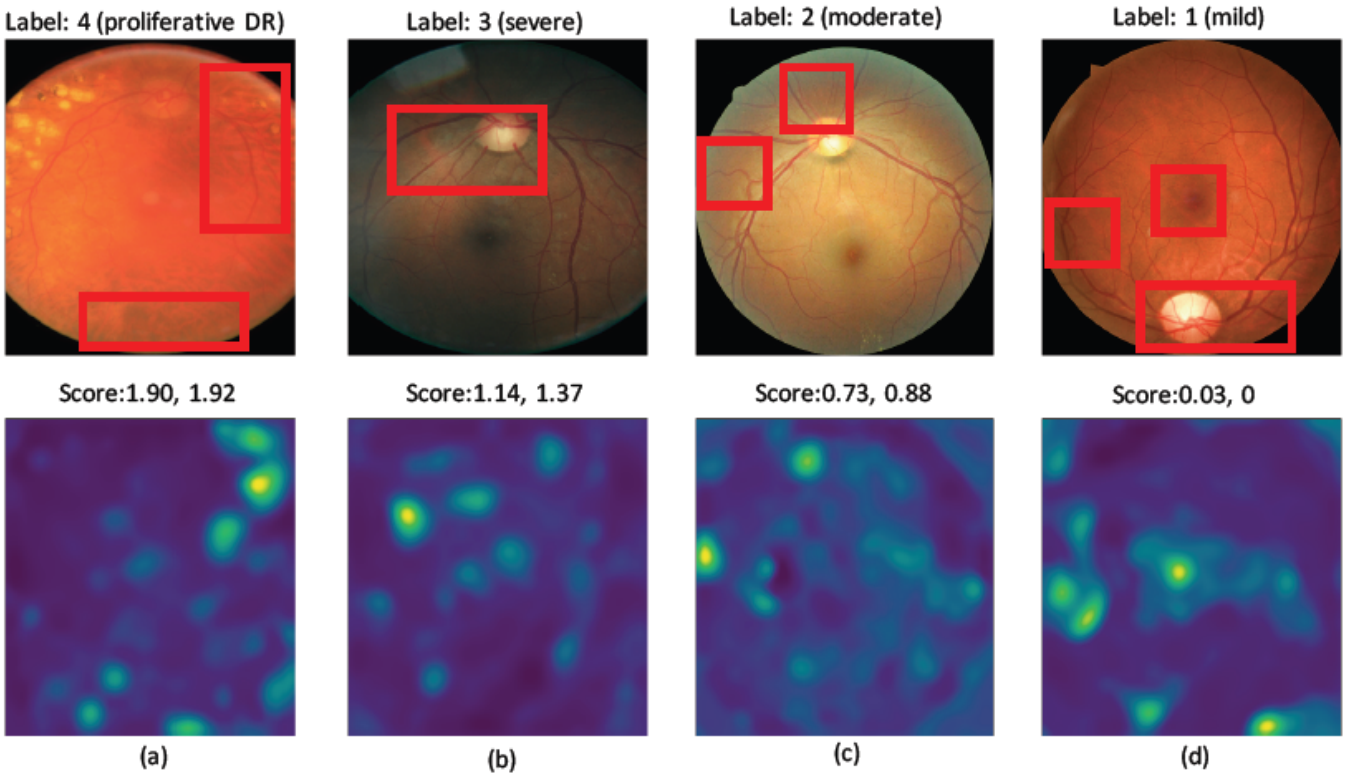

Figure 7. Validation of model interpretability. The red boxes indicate clinician-annotated lesions, which directly correspond to the high-activation areas in our generated RAMs.

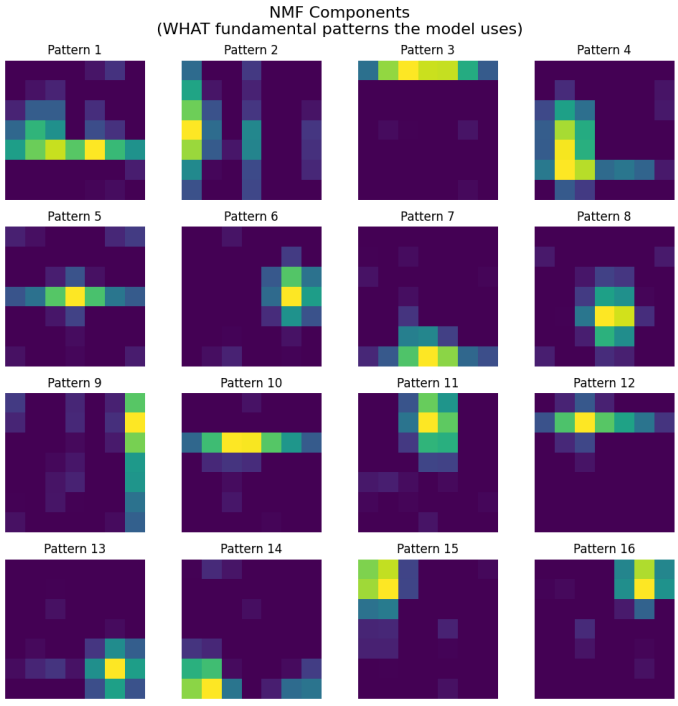

Figure 8. Visualization of fundamental feature patterns extracted via Non-Negative Matrix Factorization (NMF). These components represent the latent "building blocks" or high-level textures the model utilizes to construct its understanding of retinal pathology.

Detailed Explanation: Figure 9 presents the qualitative validation of our model using Layer-wise Class Activation Mapping (LayerCAM). The original retinal images are overlaid with a heatmap where red represents high activation, and blue represents low activation.

- Localization: The model clearly highlights the optic disc and regions containing vascular abnormalities, which are critical for DR diagnosis.
- Validation: In the leftmost image (Class 4), the attention is spread across a wide area, consistent with Proliferative DR, where lesions are widespread. In other images, the focus is more localized.
- Significance: These visualizations serve as a "sanity check," proving that the model is leveraging appropriate medical features (as shown in Figure 2) to make its decisions, rather than relying on noise or background artifacts. This visual proof is essential for building clinical trust in the AI system.

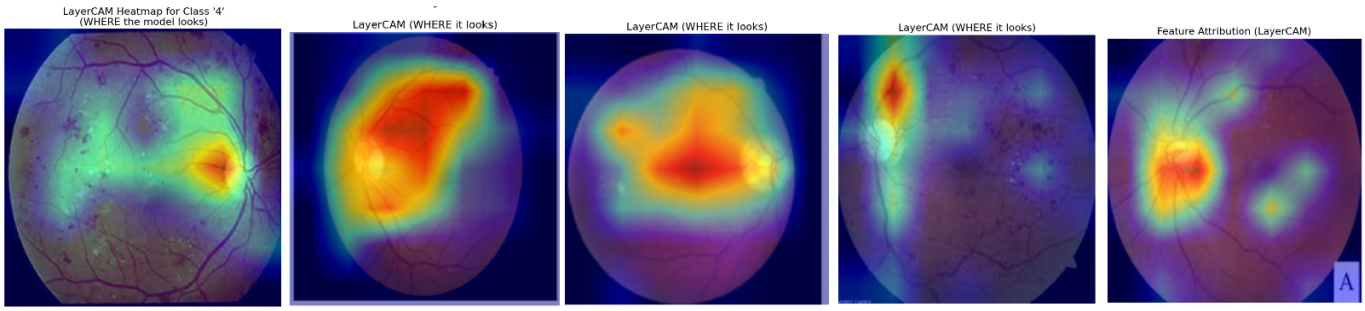

Figure 9. LayerCAM feature attribution maps demonstrating the model's region of interest. The heatmaps (red/yellow regions) indicate the pixels that most strongly influenced the model's prediction, confirming that attention is focused on clinically relevant areas.

5 Discussion

The results presented in this study offer a compelling validation of lightweight, interpretable deep learning architectures for medical diagnostics. By systematically comparing our Global Average Pooling (GAP) based model against traditional dense architectures, we have highlighted three critical advancements: algorithmic efficiency, diagnostic precision, and, most importantly, decision-making transparency.

5.1 Breaking the Trade-off Between Efficiency and Accuracy

A prevailing paradigm in deep learning has been that higher accuracy requires deeper, more parameter-heavy networks. Our findings challenge this assumption in the context of diabetic retinopathy grading. The proposed model achieved a Quadratic Weighted Kappa score of 0.8412, statistically comparable to the complex baseline model (Kappa 0.8448). This indicates that the dense, fully connected layers found in standard CNNs, which contribute the vast majority of parameters, are often redundant for this specific task. By replacing them with GAP, we achieved a parameter reduction of 21.8% (dropping from 12.4 million to 9.7 million parameters). This structural optimization acts as a regularizer, reducing the risk of overfitting while maintaining the model's capacity to learn subtle, discriminative features of the retina.

5.2 Implications for Global Health and Edge Computing

The efficiency gains realized in this study have profound practical implications for deployment in resource-constrained environments. Diabetic retinopathy is a growing epidemic in developing regions where high-end computational infrastructure (such as cloud-based GPU clusters) is often inaccessible. The reduced computational footprint and 13% faster training time of our model suggest that it is highly suitable for "Edge AI" applications. This architecture could feasibly be deployed on portable diagnostic devices or mid-range laptops in rural clinics, enabling real-time screening without reliance on stable internet connectivity for cloud processing.

5.3 Resolving the "Black Box" Dilemma with RAMs

Perhaps the most transformative contribution of this research is the integration of Regression Activation Maps (RAMs). The clinical adoption of AI has historically been stalled by the "black box" nature of deep learning; clinicians are reluctant to trust a diagnosis derived from an opaque process. RAMs bridge this trust gap by generating a visual audit trail. As demonstrated in our qualitative analysis, the model does not merely output a score; it visually localizes the pathological evidence, such as microaneurysms, hemorrhages, and neovascularization, that supports its conclusion. This feature transforms the AI from a replacement tool into a collaborative partner. In a clinical workflow, this allows for a "human-in-the-loop" system where the AI flags specific regions of interest, and the ophthalmologist verifies the findings, significantly reducing the time required for manual grading while ensuring accountability.

5.4 Limitations and Future Scope

While these results are promising, specific limitations must be acknowledged to guide future research:

- **Dataset Bias and Generalizability**: This study relied exclusively on the Kaggle detection dataset. Medical imaging data can vary significantly depending on the fundus camera model, lighting conditions, and patient demographics. Future work must validate this architecture on external datasets (cross-domain validation) to ensure it remains robust against varied imaging protocols.

- **Weakly Supervised Segmentation**: Currently, the RAMs provide coarse localization of lesions. A promising avenue for future research is utilizing these activation maps as "pseudo-labels" for semi-supervised learning. By treating the high-activation regions as ground truth masks, we could train distinct segmentation models to outline lesions pixel-by-pixel, eliminating the need for expensive, manually annotated segmentation datasets.

6 Conclusion

We have discussed the essential trade-off between the transparency of the model and computational efficiency in the automated Diabetic Retinopathy (DR) grading in this comprehensive research. We achieved a lean Convolutional Neural Network architecture by redefining the severity assessment problem as a continuous regression problem and replacing dense layers with Global Average Pooling (GAP), engineering a leaner convolutional neural network architecture. By redefining the severity assessment as a continuous regression problem and replacing dense layers with Global Average Pooling (GAP), our approach yields a substantially more efficient architecture. As evidenced by the comparative analysis, this structural optimization significantly reduces parameter counts compared to traditional fully-connected designs, while delivering diagnostic performance comparable to established baselines. Combined with the explanatory power of Regression Activation Maps (RAMs), this enables a transparent, reliable, and scalable solution for automated DR screening. The most radical of all the results of the research, though, is the creation of Regression Activation Maps (RAMs). The visualization method is quite successful in breaking the black box paradigm, as the focus of the model is reversed onto the retinal image. RAMs give the explanatory evidence required to develop clinical trust by identifying pathological markers in a distinct way, e.g., hemorrhages, exudates. This would enable the ophthalmologists to authenticate AI decisions, which would create a human-in-the-loop ecosystem. Finally, this combination of operational performance and visual interpretability will be a breakthrough in medical AI, and an accountable, scalable solution to the rollout of automated screening in resource-limited settings to address a preventable form of blindness.

Data Availability Statement

Data will be made available on request.

Funding

This work was supported without any funding.

Conflicts of Interest

The authors declare no conflicts of interest.

Ethical Approval and Consent to Participate

Ethical approval and informed consent were not required for this study, as it exclusively utilized the publicly available Kaggle Diabetic Retinopathy Detection dataset, which has been previously de-identified and approved for research purposes.

References

[1] Kaggle. (2015). *Diabetic retinopathy detection*. Retrieved from https://www.kaggle.com/c/diabetic-retinopathydetection

[2] Antony, M., & Brggemann, S. (2015). Kaggle diabetic retinopathy detection: team o_O solution. Competition Report Github. Retrieved from https://github.com/sveitser/kaggle_diabetic

[3] Bazzani, L., Bergamo, A., Anguelov, D., & Torresani, L. (2016, March). Self-taught object localization with deep networks. In *2016 IEEE winter conference on applications of computer vision (WACV)* (pp. 1-9). IEEE. [CrossRef]

[4] Bengio, Y. (2009). Learning deep architectures for AI. *Foundations and trends® in Machine Learning*, *2*(1), 1-127. [CrossRef]

[5] Deng, L. (2014). A tutorial survey of architectures, algorithms, and applications for deep learning. *APSIPA Transactions on Signal and Information Processing*, *3*, e2. [CrossRef]

[6] Dosovitskiy, A., & Brox, T. (2016, June). Inverting Visual Representations with Convolutional Networks. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 4829-4837). IEEE. [CrossRef]

[7] Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems* (pp. 1097–1105).

[8] LeCun, Y., Boser, B., Denker, J. S., Henderson, D., Howard, R. E., Hubbard, W., & Jackel, L. D. (1989). Backpropagation applied to handwritten zip code recognition. *Neural Computation*, *1*(4), 541–551. [CrossRef]

[9] LeCun, Y., Bottou, L., Orr, G. B., & Müller, K. R. (2002). Efficient backprop. In *Neural networks: Tricks of the trade* (pp. 9-50). Berlin, Heidelberg: Springer Berlin Heidelberg. [CrossRef]

[10] Lim, G., Lee, M. L., Hsu, W., & Wong, T. Y. (2014, July). Transformed representations for convolutional neural

networks in diabetic retinopathy screening. In *AAAI workshop: modern artificial intelligence for health analytics* (pp. 21-25).

[11] Mahendran, A., & Vedaldi, A. (2015, June). Understanding deep image representations by inverting them. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 5188-5196). IEEE.[CrossRef]

[12] Oquab, M., Bottou, L., Laptev, I., & Sivic, J. (2014, June). Learning and Transferring Mid-level Image Representations Using Convolutional Neural Networks. In *2014 IEEE Conference on Computer Vision and Pattern Recognition* (pp. 1717-1724). IEEE. [CrossRef]

[13] Pinz, A., Bernogger, S., Datlinger, P., & Kruger, A. (1998). Mapping the human retina. *IEEE Transactions on medical imaging, 17*(4), 606-619. [CrossRef]

[14] Pratt, H., Coenen, F., Broadbent, D. M., Harding, S. P., & Zheng, Y. (2016). Convolutional neural networks for diabetic retinopathy. *Procedia computer science, 90*, 200-205. [CrossRef]

[15] Silberman, N., Ahrlich, K., Fergus, R., & Subramanian, L. (2010, March). Case for Automated Detection of Diabetic Retinopathy. In *AAAI Spring Symposium: Artificial Intelligence for Development*.

[16] Sopharak, A., Uyyanonvara, B., & Barman, S. (2009). Automatic exudate detection from non-dilated diabetic retinopathy retinal images using fuzzy c-means clustering. *sensors, 9*(3), 2148-2161. [CrossRef]

[17] Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., ... & Rabinovich, A. (2015, June). Going deeper with convolutions. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 1-9). IEEE. [CrossRef]

[18] Wang, S., Yin, Y., Cao, G., Wei, B., Zheng, Y., & Yang, G. (2015). Hierarchical retinal blood vessel segmentation based on feature and ensemble learning. *Neurocomputing, 149*, 708–717. [CrossRef]

[19] Wang, Z., Yan, W., & Oates, T. (2017). Time series classification from scratch with deep neural networks: A strong baseline. In *2017 International Joint Conference on Neural Networks (IJCNN)* (pp. 1578–1585). [CrossRef]

[20] Wu, D., Zhang, M., Liu, J. C., & Bauman, W. (2006). On the adaptive detection of blood vessels in retinal images. *IEEE Transactions on Biomedical Engineering, 53*(2), 341-343. [CrossRef]

[21] Yang, J. B., Nguyen, M. N., San, P. P., Li, X. L., & Krishnaswamy, S. (2015, July). Deep convolutional neural networks on multichannel time series for human activity recognition. In *Proceedings of the 24th International Conference on Artificial Intelligence* (pp. 3995-4001).

[22] 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. [CrossRef]

[23] Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., & Torralba, A. (2014). Object detectors emerge in deep scene cnns. *arXiv preprint arXiv:1412.6856*.

[24] Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., & Torralba, A. (2016, June). Learning Deep Features for Discriminative Localization. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 2921-2929). IEEE. [CrossRef]

[25] Lin, M., Chen, Q., & Yan, S. (2013). Network in network. *arXiv preprint arXiv:1312.4400*.

[26] Howard, A. G., Zhu, M., Chen, B., Kalenichenko, D., Wang, W., Weyand, T., ... & Adam, H. (2017). Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv preprint arXiv:1704.04861*.

[27] Tjoa, E., & Guan, C. (2020). A survey on explainable artificial intelligence (XAI): Toward building trustable AI. *IEEE Transactions on Neural Networks and Learning Systems, 32*(11), 4793–4813. [CrossRef]

[28] Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. (2017, October). Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. In *2017 IEEE International Conference on Computer Vision (ICCV)* (pp. 618-626). IEEE. [CrossRef]

[29] Ghassemi, M., Oakden-Rayner, L., & Beam, A. L. (2021). The false hope of current approaches to explainable artificial intelligence in health care. *The lancet digital health, 3*(11), e745-e750. [CrossRef]

[30] Suryani, A. I., Chang, C. W., Feng, Y. F., Lin, T. K., Lin, C. W., Cheng, J. C., & Chang, C. Y. (2022). Lung tumor localization and visualization in chest X-ray images using deep fusion network and class activation mapping. *IEEE Access, 10*, 124448-124463. [CrossRef]

[31] Suvalakshmi, S., & Vinoth Kumar, B. (2025, April). Diabetic Retinopathy Classification using Transformer Models: An Comprehensive Survey. In *International Conference on Computer Vision and Robotics* (pp. 58-72). Cham: Springer Nature Switzerland. [CrossRef]

[32] Yuan, H., Kang, L., & Li, Y. (2025). Opening the black box of deep learning: Validating the statistical association between explainable artificial intelligence (XAI) and clinical domain knowledge in fundus image-based glaucoma diagnosis. *arXiv preprint arXiv:2504.04549*.

[33] Alonso-Caneiro, D., Kugelman, J., Tong, J., Kalloniatis, M., Chen, F. K., Read, S. A., & Collins, M. J. (2021, November). Use of uncertainty quantification as a surrogate for layer segmentation error in Stargardt disease retinal OCT images. In *2021 Digital Image Computing: Techniques and Applications (DICTA)* (pp. 1-8). IEEE. [CrossRef]

Muhammad Imran received the B.S. degree in Information Technology from the University of Education, Pakistan, in 2018, and the M.S. degree in Computer Science from the University of Okara, Pakistan, in 2022. He is currently pursuing a Ph.D. degree in Computer Science at the University of Posts and Telecommunications, China. His research during the B.S. focused on face recognition, while his M.S. research was in the field of image processing. He is currently conducting research in Explainable Artificial Intelligence (XAI). In addition, he has undertaken research and coursework in networking and cybersecurity. (Email: L202310008@stu.cqupt.edu.cn, imran.khalid292@gmail.com)

I. Ahmad received the M.S. degree in Pattern Recognition and Intelligent Systems from Beihang University, Beijing, China. His research interests include remote sensing object detection for smart cities, adversarial machine learning, medical image analysis, and hybrid modeling methodologies. (Email: israrcsc5@gmail.com)

Sohaib Saleem received the MS degree in Computer Science from Government College University Faisalabad (GCUF), Pakistan, in 2020, where his research focused on blockchain technology, privacy-preserving systems, and healthcare security. He is currently pursuing the Ph.D. degree in Smart Computing under an integrated program offered jointly by the University of Florence, the University of Pisa, and the University of Siena, Italy. He is hosted at the Department of Information Engineering at the University of Pisa. His current research focuses on cryptography, blockchain security, multimedia authentication, BLS signatures, decentralized storage, and IoT security. His broader research interests include digital forensics, secure system design, and explainable artificial intelligence (XAI). (Email: sohaib.saleem@phd.unipi.it, sohaib.saleem@unifi.it)

Altaf Hussain received his Bachelor Degree in Computer Science from University of Peshawar, Pakistan in 2013 & Master Degree in Computer Science from The University of Agriculture Peshawar, Pakistan in 2017, respectively. He has more than 6 years of teaching & research experience. He worked at The University of Agriculture Peshawar in Faculty of IT as Researcher from 2017 to 2019. He has supervised many bachelor's and master's degree level students and helped them with their final year projects and research. During his Master study, he has completed his research in drone communication systems. Currently, he is a PhD Scholar in School of Computer Science and Technology, Chongqing University of Posts and Telecommunications, Chongqing, China, specializing in Underwater Acoustic Sensor Networks with focus on energy efficiency, channel optimization, localization and dynamic scheduling integrating 6G and quantum computing. He has served as a Lecturer in Computer Science Department in Government Degree College Lal Qilla Dir Lower, KPK Pakistan from 2020 to 2021. He has worked as Research Assistant with the Department of Accounting and Information Systems, College of Business and Economics, Qatar University, Doha, Qatar. He also worked as IT clerk in the Court of District and Session Judge Timergara Dir Lower from 2022 to 2023. He has published several notable research papers. He has reviewed many articles and is serving as reviewer for Cluster Computing, Computing, Cybernetics and Systems, Journal of Cloud Computing, Knowledge and Information Systems, Peer-to-Peer Networking and Applications, SN Applied Sciences, The Imaging Science Journal, The Journal of Supercomputing, Transactions on Emerging Telecommunications Technologies, Wireless Personal Communications, Frontiers in Big Data, CMC-Computers, Materials & Continua, and Bulletin of Electrical Engineering and Informatics (BEEI). His Research interest includes Artificial Intelligence, Machine Learning, Deep Learning, Gesture Detection, Wireless Networks, Underwater Sensor Networks, and Unmanned Drone Communication Systems. (Email: altafkfm74@gmail.com)

Sayed Akif Hussain received his Bachelor of Science degree in Accounting and Finance from the University of Wah, Pakistan. He is currently pursuing his Master's degree at the School of Economics and Management Sciences, where his research focuses on Financial Engineering, with a particular interest in the application of advanced computational techniques, machine learning, and data-driven models to financial markets, Corporate Finance, Financial Statement Analysis and investment strategies. His academic and research background combines expertise in finance with modern analytical methods, aiming to bridge theory and practice in innovative financial solutions. (Email: L202320034@stu.cqupt.edu.cn, syedakifhussain110@gmail.com)

Atif Ali Wagan is a PhD Scholar at School of Computer Science and Technology, Chongqing University of Posts and Telecommunications, Chongqing 400065, China specializing in IIoT security with a focus on cybersecurity deception. His broader research interests include artificial intelligence, computer vision, deep learning, game theory, and reinforcement learning. (Email: atif.wagan2@yahoo.com)

Muzammil Khan received the B.Sc. degree in Electronic Engineering from the Balochistan University of Information Technology, Engineering and Management Sciences (BUITEMS), Pakistan. He is currently pursuing M.Sc. in Control Science and Engineering at the College of Automation, Chongqing University of Posts and Telecommunications (CQUPT), Chongqing, China. His research interests include biomimetic MEMS-based wearable sensors for health monitoring, sensor design and microfabrication, and signal processing for biomedical applications. (Email: m.khan25396@gmail.com)