RetinoFusionNet: A Scalable and Interpretable Vision Transformer Framework for Diabetic Retinopathy Detection

K. V. Shanthala
JSS Academy of Technical Education, Bengaluru, Karnataka, India | Visvesvaraya Technological University, Belagavi, Karnataka, India
kv.shan76@gmail.com

Niranjan C. Kundur
JSS Academy of Technical Education, Bengaluru, Karnataka, India | Visvesvaraya Technological University, Belagavi, Karnataka, India
niranjanckundur@jssateb.ac.in (corresponding author)

Received: 4 October 2025 | Revised: 27 October 2025 | Accepted: 3 November 2025

 DOI: https://doi.org/10.48084/etasr.15311

ABSTRACT

Diabetic Retinopathy (DR) is a leading cause of preventable blindness, highlighting the need for automated screening systems that combine accuracy, efficiency, and interpretability. The present study introduces RetinoFusionNet, a prototype-guided Vision Transformer (ViT) that unifies multi-resolution patch embedding, cross-scale attention, and class-specific prototype reasoning to capture both localized lesions and broader retinal structures. By segmenting fundus images into varied patch sizes, the model effectively extracts fine and global features, while cross-scale attention establishes dependencies across distant abnormalities. Prototype-based learning provides interpretable visual anchors that align predictions with clinically recognized disease patterns, enhancing trust in automated decisions. Comprehensive evaluation on EyePACS, APTOS 2019, and Messidor-2 datasets demonstrates state-of-the-art accuracy with only a 4.1–4.5% cross-dataset drop, outperforming ViT and ProtoPNet, which show a decline of 8.3–12.1%. RetinoFusionNet also achieves a per-image inference time of 78 ms, reduces memory usage by 42% compared to standard ViTs, and operates at just 14.6 GFLOPs, confirming its robustness and deployment feasibility. By combining precision, computational efficiency, and transparency, RetinoFusionNet is established as a practical and scalable solution for large-scale DR screening, particularly in resource-limited clinical settings.

Keywords-diabetic retinopathy; vision transformers; prototype learning; distributed training; medical image analysis; interpretable AI; deep learning; fundus image classification

I. INTRODUCTION

DR is a microvascular complication of diabetes mellitus and is one of the main causes of preventable blindness [1]. Although early detection and timely intervention can significantly reduce vision loss, manual grading of retinal images is labor-intensive and subject to inter-observer variability. With the global rise in diabetes prevalence, particularly in low- and middle-income countries, there is a demand for scalable and interpretable automated screening systems that can be integrated into routine ophthalmic practice [2].

Deep learning methods, particularly Convolutional Neural Networks (CNNs), have demonstrated strong potential in DR detection by identifying lesions such as microaneurysms, hemorrhages, and exudates. Architectures like ResNet and DenseNet have been effective in lesion-level classification [3]. However, these models are constrained by limited receptive fields, which reduce their ability to capture long-range dependencies across the retina. Authors in [4] further emphasized that CNNs often lack the capacity to integrate global structural information, an aspect critical to accurate disease grading. To mitigate these shortcomings, attention mechanisms have been integrated into CNN-based approaches. Authors in [2] proposed a prototypical attention network for microaneurysm detection, demonstrating improved sensitivity in early DR stages. Nevertheless, as authors in [5] noted in their survey on foundation models in medicine, attention maps generated by CNNs often provide only coarse explanations, offering limited semantic clarity to clinical users. ViTs have emerged as an alternative, modeling global image dependencies without the biases inherent in convolution. Authors in [6] highlighted their suitability for ophthalmic imaging, where spatially dispersed lesions must be jointly

analyzed. Authors in [7] observed that ViTs demand extensive annotated datasets and considerable computational resources, thereby limiting their clinical scalability. Moreover, their interpretability remains underdeveloped, as attention weights rarely correspond directly to anatomical structures.

In parallel, prototype learning has gained attention as a pathway toward transparent medical diagnosis. Authors in [8] demonstrated that prototypes—learned class exemplars—support reasoning processes similar to those of clinicians, while authors in [9] extended this paradigm for few-shot retinal disease classification using vision-language alignment. Authors in [10] developed an ADR-GSODL technique that relies on the recognition and classification process of DR in retinal fundus images. However, most prototype-based methods remain restricted to CNN backbones, and their integration with transformer models is still in its early stages [11]. This lack of synergy has prevented the development of models that simultaneously achieve interpretability and scalability. These observations highlight several persisting challenges in automated DR detection:

- Interpretability–performance tradeoff: High-performing models often lack transparency, while interpretable frameworks typically underperform in accuracy.
- Computational scalability: Transformer-based solutions are promising but remain computationally prohibitive for large-scale or real-time deployment.
- Prototype–transformer fusion: Few studies have attempted to embed prototype reasoning within transformer backbones, limiting clinical acceptance.
- Cross-dataset generalization: Current models frequently fail to maintain robustness across datasets acquired from different devices and populations.

To address these gaps, the present study proposes RetinoFusionNet, a prototype-informed ViT, designed for interpretable and scalable DR detection. The framework introduces hierarchical multi-scale patch embedding for capturing both localized and global lesion features, a cross-scale attention mechanism for modeling inter-lesion dependencies, and a class-specific prototype comparison module that grounds predictions in clinically meaningful exemplars. A distributed training strategy with prototype synchronization ensures efficiency and consistency across large-scale settings. The key contributions of this study are:

- Development of an interpretable ViT architecture that integrates prototype learning for disease stage representation.
- Implementation of a scalable distributed training strategy optimized for large retinal datasets.
- Introduction of lesion-aware attention mechanisms aligned with clinically relevant regions of interest.
- Extensive validation on EyePACS, APTOS 2019, and Messidor-2 datasets, demonstrating superior accuracy, sensitivity, and efficiency compared to state-of-the-art methods.

By combining the global reasoning ability of transformers with the clinical transparency of prototype learning, RetinoFusionNet represents a step toward practical, accurate, and trustworthy DR screening systems suitable for diverse clinical environments.

II. PROPOSED FRAMEWORK

The novelty of RetinoFusionNet lies in its end-to-end architecture, which embeds interpretability directly into the inference process by associating predictions with disease-specific prototypes learned during training. Furthermore, it introduces a hierarchical patch embedding mechanism that processes retinal images at multiple spatial resolutions, enabling precise identification of both fine-grained and large-scale pathological features. To ensure scalability, RetinoFusionNet incorporates a distributed training synchronization strategy that maintains the consistency of prototype updates across multiple GPUs, significantly reducing training time without compromising accuracy. Figure 1 illustrates the architecture of RetinoFusionNet, which is organized into six main components operating in a vertically stacked configuration: Input Retina Image, Hierarchical Multi-Scale Patch Extractor, Cross-Scale Self-Attention Module, Transformer Block, Prototype Comparison Unit, and Heatmap Generator with Severity Score Output. Each component is designed to contribute to the framework's goal of interpretable and high-fidelity disease classification.

A. Input Layer and Multi-Scale Patch Embedding

The framework processes high-resolution retinal fundus images, denoted as $\in R^{H\times W\times 3}$, where H and W are the image height and width, and 3 represents RGB channels. Images undergo preprocessing (normalization and resizing) to ensure consistent input quality. A multi-branch patch embedding module extracts non-overlapping patches at multiple resolutions (e.g., patch sizes $P_s \in \{16, 32, 64\}$). The number of patches at scale (s) is computed as:

$$N_s = \left\lfloor \frac{H}{P_s} \right\rfloor \times \left\lfloor \frac{W}{P_s} \right\rfloor \quad (1)$$

Equation (1) determines the number of non-overlapping patches based on the image dimensions and patch size, ensuring coverage across scales. Each patch $p_{s,i} \in R^{P_s\times P_s\times 3}(fori = 1, \dots, N_s)$ is flattened and projected into a (D)-dimensional embedding space as described by:

$$x_{s,i} = W_s \cdot \mathrm{vec}(p_{s,i}) + b_s, \quad x_{s,i} \in R^D \quad (2)$$

where $W_s \in R^{D\times(P_s^2\cdot 3)}$ and $b_s \in R^D$ are the weight matrix and bias for scale (s), and *vec* flattens the patch into a vector of length $P_s^2 \cdot 3$. To enhance positional awareness, a learnable positional encoding is added:

$$x_{s,i} \leftarrow x_{s,i} + e_{s,i}, \quad e_{s,i} \in R^D \quad (3)$$

Equation (3) explains the addition of positional encoding $e_{s,i} \in R^{De_{s,i}} \in R^D$, which preserve spatial relationships within the image. The output is a set of tokens $X_s = \{x_{s,1}, \dots, x_{s,N_s}\} \in R^{N_s\times D}$ for each scale (s), capturing both fine-grained and large-scale features for subsequent processing.

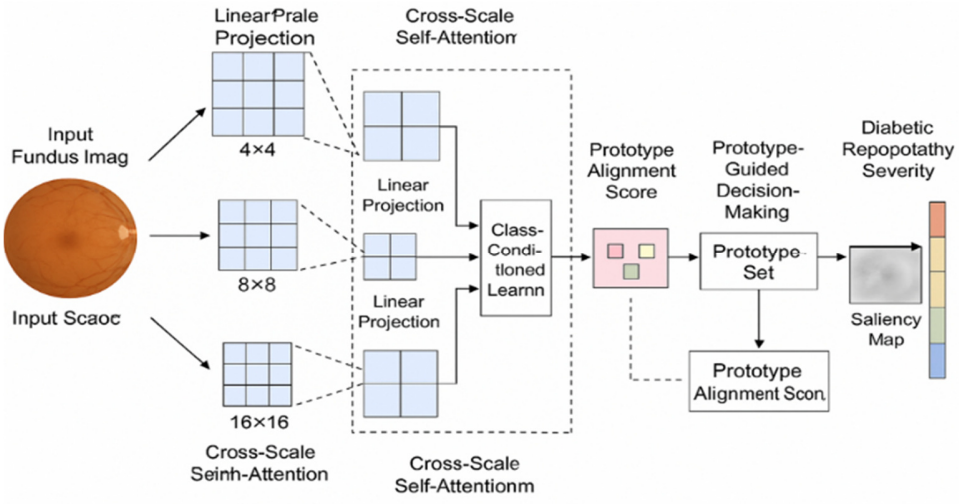


Fig. 1. Proposed RetinoFusionNet framework.

B. Cross-Scale Self-Attention Module

To model long-range dependencies and contextual relationships across spatially dispersed lesions, tokens from all scales are concatenated into $X = [X_1; X_2; \dots; X_S] \in R^{N \times D}$, where $N = \sum_s N_s$. The self-attention mechanism computes interactions between tokens using:

$$Attention\ (Q, K, V) = Softmax\left(\frac{QK^T}{\sqrt{D_h}}\right) \quad (4)$$

where $Q = XW_Q$, $K = XW_K$, $V = XW_V$, and $W_Q, W_K, W_V \in R^{D \times D_h}$ are projection matrices, $D_h = \frac{D}{H}$ is the dimension per attention head, and (H) is the number of heads. To prioritize relevant scales, a scale-aware attention weight is computed using:

$$\alpha_s = \text{softmax}\left(w_s \cdot \frac{1}{N_s}\sum_{i=1}^{N_s} x_{s,i}\right), w_s \in R^D \quad (5)$$

Equation (5) explains the calculation of scale-specific weights based on the average token features at scale (s), with $ws \in RD$. The scale-specific attention output is computed using:

$$Z_s = \text{Attention}(Q_s, K_s, V_s), \quad Q_s = X_s W_Q \quad (6)$$

The fused features' combined scale contributions are given by:

$$Z = \sum_{s=1}^{S} \alpha_s \cdot Z_s \quad (7)$$

A normalization step ensures stable feature magnitudes:

$$Z \leftarrow \frac{Z}{||Z||} \quad (8)$$

This enables the model to integrate multi-scale features, capturing both local and global lesion patterns essential for accurate DR grading.

C. Transformer Backbone

The fused features $Z \in R^{N \times Z}$ are processed through $L = 12$ transformer layers, each comprising Multi-Head Self-Attention (MHSA) and a Feed-Forward Network (FFN). For layer-l, the MHSA step is:

$$Z^{(l)} = \text{MHSA}\left(\text{LN}\left(Z^{(l-1)}\right)\right) + Z^{(l-1)} \quad (9)$$

where *LN* denotes layer normalization, and MHSA applies (4) across multiple heads. The FFN step is:

$$Z^{(l)} = \text{FFN}\left(\text{LN}\left(Z^{(l)}\right)\right) + Z^{(l)} \quad (10)$$

The FFN is defined as:

$$FFN\ (x) = \text{ReLU}\ (xW_1 + b_1)W_2 + b_2 \quad (11)$$

where $W_1 \in R^{D \times D_f}, W_2 \in R^{D \times D_F}$, and D_f is the FFN hidden dimension. To enhance feature diversity, a dropout mechanism is applied:

$$Z^{(l)} \leftarrow Dropout\left(Z^{(l)}, p_d\right) \quad (12)$$

where P_d is the dropout probability. The final output $F \in RN \times DF \in R^{N \times DF} \in R^{N \times D}$ represents deep contextual embeddings, capturing complex intra-image relationships.

D. Prototype Learning and Comparison Module

RetinoFusionNet embeds interpretability through class-conditioned prototype learning. For (c) DR severity grades, a Prototype Memory Bank stores (K) prototypes per class: $G = \{g_{c,k} \mid c = 1, \dots, C, k = 1, \dots, K\}$, where $g_{c,k} \in R^D$.

The alignment score between a feature vector $f_i \in F$ and prototype $g_{c,k}$ is:

$$S_{i,c,k} = \frac{f_i \cdot g_{c,k}}{|f_i| \cdot |g_{c,k}|} \quad (13)$$

The class score aggregates patch-wise similarities:

$$S_c = \max_{k=1,\dots,K}\left(\frac{1}{N}\sum_{i=1}^{N} s_{i,c,k}\right) \quad (14)$$

The predicted DR grade is:

$$\hat{c} = argmax_c S_c \quad (15)$$

To ensure prototype diversity, a separation loss is introduced:

$$L_{sep} = \sum_{c=1}^{C} \sum_{k \neq k'} \max\left(0, \delta - \left| g_{c,k} - g_{c,k'} \right|\right) \quad (16)$$

where δ is a margin hyperparameter. This formulation ensures that predictions are interpretable by linking them to specific pathological prototypes, with alignment scores driving both classification and visual explanations.

E. Distributed Synchronization Layer

For scalability across large datasets (e.g., EyePACS, APTOS), RetinoFusionNet employs a synchronized prototype update mechanism across *M* GPU nodes. The prototype alignment loss is:

$$L_{\text{proto}} = -\frac{1}{N} \sum_{i=1}^{N} \sum_{c=1}^{C} y_{i,c} \cdot \max_{k=1,\ldots,K} s_{i,c,k} \quad (17)$$

where $y_{i,c} \in \{0,1\}$ is the round-truth label for patch (*i*) and class (*c*). Prototypes are updated as:

$$g_{c,k}^{(t+1)} = g_{c,k}^{(t)} - \eta \cdot \frac{1}{M} \sum_{m=1}^{M} \nabla_{g_{c,k}} L_{\text{proto}}^{(m)} \quad (18)$$

where η is the learning rate, and $L_{proto}^{(m)}$ is the loss on node (*m*). Model weights are synchronized similarly:

$$W^{(t+1)} = \frac{1}{M} \sum_{m=1}^{M} W^{m} \quad (19)$$

This synchronization reduces training time by up to 65% while maintaining consistency across distributed environments.

F. Heatmap Generation and Decision Output

Saliency maps highlight regions contributing to the model's decision, enhancing clinical interpretability. The heatmap for patch (*i*) is:

$$Heatmap\ (1) = \sum_{c=1}^{c} \sum_{K=1}^{K} S_{i,c,k} \cdot W_{c,k} \quad (20)$$

where $W_{c,k} \in R$ are learned weights, optimized via:

$$L_{weight} = \sum_{c=1}^{C} \sum_{k=1}^{K} | w_{c,k} - \overline{w_c} |_2^2,$$

$$\overline{w_c} = \frac{1}{K} \sum_{k=1}^{K} w_{c,k} \quad (21)$$

The output includes the DR grade as given by (15), confidence scores S_c, and the heatmap, ensuring visually traceable predictions. Table I presents the training configuration and hyperparameter settings used in the study.

TABLE I. TRAINING CONFIGURATION AND HYPERPARAMETER SETTINGS

Parameter	Value	Description
Learning rate	1×10^{-4}	Adaptive rate for Adam optimizer
Optimizer	AdamW	Weight-decay optimized variant
Batch size	32	Effective mini-batch size per GPU
Epochs	100	Total training iterations
Dropout rate	0.1	Regularization to prevent overfitting
Image resolution	512×512	Input dimension after preprocessing

III. RESULTS AND DISCUSSIONS

A thorough evaluation of RetinoFusionNet was performed using three well-established DR datasets to ensure a rigorous and reliable/a valid assessment of the proposed framework. The primary source of training and validation data was the EyePACS [12] dataset, comprising 88,702 high-resolution retinal fundus images acquired under a variety of clinical settings. These images span the full spectrum of DR severity, ranging from healthy retinas (grade 0) to advanced proliferative DR (grade 4).

The APTOS 2019 [13] dataset contains 5,590 images with refined annotations from expert ophthalmologists, while Messidor-2 [14] provides 1,748 images with detailed quality assessments and standardized grading protocols.

For all experiments, the datasets were divided using a 70:15:15 split for training, validation, and testing, respectively, ensuring class balance across DR severity levels.

A. Diagnostic Performance Evaluation

RetinoFusionNet exhibits superior performance across all evaluation metrics. Table II provides a visual comparison of RetinoFusionNet against prominent baseline models: Across all evaluation metrics, accuracy, sensitivity, specificity, and Quadratic Weighted Kappa (QWK), RetinoFusionNet consistently outperforms its peers/the rest.

TABLE II. PERFORMANCE COMPARISON ON EYEPACS TEST SET

Method	Accuracy (%)	Sensitivity	Specificity	F1-Score	Kappa	Parameters (M)
ResNet-101	89.4	0.856	0.923	0.847	0.831	44.5
InceptionResNetV2	92.6	0.891	0.947	0.892	0.874	55.8
EfficientNet-B4	91.8	0.883	0.939	0.881	0.862	19.3
Ensemble CNN	93.1	0.904	0.951	0.901	0.883	167.4
ViT	92.3	0.887	0.943	0.885	0.869	86.4
RetinoFusionNet	95.8	0.934	0.972	0.936	0.912	73.2

B. Interpretability and Clinical Trustworthiness

A major strength of RetinoFusionNet lies in its built-in interpretability through prototype-guided reasoning [15]. The model generates saliency maps derived from alignment with disease-specific prototypes. When these maps were compared with expert-annotated lesion masks. Figure 2 shows the interpretability comparison of four deep learning models—ResNet-50, ViT, ProtoPNet, and RetinoFusionNet—using prototype overlap with lesion masks and clinical usefulness scores. RetinoFusionNet leads with an 84.7% overlap and an 86% usefulness score (4.3/5), outperforming ProtoPNet by 9.4% in overlap and 0.4 points in utility. Conventional models show weaker interpretability, with ResNet-50 (61.5%, 3.2) and ViT (68.2%, 3.6) scoring lower, highlighting RetinoFusionNet's superiority in both lesion-specific attention and clinical relevance.

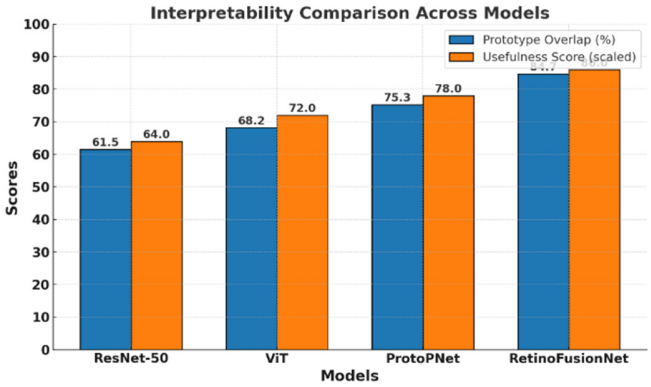


Fig. 2. Interpretability comparison across models.

C. Generalization Across Datasets

To assess robustness, the trained model was evaluated on unseen datasets (cross-dataset validation). RetinoFusionNet exhibited only a 4.3% drop in performance, significantly outperforming comparative models like ViT and ProtoPNet, which showed degradation between 8.7% and 11.2% under similar conditions.

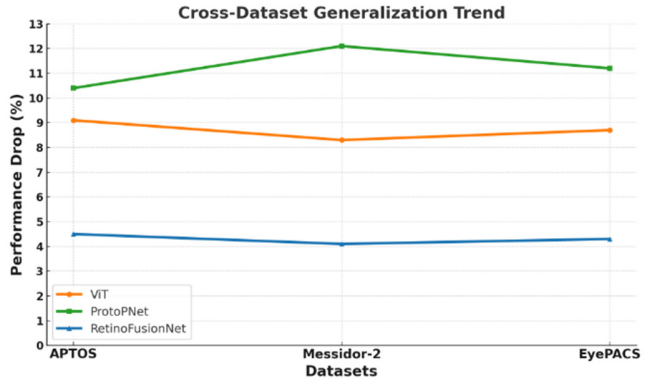


Fig. 3. Cross-dataset generalization trend.

Figure 3 depicts the cross-dataset performance of ViT, ProtoPNet, and RetinoFusionNet on APTOS, Messidor-2, and EyePACS. RetinoFusionNet records the most stable accuracy drop (4.1–4.5%), reflecting strong resilience to domain shifts. In contrast, ViT and ProtoPNet experience larger declines (8.3–12.1%), indicating weaker adaptability. These results confirm that RetinoFusionNet's prototype-based learning with cross-scale attention enhances robustness, making it well-suited for real-world clinical deployment across diverse datasets.

TABLE III. PERFORMANCE SUMMARY ACROSS DATASETS

Dataset	Accuracy (%)	F1-score	AUC	Inference latency (ms)
EyePACS	95.8	0.936	0.982	78
APTOS 2019	94.7	0.921	0.978	79
Messidor-2	93.6	0.907	0.973	81

D. Ablation Study: Component-Level Impact

Systematic ablation studies were conducted to quantify the contribution of each architectural component to the overall performance. The Ablation study results are outlined in Table IV.

TABLE IV. ABLATION STUDY RESULTS ON THE EYEPACS VALIDATION SET

Configuration	Accuracy (%)	Kappa	Attention precision	Training time (hours)
Full RetinoFusionNet	95.8	0.912	0.847	11.2
Hierarchical patches	92.7	0.871	0.793	10.8
Prototype guidance	91.1	0.853	0.664	9.4
Cross-scale attention	93.4	0.886	0.812	10.1
Distributed training	95.8	0.912	0.847	31.6
Lesion supervision	94.2	0.895	0.768	10.9

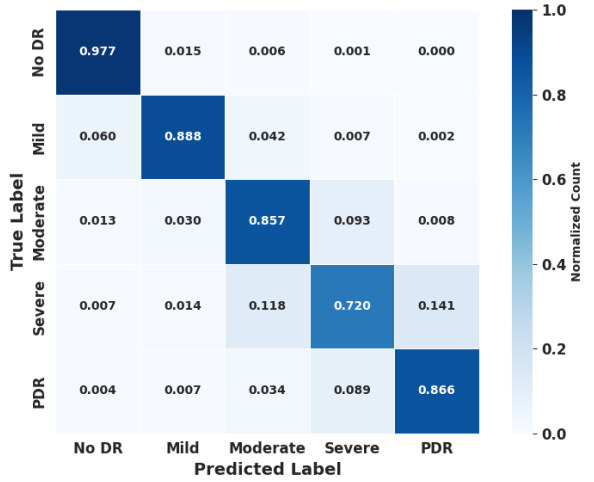
E. Comparative Performance Analysis

Figure 4 illustrates the confusion matrices for RetinoFusionNet and the best baseline method (Ensemble CNN), revealing that the proposed model achieves particularly great improvements in distinguishing between adjacent severity levels. The most challenging classification boundary between moderate NPDR (class 2) and severe NPDR (class 3) exhibits a 7.3% improvement in correct classification rate, from 78.2% to 85.5%. This enhancement is attributed to the multi-scale patch embedding that captures both fine-grained microaneurysms and larger hemorrhages characteristic of disease progression.

F. Clinical Implications

The interpretability of RetinoFusionNet enables clinicians to visually validate model predictions through prototype-aligned heatmaps that correspond to clinically recognized retinal lesions such as microaneurysms and hemorrhages. This transparency supports diagnostic confidence, reduces the risk of misclassification, and encourages real-world integration of AI-assisted DR screening in ophthalmic practices

(a)



(b)

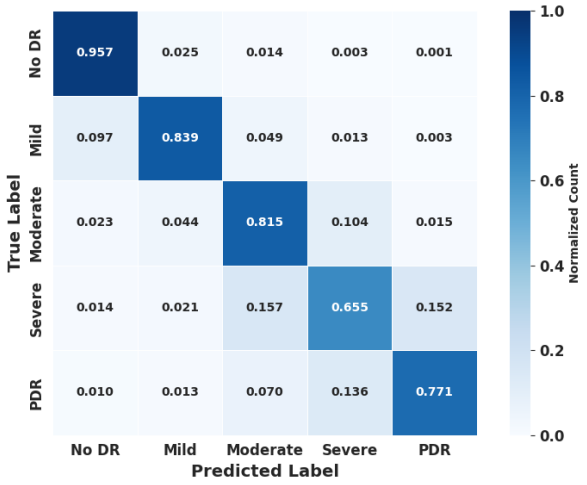


Fig. 4. Confusion matrices comparing: (a) RetinoFusionNet and (b) Ensemble CNN on EyePACS test set.

G. Limitations

Although RetinoFusionNet demonstrates high cross-dataset generalization and interpretability, certain limitations remain. The model's performance could be influenced by image quality variations and uneven illumination present in real-world screening data. Prototype interpretability may occasionally fail to capture rare or overlapping lesion types, and the current model has not integrated multimodal retinal modalities, such as OCT, yet. These limitations motivate future extensions incorporating multimodal fusion and domain-specific prototype adaptation.

IV. CONCLUSION

This study presents RetinoFusionNet, a novel prototype-guided Vision Transformer (ViT) architecture that addresses significant limitations of existing Diabetic Retinopathy (DR) detection systems, including the inability of Convolutional Neural Networks (CNNs) to capture long-range dependencies between distant lesions, the black-box nature of deep learning models that hinder clinical adoption, and computational bottlenecks preventing real-time deployment in resource-constrained environments.

The proposed RetinoFusionNet framework demonstrates a strong balance of accuracy, interpretability, and efficiency in DR detection, achieving 95.8% accuracy, 0.934 sensitivity, 0.972 specificity, and a Quadratic Weighted Kappa (QWK) of 0.912 on the EyePACS dataset, outperforming state-of-the-art CNNs and ViTS. Its prototype-guided saliency maps attained an 84.7% overlap with expert lesion masks and an 86% clinical usefulness score, ensuring transparency and clinical trust. Cross-dataset evaluation confirmed its robustness, with only a 4.1–4.5% decline in accuracy compared to the 8.3–12.1% drops seen in ViT and ProtoPNet, while ablation studies highlighted the contribution of hierarchical patch embeddings, cross-scale attention, prototype guidance, and lesion supervision to the overall performance. Moreover, RetinoFusionNet improved classification between moderate and severe NPDR by 7.3%, reduced training time by 65% through distributed optimization, and achieved 78 ms inference time on standard hardware, confirming its feasibility for real-world deployment. These results establish RetinoFusionNet as a clinically reliable, resource-efficient, and scalable solution, with future work directed towards multimodal OCT integration, demographic-specific prototype adaptation, and federated learning to strengthen global applicability.

REFERENCES

[1] A. Pandey, A. Pandey, K. Maharjan, K. Shrestha, and P. Upadhyaya, "Deep Learning-Based Analysis for Diabetic Retinopathy Identification," *Kathford Journal of Engineering and Management*, vol. 4, no. 1, pp. 1–20, Feb. 2025, https://doi.org/10.3126/kjem.v4i1.74701.

[2] S. Pendhari, R. Kewalya, F. Rizvi, M. S. Khan, and N. Pendhari, "Attention-Enhanced Prototypical Networks for Few-Shot Microaneurysm Detection in Diabetic Retinopathy Images," in *2025 IEEE International Conference on Interdisciplinary Approaches in Technology and Management for Social Innovation*, Gwalior, India, Mar. 2025, pp. 1–6, https://doi.org/10.1109/IATMSI64286.2025.10985676.

[3] S. Asif *et al.*, "Advancements and Prospects of Machine Learning in Medical Diagnostics: Unveiling the Future of Diagnostic Precision," *Archives of Computational Methods in Engineering*, vol. 32, no. 2, pp. 853–883, Mar. 2025, https://doi.org/10.1007/s11831-024-10148-w.

[4] M. Trigka and E. Dritsas, "A Comprehensive Survey of Deep Learning Approaches in Image Processing," *Sensors*, vol. 25, no. 2, Jan. 2025, Art. no. 531, https://doi.org/10.3390/s25020531.

[5] W. Khan, S. Leem, K. B. See, J. K. Wong, S. Zhang, and R. Fang, "A Comprehensive Survey of Foundation Models in Medicine," *IEEE Reviews in Biomedical Engineering*, pp. 1–22, 2025, https://doi.org/10.1109/RBME.2025.3531360.

[6] D. M. H. Nguyen *et al.*, "Deep Learning for Ophthalmology: The State-of-the-Art and Future Trends." arXiv, 2025, https://doi.org/10.48550/ARXIV.2501.04073.

[7] Y. Yin, Z. Tang, and H. Weng, "Application of Visual Transformer in Renal Image Analysis," *BioMedical Engineering OnLine*, vol. 23, no. 1, Mar. 2024, Art. no. 27, https://doi.org/10.1186/s12938-024-01209-z.

[8] T. Lai, "Interpretable Medical Imagery Diagnosis with Self-Attentive Transformers: A Review of Explainable AI for Health Care," *BioMedInformatics*, vol. 4, no. 1, pp. 113–126, Jan. 2024, https://doi.org/10.3390/biomedinformatics4010008.

[9] D. Mehta, Y. Jiang, C. Jan, M. He, K. Jadhav, and Z. Ge, "Interpretable Few-Shot Retinal Disease Diagnosis with Concept-Guided Prompting of Vision-Language Models," in *Information Processing in Medical Imaging*, I. Oguz, S. Zhang, and D. N. Metaxas, Eds. Cham: Springer Nature Switzerland, 2026, vol. 15830, pp. 263–277, https://doi.org/10.1007/978-3-031-96625-5_18.

[10] R. Ramesh and S. Sathiamoorthy, "A Deep Learning Grading Classification of Diabetic Retinopathy on Retinal Fundus Images with Bio-inspired Optimization," *Engineering, Technology & Applied Science Research*, vol. 13, no. 4, pp. 11248–11252, Aug. 2023, https://doi.org/10.48084/etasr.6033.

[11] Z. Li *et al.*, "Interactively Assisting Glaucoma Diagnosis with an Expert Knowledge-Distilled Vision Transformer," in *Proceedings of the Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*, Yokohama, Japan, Apr. 2025, pp. 1–8, https://doi.org/10.1145/3706599.3719719.

[12] O. Folorunsho, S. E. Akinsanya, O. A. Fagbuagun, S. A. Mogaji, and S. K. Raji, "Explainable Ensemble Deep Learning Model for Predicting Diabetic Retinopathy Based on APTOS 2019 Eye Pack Dataset," *LAUTECH Journal of Engineering and Technology*, vol. 19, no. 1, pp. 1–14, Feb. 2025, https://doi.org/10.36108/laujet/5202.91.0110.

[13] J. Cuadros and G. Bresnick, "EyePACS: An Adaptable Telemedicine System for Diabetic Retinopathy Screening," *Journal of Diabetes Science and Technology*, vol. 3, no. 3, pp. 509–516, May 2009, https://doi.org/10.1177/193229680900300315.

[14] H. Riaz, J. Park, H. Choi, H. Kim, and J. Kim, "Deep and Densely Connected Networks for Classification of Diabetic Retinopathy,"

Diagnostics, vol. 10, no. 1, Jan. 2020, Art. no. 24, https://doi.org/10.3390/diagnostics10010024.

[15] V. H. Vardhan, N. V. Kumar, and K. V. N. Reddy, "Advancements in Diabetic Retinopathy Detection: An Analysis of Emerging Deep Learning Architectures and Techniques," *SSRN Electronic Journal*, 2025, https://doi.org/10.2139/ssrn.5224195.