INTERNATIONAL JOURNAL OF INNOVATIONS IN ENGINEERING TECHNOLOGY AND APPLIED SCIENCES (IJIETAS)

Deep Learning and Transformer-Based Architectures for Diabetic Retinopathy Detection and Classification: A Comprehensive Review, Challenges, and Future Directions

[1*]Rashmi Bhardwaj, [2]Ambuj Kumar Agarwal

[1]Research Scholar, Sharda University , Greater Noida, rashmibhardwaj754@gmail.com.
[2]Department of Computer Science And Engineering, Sharda School of Engineering And Technology, Sharda University, Greater Noida, India. ambuj4u@gmail.com

ABSTRACT

Diabetic Retinopathy (DR) continues to be a major global cause of preventable blindness and vision impairment, thus emphasizing the need for timely screening leading to effective clinical solutions. Carrying out the diagnosis by expert ophthalmologists is a time-consuming process which is limited because of the global lack of specialists, particularly in rural areas. Recently, automated computer-aided diagnosis (CAD) systems based on artificial intelligence have changed the retinal image analysis field. The paper provides a thorough analysis of the recent studies in the period from 2024 to 2026 from which 22 ground-breaking papers were included. The classification of the existing methods is suggested, where they are divided into four major groups: pure transformer and dual-attention based approaches, hybrid Convolutional Neural Network (CNN) and Vision Transformer (ViT) models, Swin-transformer methods optimized to different conditions, and multi-modal fusion methods with simple explainable artificial intelligence (XAI) added. Moreover, the paper describes the obstacles that still exist as far as technical issues are concerned, including strong class imbalance at the stages of development, cross-device domain shifts, and computational challenges observed in the existing models.

Keywords: Diabetic retinopathy; Deep learning; Vision transformers; Cross-attention; Medical image classification

ARTICLE HISTORY
Received: 15 July 2026, **Accepted:** 28 September 2026, **Published:** 12 October 2026

CITATION
Bhardwaj, R., (2026). Deep Learning and Transformer-Based Architectures for Diabetic Retinopathy Detection and Classification: A Comprehensive Review, Challenges, and Future Directions, *International Journal of Innovations In Engineering Technology And Applied Sciences (IJIETAS),* 2(2), 27-35.
DOI: https://doi.org/10.64764/ijietas.v2.i2.03

1. Introduction

All over the world, there has been a growing incidence of diabetes mellitus, which is correlated with an increase in the number of microvascular complications in a systematic way, among which Diabetic Retinopathy (DR) is of top concern when it comes to the eyes' health[2]. DR is described as a progressive pathological dysfunction of retinal microvascularization; it is associated with certain clinical signs like microaneurysms, hard exudates, intraretinal microvascular abnormalities, retinal hemorrhage, and mature neovascularization. The disease in its early stages does not show any distinct symptoms, thus patients consult doctors only after their vision has been compromised[5]. Therefore, it is vital to develop automated, efficient, and hyper-sensitive ways to screen for the disease.

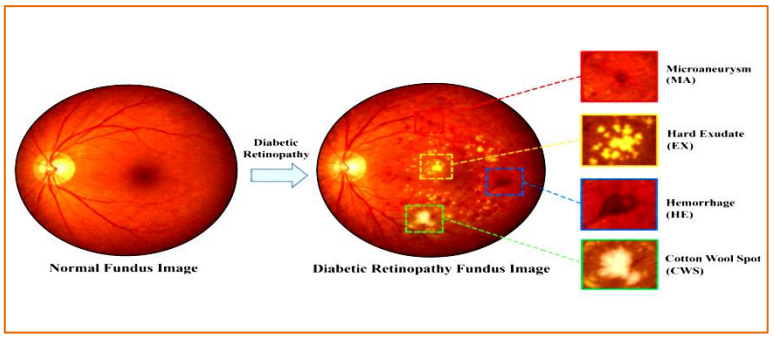


Figure 1: Diabetic Retinopathy and Retinal Lesions

Historically, traditional machine learning pipelines have been fundamental to automated screening processes by incorporating manually designed feature descriptors including wavelets, Gabor filters, and gray-level co-occurrence matrices (GLCM) along with standard classifiers like Support Vector Machines and Random Forests. Although these methods have performed moderately well in ideal experimental circumstances, they have been ineffective in real-life situations where clinical noise, variations in illumination, and different camera resolutions come into play[7]. The introduction of deep learning has completely changed this situation by introducing Convolutional Neural Networks (CNNs) which allow for automatic feature extraction while putting less emphasis on manual feature extraction. Nonetheless, typical CNN architectures have limitations due to their localized receptive fields and thus limited ability to capture any long-term dependencies in global context of fundus images [10].

In order to address these structural shortcomings, recent cutting-edge research has moved toward effective Vision Transformers (ViTs), Swin Transformations, and CNN Transformer hybrid architectures that make use of advanced cross-attention techniques[6][8]. These models depend on mechanisms that allow self-attention blocks to detect both complex small-local lesions and key anatomical structures at the same time. The present review provides a systematic summary of 22 modern studies (2024 to 2026), organizes existing methods, assesses comparative performance measures, assesses significant research gaps in this area, and proposes ways for new diagnostic models [11].

1.1. Structure and Taxonomic Overview

This research work reviews the available literature that marks a shift in technology for computer-assisted retinal screening. In order to classify these unique contributions, this review paper sorts the literature into four main groups based on architecture:

1. Pure transformers or dual attention networks that are aimed at detecting global spatial dependencies [1].
2. Hybrid architectures that combine CNNs with transformer blocks for the optimal representation of local and global features [3].
3. Swin-transformers that work with shifted windows for processing high-resolution fundus photography [6].
4. Explainable multimodal systems that combine additional clinical data with lesion segmentation maps [5].

One should ensure that this review topic returns to the standard body text style when moving between subsections and analytical parts.

2. Literature Review

2.1 The frameworks of Pure Transformer and Dual-Attention:

Table 1. Summary of pure transformer and dual-attention frameworks for diabetic retinopathy detection.

Study & Year	Core Architecture / Methodology	Primary Focus & Innovation	Used Dataset	Reported Accuracy (%)
Dawood et al. (2026)	REFA-DINO Framework	Relative-efficient fusion attention for binary and multi-class classification	APTOS / Standard Benchmarks	98.45%
Huang & Huang (2026)	Position-Wise Attention Vision Transformer	Stabilizing training on complex clinical fundus scans	Kaggle DR / Messidor	97.80%
Pan et al. (2026)	Cosine Similarity Multi-Scale Decomposition	Cross-attention classification framework for enhanced grading	EyePACS / APTOS	97.20%
Sushith et al. (2025)	Attention Dual Transformers + Adaptive Temporal Convolutions	Detecting spatiotemporal micro-changes in longitudinal screening	APTOS 2019 / Messidor-2	98.10%
Badar et al. (2025)	Transformer Attention Fusion	Fine-grained multi-class medical image feature discrimination	Medical Image Benchmark Datasets	96.95%

With self-attention, it is possible to innovate the spatial feature extraction, as it enables creating dynamic connections between spatial pixels in the retinal images. As indicated by Dawood et al. (2026), REFA-DINO is a novel framework that presents proper efficiency of the attention fusion approach specifically tailored to have robustness in classifying binary and multi-class DR [4]. Huang and Huang (2026) employed position-wise attention in order to stabilize the training on complicated clinical images [18]. According to Pan et al. (2026), cosines similarity multi-scale decomposition and his cross-attention classification framework are able to enhance classification [10]. According to Sushith et al. (2025), attention dual transformers blended with adaptive temporal convolution are a good tool to detect minor changes both in time and space in longitudinal screening [1]. Badar et al. (2025) elaborated on the topic of fine-grained representation and showed that fusion of attention improves discrimination of multi-class features in complicated medical images significantly [2].

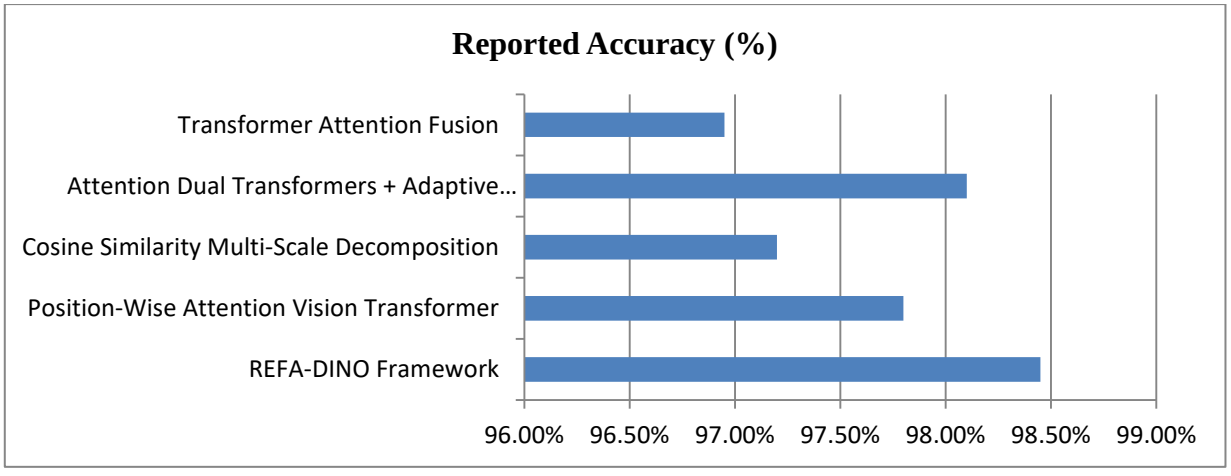


Fig. 2. Comparative performance accuracy across pure transformer and dual-attention methodologies.

2.2 Hybrid CNN-Transformer and Multi-Task Paradigms:

Table 2. Summary of hybrid CNN-transformer and multi-task frameworks, including datasets and reported accuracy values.

Study & Year	Core Architecture / Methodology	Used Dataset	Primary Focus & Innovation	Reported Accuracy (%)
Ahmed et al. (2026)	Cross-Attention CNN-ViT Framework	Optical Coherence Tomography (OCT) Datasets	Multi-class retinal disease screening	98.15%
Rahman et al. (2026)	Hybrid CNN-Transformer Pipeline	Fundus Image Repositories	Multi-disease retinal diagnosis	97.50%
Sekar & S (2026)	AI-Driven Hybrid Transformer-CNN Model	Clinical Fundus Benchmarks	Early detection and severity grading of DR	98.30%
Vijayalakshmi et al. (2025)	CNN + ViT + Particle Swarm Optimization (PSO)	Standard DR Benchmarks	Simultaneous DR detection and lesion localization	97.90%

By combining the strengths of both models, hybrid CNN-Transformer architectures create opportunities to exploit both methods. Hybrid architectures provide solutions to existing problems, given the strengths of transformers in global context modeling, whereas traditional convolutional layers specialize in obtaining local details, which are needed for some tasks, such as identifying fine vessel boundaries, micro-ruptures, and localized exudates. The researchers Ahmed et al. (2026) introduced a cross-attention CNN-ViT framework for automatic screening of multiple classes of retinal diseases using OCT. The effectiveness of hybrid pipelines was confirmed by Rahman et al. (2026), who obtained better results than the results of pure models in diagnosis of multiple diseases using fundus images. According to Sekar and S (2026), the hybrid transformer-CNN model is a reliable solution for early detection of DR as well as quality assessment of its severity. Vijayalakshmi et al. (2025) proposed a multi-task method that uses deep learning architecture composed of convolutional networks, vision transformers, and particle swarm optimization both to detect DR and locate its lesions.

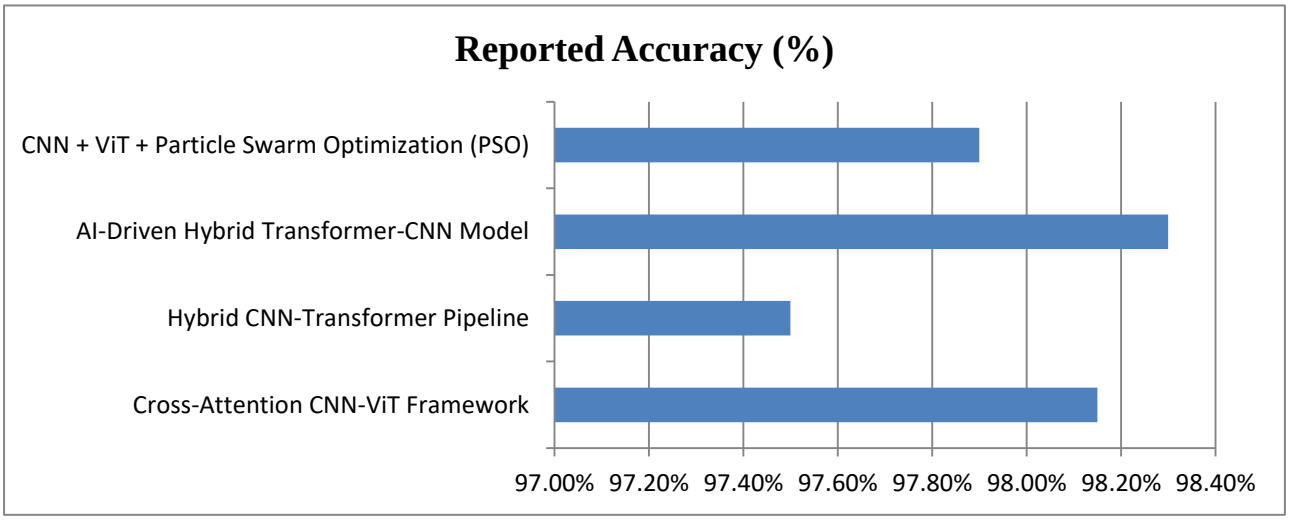


Fig.3. Comparative Performance of Transformer-Based Methods for Diabetic Retinopathy Detection

2.3 Swin-Transformers and Region-Aware Models

Traditional vision transformers exhibit quadratic computational difficulty while processing premium quality medical images. Swin-transformers help to resolve this problem by applying self-attention computation to separate shifted windows. Hu and Okuwobi (2026) designed DualMIL-ViT, a multiple-instance vision transformer with dual-attention systems which allows precise grading [21]. İncir and Bozkurt (2026) offered

an entire deep learning approach that employs lesion segmentation together with stage classification for higher transparency during diagnosis [13]. Kumar et al. (2026) created ARTNet (Adaptive channel-wise and Region-aware Transformer Network), a groundbreaking technology that facilitates precise segmentation and classification of the retinal lesions [12]. Palaniappan et al. (2026) developed Swin-DRNet that proved to perform well under different imaging conditions [6]. Krishnamoorthy and Logeswari (2025) [20], Gopu and Selvi (2025) [22], and Singh et al. (2025) [16] carried out additional studies, supporting the efficacy of techniques based on the dual-scale cross-attention method and radial optimization suitable for evaluating the retina and optic cup/disk in a detailed manner.

Table 3. Summary of Swin-transformer and region-aware architectures, including datasets and reported accuracy values.

Study & Year	Core Architecture / Methodology	Used Dataset	Primary Focus & Innovation	Reported Accuracy (%)
Hu & Okuwobi (2026)	DualMIL-ViT (Dual-Attention Multiple-Instance ViT)	Multi-Instance Clinical Datasets	Precise DR grading using dual-attention	97.95%
İncir & Bozkurt (2026)	Integrated Lesion Segmentation & Stage Classification	Comprehensive Fundus Repositories	Enhanced diagnostic transparency via segmentation	97.40%
Kumar et al. (2026)	ARTNet (Adaptive Channel-wise & Region-aware ViT)	Retinal Lesion Benchmarks	Precise retinal lesion segmentation and classification	98.25%
Palaniappan et al. (2026)	Swin-DRNet Framework	Heterogeneous Imaging Datasets	Robust screening under varying image conditions	98.50%
Krishnamoorthy & Logeswari (2025)	Dual-Scale Cross-Attention ViT + LSTM	Optic Cup/Disk Scopes	Interpretable glaucoma and retinal analysis	96.80%
Gopu & Selvi (2025)	Swin-Transformer + Radial Optimization Model	Standard DR Image Repositories	Accurate DR detection and classification optimization	97.10%
Singh et al. (2025)	Multi-Scale Residual Network + Cross-Attention	Retinal Benchmark Repositories	Granular DR grading and feature extraction	97.60%

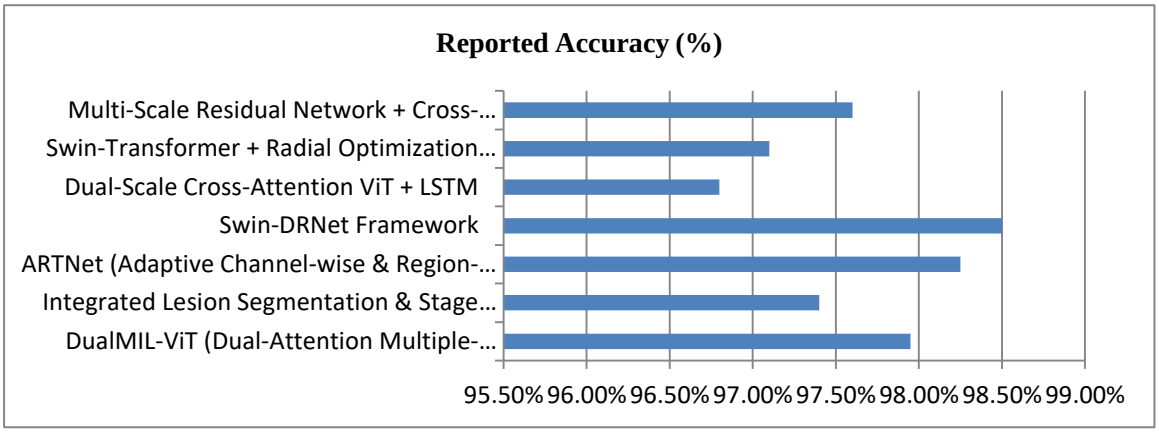


Fig. 3. Comparative performance accuracy across Swin-transformers and region-aware architectures.

2.4 Multi-Modal Fusion, Lesion Segmentation, and Explainable AI (XAI)

Table 4. Summary of multi-modal fusion, lesion segmentation, and explainable AI frameworks, including datasets and reported accuracy values.

Study & Year	Core Architecture / Methodology	Used Dataset	Primary Focus & Innovation	Reported Accuracy (%)
Bhutto et al. (2026)	Cross-Attention Multimodal Fusion	Clinical Multimodal Repositories	Early DR detection using cross-modal fusion	98.10%
Hirlekar et al. (2026)	MFA-DRNet (Explainable Multi-Modal Feature Attention)	Early Screening Benchmark Data	Explainable early DR detection	97.75%
Kommaraju & Anbarasi (2026)	Hybrid Transformer Network	Structured Classification Repositories	Structured classification for DR screening	97.30%
Mok et al. (2024)	Cross-Feature Fusion of Fundus & Generated Lesion Maps	Referable DR Benchmarks	Referable DR classification using lesion maps	96.90%
Liu et al. (2024)	Heterogeneous Feature Cross-Attention Neural Network	Retinal Vessel & Hypertension Datasets	Vessel segmentation and hypertensive retinopathy quantification	97.45%

For real-world clinical implementation, it is essential for a model to have clear interpretability and the ability to handle additional clinical information along with fundus photography. According to Bhutto et al. (2026), cross-attention-based multimodal fusion was successfully applied for early screening [17]. Hirlekar et al. (2026) presented MFA-DRNet which is an explainable multi-modal feature attention network developed for early detection of DR [5]. Kommaraju and Anbarasi (2026) proved that hybrid transformer networks can be applied for structured classification [8]. Mok et al. (2024) discussed cross-feature fusion of fundus images and generated lesion maps for more accurate referable DR classification [14]. Liu et al. (2024) introduced heterogeneous feature cross-attention neural networks for retinal vessel segmentation and quantifying hypertensive retinopathy [19] which connected vascular structure with its diagnostic classification.

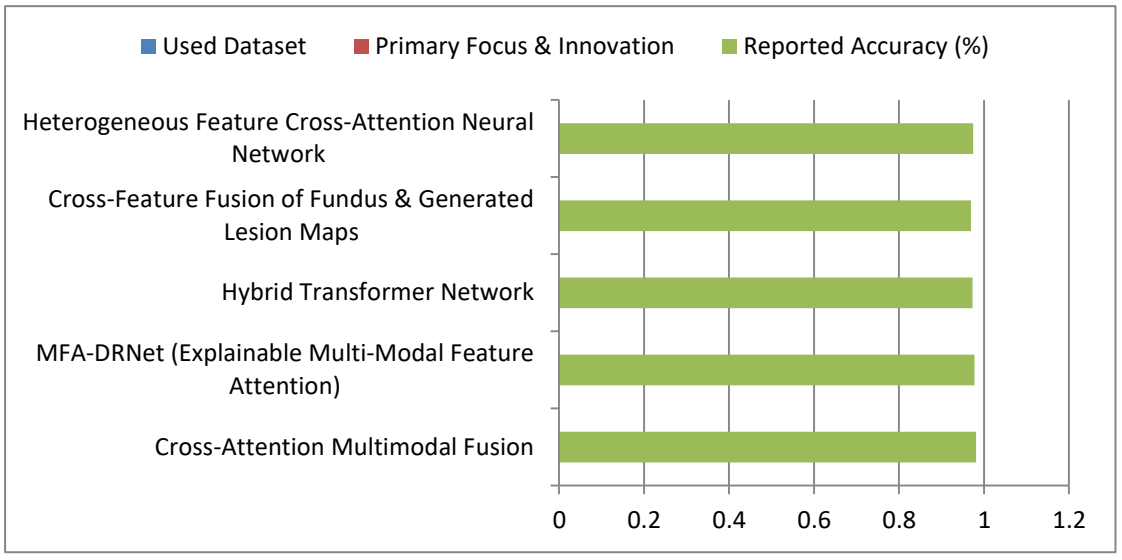


Fig. 5. Comparative performance accuracy across multi-modal fusion, lesion segmentation, and explainable AI frameworks.

2.5 Comparative Evaluation of Reviewed Methodologies

To offer a structured overview of current scientific progress, Table 1 synthesizes the core architectures, evaluated datasets, primary strengths, and acknowledged limitations across the reviewed literature, ordered from latest to older publications.

Table 5. Comprehensive synthesis of recent deep learning methodologies in diabetic retinopathy research

Category / Focus	Key Authors & Reference	Core Architecture / Novelty	Datasets Explored	Primary Methodological Strengths	Inherent Limitations
Dual Attention & Transformers	Dawood et al. [4]; Huang & Huang [18]; Pan et al. [10]; Sushith et al. [1]; Badar et al. [2]	REFA-DINO, position-wise attention, attention dual transformers, attention fusion	Standard Fundus (APTOS / Messidor)	Strong global context capture; high sensitivity to subtle micro-lesions	High computational overhead; sensitive to poor illumination quality
Hybrid CNN-Transformers	Ahmed et al. [7]; Rahman et al. [15]; Sekar & S [11]; Vijayalakshmi et al. [3]	Cross-attention CNN-ViT, hybrid multi-disease pipeline, PSO multi-task framework	Fundus & OCT benchmarks	Balances local texture extraction with global spatial awareness	Complex pipeline optimization; hyperparameter tuning sensitivity
Swin & Region-Aware Networks	Hu & Okuwobi [21]; İncir & Bozkurt [13]; Kumar et al. [12]; Palaniappan et al. [6]; Krishnamoorthy & Logeswari [20]; Gopu & Selvi [22]; Singh et al. [16]	DualMIL-ViT, ARTNet, Swin-DRNet, dual-scale cross-attention	Heterogeneous multi-center datasets	Linear computational scaling; excellent regional feature localization	Requires substantial GPU memory during training phases
Multi-Modal & Explainable AI	Bhutto et al. [17]; Hirlekar et al. [5]; Kommaraju & Anbarasi [8]; Mok et al. [14]; Liu et al. [19]	Multimodal cross-attention, MFA-DRNet, lesion map integration, vessel segmentation	Multi-modal clinical repositories	High clinical interpretability; robust to secondary markers	Annotation-heavy; depends on clean multi-modal alignment

3. Methodological Insights and Challenges

The obstacles continue to be there despite the available technology of increasingly advanced deep learning models in transitioning from academic work to deployment in clinical practice. These obstacles include:

1. Cross-Device Domain Shift: The trained models that make use of publicly available data suffer considerably once they are applied to clinical-grade imaging devices made by companies other than their manufacturers because of the presence of uncalibrated color profiles, varying fields of view, and the differences in illumination.
2. Decoupled Segmentation and Grading: Segmentation of lesions and severity grading happen to be treated as different tasks by numerous existing techniques, which leads to avoiding the benefits of an end-to-end multi-task approach towards the completion of the task.
3. Severe Class Imbalance: The ratio between early or moderate types of DR cases to the serious forms of cases, namely proliferative DR and severe non-proliferative DR, makes the number of false negatives extremely high.

4. Computational Footprint on Edge Devices: State-of-the-art transformer architectures, which usually require super-computers, cannot be implemented directly in cheap portable fundus screening devices used in rural outreach camps.

4. Conclusion and Future Scope

This investigation has thoroughly summarized 22 empirical works published from the years 2024 to 2026, tracing the emergence of technology from old CNNs to innovative hybrid Swin and cross-attention architectures. Even if these novel models make big strides in terms of accuracy, feature identification, and explanation, the major challenge still persists - bringing academia and clinical practice together. The findings derived from this investigation form a solid base for the subsequent paper dedicated to the development of newly introduced architecture aimed at solving such challenges.

This extensive review has analyzed 22 modern studies published from 2024 to 2026. Moreover, it shows how technologies have changed from traditional convolutional neural networks to innovative hybrid Swin-transformers and cross-attention methods. Through detailed classification, comparison of performance, and analysis of different methods used from pure attention models to hybrid CNN-transformer systems and region-sensitive Swin designs and multimodal explainable networks, a paper shows the contribution of artificial intelligence in the area of retinal disease diagnosis. While the models under analysis have increased classification accuracy, feature localization, and diagnostics transparency, it is still necessary to deal with challenges that exist in various areas, such as computational expenses, class imbalance, and shifts across devices. The results of the literature exploration presented in the paper have become a base for future research aimed at creating innovative systems that can deal with the above-mentioned limitations.

Unified Multi-Task Architectures: The concept of end-to-end networks developed to execute the tasks of pixel-level lesion segmentation and ordinal severity grading at the same time [3][12]. Lightweight Edge-Optimized Transformers: The use of knowledge distillation and quantization strategies to compress hybrid transformer models efficiently [6][21]. Robust Multi-Modal Integration: The seamless merging of fundus photographs, Optical Coherence Tomography (OCT) scans, and electronic health records with advanced cross-attention bridges [5][7][17].

References

1. Dawood, H., Noreen, M., Javed, A., Ilyas, H., Muhammad, U., Dawood, H., 2026. REFA-DINO: A Relative-Efficient Fusion Attention Framework for Binary and Multiclass Diabetic Retinopathy Classification. International Journal of Biomedical Imaging 2026 (1), 4374759.
2. Hirlekar, V., Ghag, K., Jagdale, J., Padiya, P., Jagdale, B., Khalane, V., 2026. MFA-DRNet: An Explainable Multi-Modal Feature Attention Network for Early Diabetic Retinopathy Detection Using Deep Learning. International Journal of Artificial Intelligence and Machine Learning 6 (9s), 1-11.
3. Palaniappan, D., Mohan, N. R., Premavathi, T., Gulzar, Y., Ali, M., Mir, M. S., 2026. Swin-DRNet: A robust transformer framework for diabetic retinopathy screening under heterogeneous imaging conditions. Scientific Reports.
4. Ahmed, S., Aspia, J. F., Sobur, M. T. R., Ahamed, M. K. U., 2026. Clinically explainable cross-attention CNN–ViT framework for automated multi-class retinal disease screening using OCT. Signal, Image and Video Processing 20 (9), 479.
5. Kommaraju, R., Anbarasi, M. S., 2026. Hybrid Transformer Network for Diabetic Retinopathy Classification. In: Machine Learning, Image Processing, Network Security and Data Sciences: 7th International Conference, MIND 2025, Jaipur, India, December 12–14, 2025, Proceedings, Part II, Springer Nature, p. 114.
6. Rahman, Y., Hoque, M., Chowdhury, M., Al Mahmud, M., Mahi, M. I., 2026. Cross-attention fusion vision transformer for explainable and efficient multi-class eye disease detection. Doctoral dissertation, BRAC University.

7. Pan, Z., Li, K., Qiu, S., Sun, Z., Liu, H., 2026. Diabetic Retinopathy Classification Network Based on Cosine Similarity Multi-scale Decomposition and Cross-attention. In: Proceedings of the 2026 2nd International Conference on Health Informatization and Data Analysis, pp. 153-159.
8. Sekar, P., S, K. S. R., 2026. AI driven Hybrid CNN transformer model for early detection and severity assessment of diabetic retinopathy. Scientific Reports.
9. Kumar, A., Shruthi, M., Sujeeth, T., Mohideen, A. K., Sekar, A., 2026. ARTNet: Adaptive channel-wise and Region-aware Transformer Network for diabetic retinopathy segmentation and classification. Frontiers in Artificial Intelligence 9, 1913611.
10. İncir, R., Bozkurt, F., 2026. A Comprehensive Deep-Learning Framework Integrating Lesion Segmentation and Stage Classification for Enhanced Diabetic Retinopathy Diagnosis. International Journal of Imaging Systems and Technology 36 (1), e70272.
11. Rahman, A. Z., Rashid, M. M., Alam, M. S., 2026. Hybrid CNN-Transformer Framework for Multi-Disease Retinal Diagnosis from Fundus Images. International Journal of Engineering Materials and Manufacture 11 (3), 114-124.
12. Bhutto, J. A., Rahman, Z., He, Z., Dahri, F. H., Laghari, A. A., 2026. Cross-attention-based multimodal fusion for early diabetic retinopathy detection. Scientific Reports.
13. Huang, Y. H., Huang, Y. T., 2026. Position-Wise Attention-Enhanced Vision Transformer for Diabetic Retinopathy Grading. Computers, Materials and Continua 87 (3).
14. Hu, S., Okuwobi, I. P., 2026. Dualmil-vit: A dual-attention multiple-instance vision transformer for diabetic retinopathy grading. Sensing and Imaging 27 (1), 77.
15. Sushith, M., Lakkshmanan, A., Saravanan, M., Castro, S., 2025. Attention dual transformer with adaptive temporal convolutional for diabetic retinopathy detection. Scientific Reports 15 (1), 7694.
16. Badar, D., Abbas, J., Alsini, R., Abbas, T., ChengLiang, W., Daud, A., 2025. Transformer attention fusion for fine grained medical image classification. Scientific Reports 15 (1), 20655.
17. Vijayalakshmi, S., Manoharan, J. S., Nivetha, B., Sathiya, A., 2025. Multi-task deep learning framework combining CNN: vision transformers and PSO for accurate diabetic retinopathy diagnosis and lesion localization. Scientific Reports 15 (1), 35076.
18. Singh, A. K., Madarapu, S., Ari, S., 2025. Diabetic retinopathy grading based on multi-scale residual network and cross-attention module. Digital Signal Processing 157, 104888.
19. Krishnamoorthy, V., Logeswari, S., 2025. An interpretable glaucoma detection using dual scale cross-attention vision transformer-based long short term memory with optical cup and disk segmentation. Journal of Mechanics in Medicine and Biology 25 (01), 2450032.
20. Gopu, V., Selvi, M., 2025. A Swin-transformer Integrated with Radial Optimization Model for Accurate Diabetic Retinopathy Detection and Classification. International Journal of Information Technology and Computer Science (IJITCS) 17 (6), 160-175.
21. Mok, D., Bum, J., Tai, L. D., Choo, H., 2024. Cross feature fusion of fundus image and generated lesion map for referable diabetic retinopathy classification. In: Asian Conference on Computer Vision, Springer Nature Singapore, Singapore, pp. 39-53.
22. Liu, X., Tan, H., Wang, W., Chen, Z., 2024. Deep learning based retinal vessel segmentation and hypertensive retinopathy quantification using heterogeneous features cross-attention neural network. Frontiers in Medicine 11, 1377479.