Artificial Intelligence in The Diagnosis and Prediction of Diabetic Retinopathy

Hanyu Zheng

{hzhe212@aucklanduni.ac.nz}

Faculty of Science, University of Auckland, Auckland, New Zealand

Abstract. With population growth, the shortage of medical resources is inevitable and irresistible. Interdisciplinary research between artificial intelligence and the medical field is a current research hotspot, and diabetic retinopathy is one of the more mature areas of research. This paper reviews the applications of artificial intelligence (AI) in the diagnosis and prediction of diabetic retinopathy from three perspectives: attention-based mechanisms, feature fusion-based methods, and methods combining both. It also analyzes how each method improves diagnostic efficiency and accuracy. This paper aims to discuss the development and latest achievements of AI in this field, to identify the milestones and future directions of AI development.

Keywords: Diabetic retinopathy, Artificial intelligence, Deep learning.

1 Introduction

In recent years, as the global prevalence of diabetes continues to rise, diabetic retinopathy has become an increasingly significant cause of irreversible blindness among the working population worldwide, highlighting the growing tension between the large-scale clinical demand for early screening and the shortage of medical resources, with artificial intelligence technology emerging as a key solution to this clinical challenge. The digitalization of healthcare is a clear and irreversible trend for the future. The application of deep learning in the diagnosis of diabetic retinopathy has become a research hotspot in the field of artificial intelligence in medicine due to its significant improvement in diagnostic efficiency and accuracy.

With its core capabilities of multimodal feature extraction, massive data learning, and generalized reasoning, large-scale artificial intelligence models have enormous potential in optimizing medical resources. They have become a core research direction in the field of diabetic retinopathy image analysis, propelling the field from traditional deep learning-based single-task image recognition to a higher stage of full-process assisted diagnosis, multi-dimensional information fusion, and clinical decision support. Currently, after fine-tuning and adaptation in the medical field, basic large-scale models have demonstrated performance comparable to that of professional physicians in tasks such as imaging lesion detection,

ICIAAI 2026, May 22-24, Singapore, Singapore

DOI 10.4108/eai.22-5-2026.2365345

pathological classification, and prognosis prediction. Furthermore, the fusion application of multimodal large-scale models has further broken through the limitations of analyzing single-image data, enabling the collaborative interpretation of images, clinical texts, and omics data. Related technologies are gradually moving from laboratory validation to clinical application, becoming a key support for optimizing the allocation of medical resources and improving diagnostic efficiency and accuracy; however, there are also problems such as high equipment prices, poor data compatibility, and high technical barriers, which urgently need to be solved.

This paper will mainly review the results of three technical approaches: attention mechanism-based, feature fusion-based, and a combination of both, clarify the current technological milestones, systematically deconstruct and compare the core technologies, identify the key bottlenecks in current technological development, and aim to provide a reference for algorithm innovation, clinical implementation, and future research directions in this field.

2 Basic information about the disease

Diabetic retinopathy refers to damage to the microvessels of the retina caused by diabetes. It is one of the most common microvascular complications of diabetes and a leading cause of vision impairment and blindness among working- age people worldwide. This lesion is caused by damage to the retinal microvessels due to long-term hyperglycemia. Its pathological features include microaneurysm formation, increased vascular permeability, vascular occlusion, and neovascularization. The damage caused by the lesion can lead to retinal hemorrhage, exudation, and edema. In severe cases, it can lead to retinal detachment and neovascular glaucoma, ultimately resulting in irreversible vision loss. In 1984, Klein et al. [1] found that the prevalence of retinopathy was about 25% in patients with diabetes for less than 5 years, while the prevalence could be as high as 80% in patients with diabetes for more than 15 years. This finding highlighted the importance of early screening and intervention.

The classification and staging of diabetic retinopathy are of great significance for guiding clinical treatment and prognostic assessment. Non-proliferative DR is the early stage of DR, and it is divided into three grades: mild, moderate and severe according to the severity of the lesion (Table 1). Yang proposed in [2] that proliferative DR is the late stage of DR, characterized by the formation of neovascularization of the retina or optic disc.

Table 1. DR classification and its corresponding clinical characteristics.

Classification	Clinical features
Mild NPDR	Only microaneurysms
Moderate NPDR	Microaneurysms with retinal hemorrhage, hard exudates, cotton wool spots, beaded veins, or IRMA, but less severe than the criteria for severe retinal hemorrhage.
Severe NPDR	**Having any of the following features.** (1) Extensive intraretinal hemorrhage in all four quadrants. (2) Beaded veins in two quadrants. (3) Significant IRMA in one quadrant.
PDR	Neovascularization and/or vitreous body/preretinal hemorrhage

Typical fundus changes in diabetic retinopathy include: microaneurysms, retinal hemorrhages, hard exudates, cotton wool spots, beaded veins, and intraretinal microvascular abnormalities. In imaging, it typically manifests as small red dots, punctate, patchy, or flame-shaped marks in the fundus, along with yellowish-white lipid deposits. Diabetic macular edema is one of the leading causes of vision impairment and can occur at any stage.

3 Traditional diagnostic methods

From a diagnostic perspective, in the early period (1980-2000), seven-field stereoscopic fundus photography was used to determine whether lesions had occurred by direct ophthalmoscopy. This method involves taking stereoscopic images of seven 30 ° fields of view after pupil dilation, and then having a professional grader perform visual comparison and scoring on a backlit glass plate using a 15D stereoscopic observer [3]. This method is convenient, but highly dependent on manual labor, and due to the limited field of view, there is a risk of missing peripheral retinal lesions. This method was proposed by Vujosevic et al. in [4].

In the mid-term (2000-2015), non-mydriatic cameras were introduced, and their ultra-wide-angle fundus imaging can display the peripheral retina in a single shot, which helps to detect peripheral ischemia and neovascularization and improve the accuracy of DR diagnosis and classification. Meanwhile, optical coherence tomography, proposed by Ebrahimi et al., has become an important tool for assessing diabetic macular edema, which can quantitatively measure the thickness of the macular retina and identify changes such as cystoid edema and subretinal fluid accumulation [5].

4 Diagnosis and prediction based on artificial intelligence methods

Recently (from 2015 to present), a study published by Gulshan et al. in 2016 demonstrated the potential of deep learning algorithms in DR detection [6]. Since then, AI systems have gradually obtained regulatory approval and have become the first autonomous AI diagnostic system [7].

4.1 Attention-based methods

The attention mechanism simulates the selective attention characteristic of human vision, using adaptive weight allocation to focus the model on lesion areas (such as microaneurysms, hemorrhages, and exudates) in fundus images while suppressing background noise. Mathematically, the attention module learns a weight mapping function:

$$Attention(Q,K,V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{1}$$

Q (query), K (key), and V (value) are linear projections of the feature map, and V is the dimension of the key [8].

From an architectural standpoint, there are three main types: channel attention (such as SE-Net learning channel dependencies through global pooling, and CABNet learning discriminative features for each DR level to address the class imbalance problem) [9], spatial attention (such as CBAM sequentially integrating channel and spatial attention, and GAB capturing global attention features to enhance small lesion detection) [10, 11], and self-attention mechanism (such as ViT segmenting the image into patches and capturing global dependencies through MHSA). In the self-attention mechanism, CTNet adopts a two-stage attention architecture. The convolution module extracts local lesion features through residual connections, capturing the texture information of subtle lesions such as microaneurysms and exudations. The Transformer module divides the feature map into sequential patches and uses the self-attention mechanism to calculate the long-distance dependencies between different lesion regions, establishing global contextual associations. Unlike the traditional ViT which uses CLS tokens, CTNet uses a sequence pooling strategy to directly aggregate the patterns, significantly reducing the number of parameters to 823,555, thus solving the overfitting problem of the standard Transformer on small medical datasets [12]. Wu et al. demonstrated that the pure attention mechanism is applicable to DR classification and identification, achieving an accuracy of 91.4% at multiple resolutions with an AUC of 0.986 [13]. This type of method improves the accuracy of lesion localization, increases the sensitivity to small lesions, and enhances interpretability.

4.2 Feature fusion-based methods

Feature fusion integrates features from different levels, modalities, or network structures to construct a more comprehensive lesion characterization. DR lesions have multi-scale characteristics, and single-scale features are difficult to capture comprehensively.

In terms of fusion strategies, there are three main types: multi-scale feature fusion, cross-modal feature fusion, and cross-network feature fusion. Multi-scale feature fusion integrates the prediction results of different models at the decision level; cross-modal feature fusion combines the color information of 2D fundus photography with the 3D structural information of OCT, and the AUC can reach 0.969[5]. At the same time, it can also integrate fundus image features and clinical indicators to further improve the prediction accuracy; cross-network feature fusion integrates the depth features extracted by CNN with hand-designed features and also integrates the features of different architectures such as ResNet, Inception, and Xception, and weights them by Frobenius norm[14].

In terms of typical architecture, there are three main types: Bi-Directional Feature Fusion, Frobenius Deep Feature Fusion, and multi-level context fusion. In ViT-BiFusionDRNet, the BFF module fuses the global features of ViT with the local textures of CNN through a bidirectional attention mechanism; FDFF extracts feature matrices from two CNN models, selects important features using random forest, calculates feature weights based on the Frobenius norm, and achieves adaptive fusion [14]; in multi-level context fusion, JPU fuses multi-scale pyramid features to capture blood vessels and lesions of different sizes, and ASPP uses dilated convolutions with different dilation rates to capture multi-scale context. These methods achieve multi-scale lesion detection, combine complementary information from different modalities/networks to improve robustness, and the fusion model achieves an accuracy of 98.10% on the EyePACS and Messidor datasets [15].

4.3 Attention-based and feature fusion-based methods

This method integrates the adaptive weighting capability of the attention mechanism with the multi-source information integration advantage of feature fusion to construct a "selective fusion" framework. By guiding the fusion process through the attention mechanism, the model can dynamically decide which feature sources to focus on and which noise to suppress based on the content of the input image.

In terms of typical architecture, there are three main types: Hybrid CNN-Transformer with Cross-Attention Fusion, Attention-Guided Multi-Scale Feature Fusion, and Temporal-Spatial Attention Fusion. Hybrid CNN-Transformer with Cross-Attention Fusion consists of a CNN branch, a Transformer branch, and a cross-attention fusion module. Mathematical expression:

$$F_{fused} = CrossAttention(F_{CNN},\ F_{ViT}) = softmax\left(\frac{Q_{CNN}K_{ViT}^{T}}{\sqrt{d}}\right)V_{ViT} \tag{2}$$

Among them, CvT-13 introduces convolutional token embedding and convolutional projection in ViT, combining the locality of CNN with the globality of Transformer, achieving QWK=0.84 and AUC=0.93 on the APTOS dataset. The LeViT-256 hybrid architecture achieves 83.7% top 1 accuracy with only 1.1 GFLOPs on ImageNet-1k and performs well when transferred to the DR classification task [16]. D-TNet uses a 4-layer Transformer encoder and calculates the attention scores of features such as cotton spots and neovascularization through self-attention [17].

HybridFusionNet in Attention-Guided Multi-Scale Feature Fusion combines Self-Attention Network (SAN) and a Vision Transformer. SAN calculates the relationship between pixels to generate attention values, and ViT processes block images to capture long-distance dependencies. It achieves 99% accuracy in the 5-class DR task [18]. DL-LAHA nested V-Net segments hemorrhage, exudation, and microaneurysms. EfficientNet + RegNetY dual-path feature extraction and entropy-based feature selection followed by fusion achieve 98.65% accuracy on the MESSIDOR dataset [19].

In Temporal-Spatial Attention Fusion, ADT-ATC's dual-space Transformer (small-scale 16×16 + large-scale 64×64) and adaptive temporal convolutional memory unit fuse spatiotemporal features through hierarchical cross attention, achieving 98.2% and 97.4% accuracy on the DRIVE and DR datasets, respectively [20]. TADDL uses CNN to extract spatial features, LSTM, and an attention mechanism to capture temporal dependencies, and multi-scale convolutional paths to extract features of different granularities, achieving 97.5% accuracy on the DRIVE dataset [21].

These methods achieve dynamic adaptability, dynamically adjust the fusion strategy according to the content of the input image, and use different fusion weights for different lesion types; global-local collaboration, the local details captured by CNN and the global context modeled by Transformer mutually enhance each other; multi-task support, can simultaneously complete classification, segmentation and lesion localization, meeting diverse clinical needs; performance breakthrough, the hybrid architecture outperforms pure CNN and pure Transformer models on multiple benchmark datasets [10].

4.4 Existing limitations and prospects

Current methods primarily rely on static fundus images for diagnosis, making it difficult to capture the dynamic characteristics of DR evolution over time. Furthermore, the generalization performance of most models on cross-device and cross-ethnic data remains to be validated. Future research should focus on constructing a temporal image analysis framework to support disease progression prediction and developing lightweight, interpretable models to facilitate clinical deployment and regulatory approval.

5 Conclusions

Diagnostic methods for diabetic retinopathy (DR) have undergone three major paradigm shifts, from seven-field stereoscopic imaging to non-mydriatic cameras, OCT, and finally AI-automated deep learning. Currently, artificial intelligence technology has achieved initial breakthroughs in DR image screening and lesion grading, demonstrating efficiency and consistency surpassing traditional manual diagnosis.

Attention mechanisms enhance the model's sensitivity to small lesions and the interpretability of its decisions, primarily used to address class imbalance and missed detection of small lesions in DR diagnosis. Feature fusion enables multi-scale, cross-modal, and cross-network complementary information, effectively improving the model's robustness and generalisation ability. Combining attention mechanisms and feature fusion, represents the forefront of the field, maintaining high accuracy while gaining clinical acceptance through enhanced interpretability. It combines CNN's ability to capture local details with Transformer's capability to model global context, demonstrating optimal overall performance on several mainstream public datasets such as EyePACS, Messidor, and APTOS, and represents the current frontier and mainstream technological paradigm in this field.

In the future, for the integration of diabetic retinopathy and artificial intelligence to achieve a leap from image screening to full-cycle management, and from technological breakthroughs to widespread clinical application, it is necessary to focus on data privacy and security, enhance clinical interpretability, and undergo long-term clinical validation to ultimately alleviate problems such as insufficient and unevenly distributed medical resources. Future research in this field can focus on the dynamic evolution and progression prediction of the disease course (using sequential retinal imaging data from continuous patient follow-ups to construct models of DR disease progression and capture lesion progression patterns), enhancing multi-centre cross-domain generalisation (using federated learning, attention-guided domain adaptation, cross-centre feature alignment, etc., to improve the clinical universality of models), development of lightweight deployable models for primary care (compressing model parameters and computation to reduce dependence on high-performance hardware), and building clinically interpretable and compliant models (integrating lesion localisation, attention heatmap visualisation, and decision traceability modules in a hybrid architecture to make model diagnoses interpretable and traceable).

References

[1] Klein, R., Klein, B.E., Moss, S.E., Davis, M.D., DeMets, D.L.: The Wisconsin Epidemiologic Study of Diabetic Retinopathy: II. Prevalence and risk of diabetic retinopathy when age at diagnosis is less than 30 years. Arch Ophthalmol. 102(4), pp. 520-526 (1984)

[2] Yang, Z., Tan, T.E., Shao, Y., et al.: Classification of diabetic retinopathy: Past, present and future. Front Endocrinol. 13, pp. 1079217 (2022)

[3] Slakter, J.S., Schneebaum, J.W., Shah, S.A.: Digital Algorithmic Diabetic Retinopathy Severity Scoring System (An American Ophthalmological Society Thesis). Trans. Am. Ophthalmol. Soc. 113, pp. T9 (2015)

[4] Vujosevic, S., Aldington, S.J., Silva, P., et al.: Screening for diabetic retinopathy: new perspectives and challenges. Lancet Diabetes Endocrinol. 8(4), pp. 337-347 (2020)

[5] Ebrahimi, B., Le, D., Abtahi, M., Dadzie, A.K., Lim, J.I., Chan, R.V.P., Yao, X.: Optimizing the OCTA layer fusion option for deep learning classification of diabetic retinopathy. Biomed. Opt. Express. 14(9), pp. 4713-4724 (2023)

[6] Jiang, X., Xie, M., Ma, L., Dong, L., Li, D.: International publication trends in the application of artificial intelligence in ophthalmology research: an updated bibliometric analysis. Ann. Transl. Med. 11(5), pp. 219 (2023)

[7] Sedova, A., Hajdu, D., Datlinger, F., et al.: Comparison of early diabetic retinopathy staging in asymptomatic patients between autonomous AI-based screening and human-graded ultra-widefield colour fundus images. Eye. 36, pp. 510-516 (2022)

[8] Sushith, M., Lakkshmanan, A., Saravanan, M., et al.: Attention dual transformer with adaptive temporal convolutional for diabetic retinopathy detection. Sci. Rep. 15, pp. 7694 (2025)

[9] He, A., et al.: CABNet: Category Attention Block for Imbalanced Diabetic Retinopathy Grading. IEEE Transactions on Medical Imaging. 40(1), pp. 143-153 (2021)

[10] Zhang, W., Belcheva, V., Ermakova, T.: Interpretable Deep Learning for Diabetic Retinopathy: A Comparative Study of CNN, ViT, and Hybrid Architectures. Computers. 14(5), pp. 187 (2025)

[11] Woo, S., et al.: CBAM: Convolutional Block Attention Module. In: European Conference on Computer Vision (ECCV), pp. 3-19 (2018)

[12] Islam, M.T., et al.: CTNet: A Convolutional Transformer Network for Diabetic Retinopathy Classification. IEEE Access. 10, pp. 102209-102222 (2022)

[13] Wu, J., Hu, R., Xiao, Z., Chen, J., Liu, J.: Vision Transformer-based recognition of diabetic retinopathy grade. Med. Phys. 48(12), pp. 7850-7863 (2021)

[14] Priyadharsini, C., Phamila, Y.A.V.: Frobenius deep feature fusion architecture to detect diabetic retinopathy. Discov. Appl. Sci. 7, pp. 194 (2025)

[15] Bappi, M.I., Juthy, J.A., Kim, K.: Deep learning-based diabetic retinopathy recognition and grading: Challenges, gaps, and an improved approach — A survey. ICT Express. 11(5), pp. 993-1013 (2025)

[16] Zhang, W., Belcheva, V., Ermakova, T.: Interpretable Deep Learning for Diabetic Retinopathy: A Comparative Study of CNN, ViT, and Hybrid Architectures. Computers. 14(5), pp. 187 (2025)

[17] Sekar, P., Bhoopalan, R., Nagaprasad, N., et al.: D-TNet: a hybrid Dense Net-transformer model for robust diabetic retinopathy detection. Sci. Rep. 15, pp. 39594 (2025)

[18] Shukla, A., Tiwari, S., Jain, A.: HybridFusionNet: Deep Learning for Multi-Stage Diabetic Retinopathy Detection. Technologies. 12(12), pp. 256 (2024)

[19] Karuppusamy, K., Thangavelu, B., Subramaniam, K.M., et al.: Deep Learning-Based Lesion-Aware Hybrid Attention Framework for Diabetic Retinopathy Classification. Int. J. Comput. Intell. Syst. 19, pp. 33 (2026)
[20] Sushith, M., Lakkshmanan, A., Saravanan, M., et al.: Attention dual transformer with adaptive temporal convolutional for diabetic retinopathy detection. Sci. Rep. 15, pp. 7694 (2025)
[21] Sushith, M., et al.: A hybrid deep learning framework for early detection of diabetic retinopathy using retinal fundus images. Scientific Reports. 15(1), pp. 99309 (2025)