# Deep Learning Models for Diabetic Retinopathy Detection: A Review of CNN and Transformer-Based Approaches

Guanglongyu Huo
*Chengdu University of Technology, Erxianqiao Street, Chengdu City, Sichuan Province, China*

Keywords: Diabetic Retinopathy, Deep Learning, Convolutional Neural Network, Transformer, Diagnosis.

Abstract: This article reviews the progress of many deep learning neural network models in the detection of diabetes retinopathy (DR), and discusses the significance of these advances in clinical practice. It examined improved Convolutional Neural Networks (CNN) and Transformer based DR detection models. Models based on CNN, such as Romero Oraa's framework, two-layer neural networks, and weighted fusion deep learning networks, have shown promising results in addressing challenges such as lighting and image quality. Transformer based models, including dual transformer encoder models and self supervised image transformers, utilize their unique architecture to improve performance. These models improve the accuracy and efficiency of diagnosis, promoting the effectiveness of early intervention for DR treatment. In addition, the integration of these advanced technologies not only simplifies the diagnostic process, but also has the potential to alleviate the burden on the healthcare system by providing scalable solutions for extensive screening, ultimately helping to improve patient outcomes.

## 1 INTRODUCTION

The leading cause of blindness worldwide is diabetic retinopathy (DR), which is an important complication of diabetes (Solomon, 2017). It is found by Wong et al. that diabetic retinopathy occurs in about one-third of people with diabetes, with severe forms including proliferative DR and diabetic macular edema (2016). DR is associated with prolonged diabetes, high blood sugar, and hypertension. While traditionally viewed as a microvascular disease, it also involves retinal neurodegeneration. The development of DR is driven by complex mechanisms related to hyperglycemia, including genetic factors, free radicals, and inflammatory mediators. Effective control of blood glucose and blood pressure is crucial for prevention. Treatments like anti-VEGF therapy and laser photocoagulation can help manage vision loss. Increased public awareness and regular screenings are essential for improving outcomes and preventing blindness in DR patients.

As the prevalence of diabetes continues to rise, the need for effective screening and diagnostic tools becomes increasingly critical. Traditional methods of DR detection often rely on manual examination of retinal images, which can be time-consuming and subject to human error. In recent years, the way for automated systems that can enhance the accuracy and efficiency of diabetic retinopathy diagnosis has been paved by advancements in deep learning and artificial intelligence.

This paper reviews various improved convolutional neural network(CNN) models and Transformer based for the detection and classification of DR. By utilizing innovative architectures and techniques, these models are designed to improve diagnostic performance while addressing challenges such as class imbalances, new image perspectives, and the need for generalization across diverse populations and imaging conditions.

Through a comprehensive analysis of recent literature, this review emphasizes the progress of deep learning in detecting diabetic retinopathy and deliberates on the significance of these advancements for clinical practice. By summarizing the most innovative approaches and their performance metrics, this work aims to provide insights for future research directions in this critical area of medical imaging. Ultimately, the goal is to facilitate the development of more reliable and accessible tools for early detection and intervention in diabetic retinopathy, thereby improving patient outcomes and reducing the burden of this preventable disease.



Huo and G.
Deep Learning Models for Diabetic Retinopathy Detection: A Review of CNN and Transformer-Based Approaches.
DOI: 10.5220/0013533700004619
In *Proceedings of the 2nd International Conference on Data Analysis and Machine Learning (DAML 2024)*, pages 594-598
ISBN: 978-989-758-754-2

## 2 CNN-BASED MODELS

The development of various improved CNN-based models for DR detection have been led by recent advancements in deep learning, each addressing specific challenges associated with fundus image analysis. These models leverage different techniques to enhance accuracy, efficiency, and robustness in diagnosing DR.

In the field of diabetic retinopathy (DR) detection, a novel framework that harnesses the power of deep learning techniques. This framework is designed to automate the detection and grading of DR, a crucial step in early diagnosis. Incorporating an attention mechanism the model can focus on different aspects of the retinal images (Romero - Oraa et al 2024). Specifically, it separates dark structures and bright structures. This focused attention enhances the classification accuracy as it enables the model to better distinguish between various features and patterns associated with DR. The framework further decomposes the input images. This decomposition serves to improve the visibility of lesions within the images. Additionally, it generates interpretable attention maps. These maps provide valuable insights into the model's predictions, allowing clinicians to better understand how the model arrived at its conclusions. Verified on the Kaggle DR Detection dataset, the accuracy of the model is 83.7%, and 0.78 on the quadratic weighted Kappa. These results are significant as they outperform several state-of-the-art methods. This performance makes the framework a valuable diagnostic tool for clinicians, aiding in the early detection and grading of DR.

The bilayered neural network is another notable approach in DR detection. It employs a feedforward architecture with two fully connected layers (Islam, 2021). Retinal images often present challenges due to varying illumination and fields of vision. These factors can complicate accurate detection of DR. The bilayered approach is designed to address these challenges. It allows for enhanced feature extraction and classification, enabling the model to better discern subtle differences in DR severity.Through its unique design, the model is able to learn complex patterns in fundus images. This is reflected in its performance on the test set, where it achieved an accuracy of 93.33%. The incorporation of resubstitution validation further optimizes its performance. As a result, it holds promise as a solution for automated DR detection in clinical settings.

In contrast, weighted fusion deep learning network(WFDLN) tackles the challenges posed by low - quality fundus images, a common issue in DR diagnosis (Nneji, 2022). The dual-channel scans, namely the CLAHE which stands for contrast-limited adaptive histogram equalization images and CECED which is contrast-enhanced canny edge detection images, are processed by the network. This approach allows it to handle the complexity of low - quality images more effectively. WFDLN utilizes fine - tuned Inception V3 and VGG - 16 for feature extraction. Impressive performance metrics are the result of this combination. On the dataset of Messidor, It achieved an accuracy corresponding to 98.5 per hundred, a sensitivity of 98.9 per hundred, and a specificity of 98.0 per hundred. These results highlight its effectiveness in accurate and automated DR classification, demonstrating high accuracy and robustness while addressing common challenges in fundus image analysis.

Additionally, A deep learning - based approach has been developed to predict the risk of developing referable DR. This model utilizes a substantial dataset of 156,363 fundus images from the EyePACS database (Bora, 2021). The model's ability to predict the risk of developing referable DR is a significant advancement. When averaging scores from multiple images, it reaches an AUC value of 0.81. This indicates the potential for personalized risk assessments, which can enhance screening strategies and enable timely interventions.

The BigAug method introduces a novel approach to generalizing deep learning models for medical image segmentation across unseen domains by applying random transformations that improve robustness against variations in medical imaging data (Zhang, 2022). This approach is crucial as medical imaging data can vary significantly. Evaluations of the BigAug method demonstrate competitive performance comparable to fully supervised models. This significant contribution enhances the adaptability of deep learning techniques in diverse clinical settings, particularly in the automatic processing of fundus images for diabetic retinopathy grading, thereby improving diagnostic accuracy and efficiency.

While not exclusively focused on diabetic retinopathy, the IVGG13 model modifies the VGG16 architecture for pneumonia classification but showcases how architectural improvements can enhance training efficiency and classification performance (Jiang, 2021). This modification showcases how architectural improvements can enhance training efficiency and classification performance. It highlights the broader applicability of refined architectures in medical image analysis,

providing insights that can be applied to DR detection models as well.

Lastly, DiaNet is a dedicated architecture designed to diagnose diabetes from retinal photographs using CNNs to extract features effectively (Islam, 2021). It uses CNNs to extract features effectively from retinal photographs. The experimental results indicate significant accuracy, establishing DiaNet as a promising non - invasive diagnostic tool. This emphasizes the potential of retinal imaging in identifying diabetes - related health risks, which is relevant to the context of DR as diabetes is a precursor to DR.

This section details various improved CNN-based models for diabetic retinopathy detection. Romero-Oraa's framework uses an attention mechanism and image decomposition, achieving good accuracy on the Kaggle dataset. The bilayered neural network addresses illumination and vision challenges, with high test set accuracy. WFDLN tackles low-quality images well, showing strong performance on the Messidor dataset. A model using EyePACS data predicts DR risk. The BigAug method generalizes models. The IVGG13 model shows architectural benefits, and DiaNet diagnoses diabetes from retinal images effectively, all contributing to advancements in DR detection and related medical imaging tasks.

## 3 TRANSFORMER-BASED MODELS

Recent progress in architectures based on transformers has greatly improved the detection and segmentation of DR by means of diverse innovative methods. These models leverage the strengths of transformers, such as the ability of them to capture long-range dependencies. and contextual information, which are crucial for accurately analyzing retinal images.

A novel dual transformer encoder model specifically designed for medical image classification. It addresses the issue of fixed token size in Vision Transformer (ViT) by utilizing two encoders with different hidden sizes, enabling the model to better adapt to various medical image types (Yan, Yan, & Pei, 2023). It addresses the issue of fixed token size in Vision Transformer (ViT) by utilizing two encoders with different hidden sizes. This enables the model to better adapt to various medical image types, a crucial factor in accurately analyzing retinal images which can vary in size and content. The key contribution of this model lies in its proposed dual transformer encoder architecture. This architecture enhances the ability of them to capture long-range dependencies within medical images. Long - range dependencies are important for understanding the overall context and relationships within the image, which is essential for accurate classification. The model employs the LCA module to integrate features from all encoder layers. This integration leads to improved classification performance as it combines the information from different levels of the model's processing.The Dual Transformer Encoder Model features a unique architecture with two transformer encoders of varying hidden sizes, enabling improved multi - scale feature extraction and efficient training. It has demonstrated superior robustness across multiple datasets compared to single transformer encoders and traditional CNNs. This shows its potential for applications in real-world medical imaging, particularly in the detection of diabetic retinopathy.

Different from dual transformer encoder model utilizing two encoders with different hidden sizes, a new method to classify diabetic retinopathy (DR) using a masked autoencoder (MAE) enhanced visual transformer (ViT) (Yang, 2024). The model addresses the challenge of limited data by using a large dataset of more than 100,000 fundus images that are larger than the typical ViTs input size. The key contribution of the model is its ability to utilize self-supervised learning through MAE, which enables ViT to efficiently capture rich features from retinal images. This approach improves the model's performance in terms of referable DR Classification, especially when training data is scarce. By pre-training retinal images, the model reduces overfitting and improves generalization ability. Compared with traditional methods, the architecture shows higher classification accuracy, indicating its practical application potential in the detection of diabetic retinopathy. Overall, the integration of mask autoencoders with ViT offers a promising avenue to advance automated diagnosis of DR, especially in clinical Settings where rapid and accurate assessment is critical.

A prominent method is the Self - Supervised Image Transformer (SSiT), which utilizes self - supervised learning techniques guided by saliency maps (Huang, 2022). It enables the model to pre - train on large datasets without the need for a large amount of labeled data. This is a significant advantage as obtaining labeled data for medical images can be difficult and time-consuming. The model focuses on salient regions within fundus images and employs tasks such as saliency - guided contrastive learning

and segmentation prediction to enhance feature representation. By focusing on these important regions, the model can better understand the key features and patterns associated with DR.Under the detection of publicly available datasets on multiple platforms, the model still maintains good performance and accuracy. This shows its ability to adapt to different datasets and platforms, although challenges related to data requirements, background complexity, initialization sensitivity, generalization capabilities, and the need for domain - specific knowledge must be addressed to enhance its practical application in clinical settings.

The residual visual deformers (ResViT) model uses a generative adversarial framework which combines convolution operators with visual deformers (Dalmaz, 2022). This hybrid design combines polymerized leftover transformer blocks to enhance feature representation while maintaining computational efficiency. ResViT excels in medical image synthesis. ResViT's multimodal inheritance capability enables it to process medical images of different modes, which makes it superior to SSiT for different image processing. However, the complexity of its application and the huge occupation of computing resources make it difficult to apply it in the detection of diabetic retinopathy.

Another notable architecture is the Compact Convolution Transformer (CCT), which is specifically designed for efficient DR Detection using low-resolution images (Khan, 2023). By combining convolution tokenization with the transformer backbone, CCT achieved 84.52% accuracy, saving computational resources compared to ResNeT and other traditional models, while reducing training time without compromising diagnostic accuracy.

The CNN and MLP Mixed Transformer Model is a hybrid method that integrates Convolutional Neural Networks (CNNs) to extract features, employs transformers to capture the global context, and uses Multi-Layer Perceptrons (MLPs) for classification. (Kumar & Karthikeyan, 2021). This model addresses class imbalance through a custom loss function and achieves notable accuracy of 90.17% in predicting DR severity, highlighting its effectiveness in managing the complexities of retinal images.

Lastly, The RTNet Model focuses on enhancing segmentation accuracy by employing a dual - branch structure that includes a Global Transformer Block for extracting global features and a Relation Transformer Block for capturing interdependencies between lesion features and vascular patterns. (Huang, 2022). The GTB focuses on extracting global features, while the RTB emphasizes the relationships between different lesions and their connections to vascular structures. This dual approach enables a more thorough comprehension of the spatial connections in retinal images. This is essential for accurate segmentation. RTNet has achieved competitive performance on benchmark datasets like IDRiD and DDR, although it encounters challenges that are associated with limitations of the dataset and the requirement for comprehensive pixel-level annotations.

This section is centered around models based on transformers for the detection and segmentation of diabetic retinopathy (DR). The dual transformer encoder model addresses the token size issue in ViT and has a unique architecture that enhances capturing long-range dependencies, showing robustness for DR detection. The Self - Supervised Image Transformer (SSiT) uses self-supervised learning with saliency maps, enabling pre-training on large datasets, but faces challenges for clinical use. MAE-enhanced ViT uses a large dataset to address data limits. Self-supervised learning via MAE helps capture features, improving classification, especially with scarce data. It shows higher accuracy than traditional methods, offering promise for DR diagnosis.Residual Visual Deformers (ResViT) has a hybrid design for feature representation but is complex and resource-intensive for DR detection. The Compact Convolution Transformer (CCT) is designed for low-resolution images, achieving good accuracy while saving resources. The CNN and MLP Mixed Transformer Model combines different techniques to address class imbalance and predict DR severity effectively. The RTNet Model uses a dual-branch structure for better segmentation accuracy, facing dataset and annotation challenges. These models each have unique features and challenges, contributing to the advancement of DR detection and segmentation.

# 4 CONCLUSIONS

In conclusion, this review has explored the advancements in deep learning models for diabetic retinopathy detection. The improved CNN-based models, such as Romero-Oraa's framework, bilayered neural network, and weighted fusion deep learning network, have demonstrated effectiveness in addressing challenges related to fundus image analysis and have achieved good diagnostic accuracy. Transformer-based models, including the dual transformer encoder model, Self - Supervised Image Transformer, and others, have also shown significant progress in capturing Long-range dependencies and

contextual information enhance diabetic retinopathy detection and segmentation.

These models offer promising solutions for enhancing the accuracy and efficiency of diabetic retinopathy diagnosis, which is crucial in the face of the increasing prevalence of diabetes and the need for early detection and intervention. Nevertheless, challenges remain, such as data requirements, background complexity, initialization sensitivity, generalization capabilities, and the need for domain-specific knowledge, especially for some transformer-based models. Future research ought to concentrate on tackling these challenges in order to further enhance the practical application of these models in clinical environments and ultimately make a contribution to better patient results and a reduction in the burden of DR.

# REFERENCES


Bora, A. et al. (2021) Predicting the risk of developing diabetic retinopathy using deep learning, *The Lancet Digital Health*, 3, pp.10-19. Available at: https://doi.org/10.1016/ S2589-7500(20)30250-8 *Contributed equally

Dalmaz, O. et al. (2022) ResViT: Residual Vision Transformers for Multimodal Medical Image Synthesis, *TRANSACTIONS ON MEDICAL IMAGING*, 41(10), pp.2598-2614. Available at: https://ieeexplore.ieee.org/document/9758823.

Huang, S. et al. (2022) RTNet: Relation Transformer Network for Diabetic Retinopathy Multi-Lesion Segmentation, *TRANSACTIONS ON MEDICAL IMAGING*, 41(6), pp.1596-1607. Available at: https://ieeexplore.ieee.org/document/9684442.

Huang, Y. et al. (2022) SSiT: Saliency-guided Self-supervised Image Transformer for Diabetic Retinopathy Grading, Available at:https://arxiv.org/abs/2210.10969

Islam, M. T. et al. (2021) DiaNet: A Deep Learning Based Architecture to Diagnose Diabetes Using Retinal Images Only, *Digital Object Identifier*, 9, pp.15686-15695. Available at: https://ieeexplore.ieee.org/document/9328261

Jiang, Z. et al. (2021) An Improved VGG16 Model for Pneumonia Image Classification, *Applied Science*, 11, pp.1-19. Available at: https://doi.org/10.3390/ app112311185

Khan, I U. et al. (2023) A Computer-Aided Diagnostic System to Identify Diabetic Retinopathy, Utilizing a Modified Compact Convolutional Transformer and Low-Resolution Images to Reduce Computation Time, *Biomedicines2023,* 11, Available at: https://doi.org/10.3390/biomedicines11061566.

Kumar, N. and Karthikeyan, R. (2021) Diabetic Retinopathy Detection using CNN, Transformer and MLP based Architectures, *International Symposium on Intelligent Signal Processing and Communication Systems*, DOI: 10.1109/ISPA CS51563.2021.965102.

Nneji, G U. et al. (2022) Identification of Diabetic Retinopathy Using Weighted Fusion Deep Learning Based on Dual-Channel Fundus Scans, *Diagnostics, 12*, Available at: https:// doi.org/10.3390/diagnostics12020540

Romero-Oraa, R. et al. (2024) Attention-based deep learning framework for automatic fundus image processing to aid in diabetic retinopathy grading, *Computer Methods and Programs in Biomedicine*, 249, Available at:https://doi.org/10.1016/j.cmpb.2024.108160.

Solomon, D S. et al. (2017) Diabetic Retinopathy: A Position Statement by the American Diabetes Association, *Diabetes Care*, 40 pp.412-418. DOI:10.2337/dc16-264.

Wong, T. et al. (2016) Diabetic retinopathy. *Nat Rev Dis Primers,* 2, Available at: .https://doi.org/10.1038/nrdp.2016.12

Yan, F., Yan, B. and Pei, M. (2023) DUAL TRANSFORMER ENCODER MODEL FOR MEDICAL IMAGE CLASSIFICATION, *International Conference on Image Processing*, DOI: 10.1109/ICIP49359.2023.10222303

Yang, Y. et al. (2024) 'Vision transformer with masked autoencoders for referable diabetic retinopathy classification based on large-size retina image', *PLoS ONE,* 19(3), Available at: https://doi.org/10.1371/journal. pone.0299265

Zhang, L. et al. (2020) 'Generalizing Deep Learning for Medical Image Segmentation to Unseen Domains via Deep Stacked Transformation', TRANSACTIONS ON MEDICAL IMAGING, 39(7), pp. 2531-2540. Available at:https://ieeexplore.ieee.org/abstract/document/8995481