Int. J. Adv. Res. Sci. Technol. Volume 15, Issue 5, 2026, pp.1915-1918.

IJARST

International Journal of Advanced Research in Science and Technology

https://doi.org/10.62226/ijarst20262660

ISSN 2319 – 1783 (Print)

ISSN 2320 – 1126 (Online)

DIABETIC RETINOPATHY DETECTION SYSTEM USING RETINAL IMAGES

Pankhuri Verma, Shivam Kumar, Er Shilpi Khanna

Department of Information Technology, Shri Ramswaroop Memorial College of Engineering and Management (SRMCEM) Lucknow, India

vermapankhuri202429@gmail.com, shivam985kumar@gmail.com, shilpikhanna@srmcem.ac.in

ARTICLEINFO

Article history:

Received 29 Apr 2026

Accepted 07 May 2026

Available online 11 May 2026

Keywords:

Diabetic Retinopathy, Vision Transformer, Convolutional Neural Networks (CNNs), Retinal Fundus Images, Computer-Aided Detection, Image Classification

Indexed in:

and in major libraries

ABSTRACT

Diabetic Retinopathy is a disease that impacts the retina and happens after a long duration of diabetes. Diabetic Retinopathy is one of the primary causes of blindness worldwide. It is necessary to diagnose the retinal disease at an early stage by analyzing the images of the fundus of the eye with the help of software solutions to ensure appropriate treatment to prevent blindness. Recently, artificial intelligence has helped the researchers to implement computer-aided detection systems for Diabetic Retinopathy. This paper digs into software solutions for detecting Diabetic Retinopathy, zeroing in on Vision Transformer models. We compared different ways of preprocessing retinal fundus images, along with various feature extraction and classification methods. When you stack Vision Transformers up against Convolutional Neural Networks (CNNs), Vision Transformers really shine. They handle global features better — they don't just focus on small patches. That's a big deal when you're analyzing something as intricate as retinal images. This paper focuses on Vision Transformer models and compares them with classical and CNN-based approaches for diabetic retinopathy detection.

© 2026 International Journal of Advanced Research in Science and Technology (IJARST). All rights reserved.

I. INTRODUCTION

Artificial intelligence and deep learning have really changed the game for medical image analysis lately. CNNs like ResNet, VGGNet, and EfficientNet have hit impressive accuracy levels with automated Diabetic Retinopathy detection. These models pick up hierarchical spatial features from retinal images and can catch pathological patterns pretty well. But there's a catch with CNNs. They mainly focus on local details and use local receptive fields, which can leave them blind to relationships spread out across the whole image. That's a problem since retinal abnormalities don't always pop up just in one spot — they're often scattered in different regions. To get around this, transformer-based models have stepped into the spotlight. Vision Transformers (ViT) swap out convolution operations for self-attention mechanisms, which look at how different patches of the image relate to each other, everywhere. That means they're way better at catching those long-range connections — and that's exactly what you want for something like diabetic retinopathy grading. Bringing together retinal imaging with these transformer-based deep learning models is leading us toward automated, scalable DR screening systems. They can help flag issues early and cut down on preventable blindness. The rapid growth in AI and deep learning has totally changed how we approach medical image analysis, and CNNs like ResNet, VGGNet, and EfficientNet have already proven they're solid for DR detection. They're great at mapping out spatial hierarchies in images and spotting abnormalities. Still, their focus on local details means they aren't the best at picking up broader connections that matter in medical images, especially retinal ones with scattered issues. That's why researchers are looking toward transformer models, which are much better at seeing the big picture. Vision Transformers now show real promise for tasks like DR detection, where those long-range connections are critical. Using retinal images with deep learning models like transformers opens the door to more efficient DR detection — and, just as important, early diagnosis.

II. LITERATURE SURVEY

These days, researchers are blending CNN backbones with transformer encoders, hoping to grab the best parts of each

— sharp local feature extraction from CNNs, plus a wider, global perspective from transformers.

If you check out the progress in automated diabetic retinopathy (DR) detection over the last ten years, it's impressive. Most of the work falls into three main buckets:

- Classical image processing
- CNN-based deep learning
- Vision Transformers

A. Classical Image Processing Methods

In the early days, DR detection leaned heavily on classic computer vision — basically, hand-crafted features and straightforward processing. The focus was on spotting DR lesions: microaneurysms, hemorrhages, exudates.

Typical techniques looked like this:
- Segmenting blood vessels
- Using morphological filters
- Applying intensity thresholding
- Texture analysis with Gabor filters
- Microaneurysm detection with top-hat transforms and region growing
- Exudate detection via clustering and adaptive thresholding

These methods made sense and weren't too hard to code, but they came with headaches. They struggled when lighting changed or image quality dropped, and you couldn't always count on them. Also, as people pushed these systems to work at scale, the flaws got even more obvious.

B. CNN-Based Deep Learning Models

Deep learning totally flipped the script. With CNNs, models automatically learned which features mattered in retinal images, so nobody had to spend hours manually engineering features.

Popular CNNs include:
- VGG16
- InceptionV3
- ResNet

CNNs brought real boosts: better lesion detection, stronger results from transfer learning, and smarter, more dependable features.

C. Vision Transformer-Based Approaches

Transformers started out shaking up language processing. But now, they're doing cool stuff in vision, too.

Vision Transformers slice up images into small patches — sort of like breaking a paragraph into words. With self-attention, the model figures out how these patches relate, seeing the whole image instead of just zooming in on tiny chunks.

Recent studies highlight some strengths for Vision Transformers in DR detection:

- They're flexible, performing well across different datasets.
- They handle tricky, detailed multi-class DR grading.
- They're solid at spotting lesions that pop up in scattered patterns all over the retina.

This paper addresses these gaps through a structured comparative analysis of different approaches.

III. MODEL OVERVIEW

Here's how this diabetic retinopathy detection system works: You start by collecting retinal images with fundus cameras. Then, you clean up those images, pull out important features using Vision Transformers, and finally, classify and evaluate them.

A. Image Acquisition Module

Firstly, we get images with specialized fundus cameras. The process looks like this:

$$I(x,y) = R(x,y) \cdot L(x,y) + N(x,y)$$

That means:
- I(x,y) is the actual image you capture,
- R(x,y) is the retinal structure,
- L(x,y) is the lighting in the image,
- N(x,y) covers any noise.

B. Preprocessing Module

We need sharp, usable images, so preprocessing comes next. Basically, we resize the images, normalize them, boost contrast with CLAHE, and knock down noise using a Gaussian filter.

Gaussian Filter:

$$G(x,y)= (1/2\pi\sigma^2) e^{(-(x^2 +y^2)/(2\sigma^2))}$$

Filtered Image:

$$I \text{ filtered} = I * G$$

C. Patch Embedding (ViT Input)

Images are divided into fixed-size patches and flattened.

$$x_p = Flatten(Patch_i)$$

Embedding representation:

$$z0 = [x_{class}; x_{p1}\boldsymbol{E}; x_{p2}\boldsymbol{E}; \ldots; x_{pn}\boldsymbol{E}] + \boldsymbol{E}_{pos}$$

Where:

- E= embedding matrix
- E_{pos}= positional encoding

D. Self-Attention Mechanism

The self-attention mechanism models relationships between image patches.

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

This enables the model to capture **long-range contextual relationships** across retinal images.

E. Classification Module

The final classification layer uses a soft max function.

$$P(y = c \mid x) = \frac{e^{z_c}}{\sum_{k=1}^{K} e^{z_k}}$$

The model predicts the following DR stages:

- No DR
- Mild
- Moderate
- Severe
- Proliferative DR

F. Evaluation Metrics

Performance is evaluated using:

Accuracy

$$\frac{TP + TN}{TP + TN + FP + FN}$$

Precision

$$\frac{TP}{TP + FP}$$

Recall

$$\frac{TP}{TP + FN}$$

F1 Score

$$F1 = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall}$$

IV Experimental Setup

Experiments are conducted on standard datasets such as APTOS and Messidor. The dataset is split into training, validation, and testing sets (70:15:15).

Images are resized and normalized, with augmentation applied to handle class imbalance.

V. COMPARATIVE ANALYSIS

Model	Feature Learning	Global Context	Data Requirement	Accuracy
CNN	Local	Limited	Medium	85–92%
CNN + Attention	Improved	Moderate	Medium	90–94%
Vision Transformer	Global	Strong	High	92–97%

Key Findings

- CNN models perform well in lesion detection.
- Vision Transformers capture global spatial dependencies.
- Hybrid CNN–Transformer architectures show the best performance.

VI. CONCLUSION

This paper has discussed the developments in the field of automated detection systems for diabetic retinopathy, emphasizing the shift from conventional image processing techniques to the more robust deep learning models. Although the CNN-based models have shown promising baseline results, the Vision Transformer models have shown the potential for better global feature modeling using self-attention mechanisms.

Although the transformer-based models have shown promising results in the multi-class grading of DR, there are still some challenges to be overcome. The challenges include the data imbalance, the high computational complexity, and the lack of interpretability. The comparative values are based on existing literature and are used for analytical comparison, as this study is a review paper.

Further research needs to be done on the development of efficient transformer models, multimodal learning models, and explainable AI models. The fusion of retinal imaging techniques and Vision Transformer models is a significant step towards the development of efficient and intelligent DR detection systems that prevent avoidable blindness Vision Transformers provide better global feature learning than CNNs but require higher computational resources and larger datasets. Future work should focus on hybrid models and explainable AI to improve performance and interpretability

VII. REFERENCES

[1] A. Sopharak, B. Uyyanonvara, S. Barman, and T. H. Williamson, "Automatic detection of diabetic retinopathy exudates from non-dilated retinal images using mathematical morphology methods," Computers in Medicine and Biology, vol. 38, no. 8, pp. 843–852, 2008.

[2] M. Niemeijer, B. van Ginneken, J. Staal, M. Suttorp-Schulten, and M. Abràmoff, "Automatic detection of red lesions in digital color fundus photographs," IEEE Transactions on Medical Imaging, vol. 24, no. 5, pp. 584–592, 2005.

[3] V. Gulshan et al., "Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs," JAMA, vol. 316, no. 22, pp. 2402–2410, 2016.

[4] Kaggle, "APTOS 2019 Blindness Detection Dataset," 2019.

[5] E. Decencière et al., "Feedback on a publicly distributed database: The Messidor database," Image Analysis & Stereology, vol. 33, no. 3, pp. 231–234, 2014.

[6] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in Proc. IEEE CVPR, 2016, pp. 770– 778.

[7] A. Vaswani et al., "Attention is all you need," in Advances in Neural Information Processing Systems (NeurIPS), 2017.

[8] A. Dosovitskiy et al., "An image is worth 16x16 words: Transformers for image recognition at scale," in Proc. ICLR, 2021.

[9] Z. Liu et al., "Swin Transformer: Hierarchical vision transformer using shifted windows," in Proc. IEEE ICCV, 2021.

[10] Y. Li et al., "Vision transformer-based automated diabetic retinopathy grading," IEEE Access, vol. 10, pp. 112345–112356, 2022.

[11] K. V. Shanthala and N. C. Kundur, "RetinoFusionNet: A scalable and interpretable Vision Transformer framework for diabetic retinopathy detection," Eng., Technol. & Appl. Sci. Res., vol. 16, no. 1, pp. 31386–31392, Feb. 2026.

[12] Y. Tewari, N. S. Parihar, K. Rautela, N. Kaundal, M. Diwakar, and N. K. Pandey, "Diabetic retinopathy detection and analysis with convolutional neural networks and Vision Transformer," BISH, vol. 1, no. 1, pp. 18– 26, Jun. 2025.

[13] OcuViT: A Vision TransformerBased approach for automated diabetic retinopathy and AMD classification, J. Imaging Inform. Med., 2025, doi:10.1007/s10278-025-01676-3.

[14] K. Patni and S. Yagnik, "Cross-Dataset Unified Vision Transformer model for diabetic retinopathy detection," J. Comput. & Biomed. Inform., vol. 10, no. 1, pp. 1–14, Dec. 2025.

[15] A. Shukla, S. Tiwari, and A. Jain, "HybridFusionNet: Deep learning for multi-stage diabetic retinopathy detection," Technologies, vol. 12, no. 12, p. 256, Dec. 2024.

[16] W. Zhang, V. Belcheva, and T. Ermakova, "Interpretable deep learning for diabetic retinopathy: A comparative study of CNN, ViT, and hybrid architectures," Computers, vol. 14, no. 5, p. 187, May 2025.

[17] S. Akhtar et al., "A deep learning-based model for diabetic retinopathy grading," Sci. Rep., vol. 15, Art. no. 3763, Jan.