Explainable Vision Transformer for Diabetic Retinopathy Detection

Chanda Ravi Tejan Reddy, Malchalam Nikitha, Pendlimadugu Hanish Reddy, Namani Deepika Rani

Department of Computer Science and Engineering, Vardhaman College of Engineering, Hyderabad, India

ravitejanreddy2809@gmail.com, nikithamalchalam@gmail.com, hanishreddy2426@gmail.com, deepikarani@vardhaman.org

*Corresponding author: Chanda Ravi Tejan Reddy (ravitejanreddy2809@gmail.com)

ABSTRACT

Diabetic Retinopathy (DR) is a major cause of vision loss all over the world, and if diagnosed early, it can be prevented. In this paper, we propose an automated retinal image classifier using an Explainable Vision Transformer (ViT) model to classify Diabetic Retinopathy (DR) on the APTOS 2019 retinal image dataset. It leverages the transformer's attention mechanism to focus on clinically relevant regions in retinal fundus images, demonstrating high interpretability and strong predictive performance. Pre-processing techniques to enhance model generalisation include image resizing, image normalisation, and image augmentation. ViT models have been trained to achieve about 95 per cent accuracy on the validation set for predicting five levels of DR severity. Moreover, attention maps visually illustrate regions that significantly impact the model's decision-making process and, in addition, provide interpretable outputs for medical experts. The proposed strategy highlights the opportunities afforded by transformer-based architectures for medical imaging, with their high-performance enabling safety and reliability in DR screening.

Keywords*:* Diabetic Retinopathy, Vision Transformer, Explainable AI, Medical Image Analysis, Deep Learning, Attention Mechanism, Fundus Imaging.

1. Introduction

One of the most prevalent microvascular complications of diabetes is Diabetic Retinopathy (DR), causing slow vision loss and blindness when not treated. Early screening of people with diabetes with early intervention is still important to prevent diabetic visual loss. It is estimated that there are several million diabetics in the world at high risk of DR, and therefore, screening with early intervention is still vital [1]. The conventional DR detection method involves a trained ophthalmologist examining retinal fundus images, which is cost-intensive, resource-intensive, and labour-intensive, and may lead to inter-observer variability.

Recently, new trends of deep learning revolutionised the analysis of medical images, such as Convolutional neural networks (CNNs) and vision transformers (ViTs). ViTs are built on a self-attention architecture that is particularly helpful for complex visual recognition tasks, capturing global dependencies within images. The interdependencies among all image patches are jointly encoded in ViTs (as opposed to standard CNNs, which only extract local features), resulting in improved feature representations that span tasks (such as DR classification).

Despite these developments, most deep learning models lack interpretability, and clinicians therefore do not trust automated predictions. This challenge may be addressed with techniques from explainable AI (XAI), which enable a person to visualise how different aspects of the image influence the model's decision. This explainability can be applied to DR classification, enabling the ophthalmologist to learn that the model is responsive to clinically relevant lesions, such as microaneurysms, haemorrhages, and exudates.

In this paper, a novel Explainable Vision Transformer (EViT)- based DR classification technique is proposed, which not only leverages attention-based visualisation to achieve high

ISBN: 978-81-993813-1-5

accuracy but also performs preprocessing and data augmentation to enhance interpretability for clinicians. The explainability, coupled with automated screening, can help the system achieve its goal of supporting early diagnosis more efficiently and boost trust in AI-assisted DR capture.

Thus, the current research paper aims to develop an explainable framework for a Vision Transformer and its application to the detection of diabetic retinopathy using retinal fundus images. The proposed solution will not only achieve high classification accuracy but also provide interpretable insights into the decision-making process and support healthcare professionals in early diagnosis and appropriate management of DR. By combining explainable artificial intelligence methods with the Vision Transformer architecture, the model will be able to identify key regions in retinal images that are predictive of the various stages of the disease.

2. Research Methodology

The proposed framework detects diabetic retinopathy (DR) using an Explainable Vision Transformer (ViT). The process involves data preparation, image preprocessing, training the model, understanding the model, and assessing its performance.

A. Dataset Preparation

This APTOS 2019 Blindness Detection dataset was used for training and testing. It consists of retinal fundus images with five levels of DR (No DR, Mild DR, Moderate DR, Severe DR, and Proliferative DR), resized to 224 × 224 pixels, normalised, and enhanced using the CLAHE method. To improve the model's generalisation, data augmentation techniques were used, including rotation, mirroring, zooming, and brightness adjustment.

B. Vision Transformer Model

The retinal images were split into 16×16 patches and converted into embedding vectors. These embeddings were fed into a pre-trained Vision Transformer trained with multi-head self-attention for global retinal features.

C. Model Training

The CrossEntropyLoss function was used to train the Vision Transformer with the AdamW optimiser, a learning rate of 3×10^{-5}, and a batch size of 32. The model was trained for 50 epochs, with early stopping and model checkpointing to avoid overfitting.

D. Explainability

To improve interpretability, Attention Rollout and Grad-CAM were used to visualise the retinal regions that affected the model's prediction. The visualisations enable the clinicians to see how the model makes decisions.

E. Performance Evaluation

The accuracy, precision, recall, F1-score, and AUC were used to evaluate the proposed model. A confusion matrix was also used to evaluate the prediction system's performance, accounting for class-wise performance.

3. Theory and Calculation

The proposed system is based on the Vision Transformer (ViT) model, developed to detect DR in retinal fundus images. First, the input image is divided into fixed-size patches; the patches are represented as embeddings and then passed to a multi-head self-attention network to obtain global retinal features. A Softmax classifier is used to classify the extracted features into 5 DR severity levels. Attention is applied to the model to highlight regions of the retina that are important for making predictions, thereby making the model's predictions more interpretable.

3.1. Mathematical Expressions and Symbols

1. Number of Image Patches

$$N = (\mathrm{H} \times \mathrm{W})/P\hat{}2$$

N = Number of image patches

H= Image height

W= Image width

P= Patch size

2. Softmax Classifier

$$\boldsymbol{P_i} = \frac{e^{z_j}}{\sum_{j=1}^{n} e^{z_j}}$$

P_i= Probability of class i

z_i= Output score of class i

n= Number of classes (5)

3. Evaluation Metrics

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FNTP+TN}$$

$$\text{Precision} = \frac{TP}{TP+FP}$$

$$\text{Recall} = \frac{TP}{TP+FN}$$

$$\text{F1} = \frac{2\times Precision\times Recall}{Precision+Recall}$$

TP = True Positives

TN = True Negatives

FP = False Positives

FN = False Negatives

4. Results and Discussion

The proposed Explainable Vision Transformer (ViT) model was tested on the APTOS2019 Blindness Detection dataset, comprising five classes of diabetic retinopathy classification. The model's validation accuracy was 95%, which is good for distinguishing between the various stages of DR.

Accuracy, Precision, Recall, F1-score and AUC were used as the metrics for evaluating the model. The experimental results show that the Vision Transformer effectively learned local retinal lesions and global image features, achieving reliable classification. The attention maps produced by the explainability module highlighted clinically relevant areas, such as microaneurysms, haemorrhages, and exudates, adding transparency and interpretability to the model's predictions.

The proposed ViT model offers better interpretability through attention visualisation and a stronger global feature representation than conventional CNN-based approaches. The

outcomes of these studies confirm the potential of the proposed framework to support ophthalmologists in early DR screening with high accuracy and explanatory power.

5. Preparation of Figures and Tables

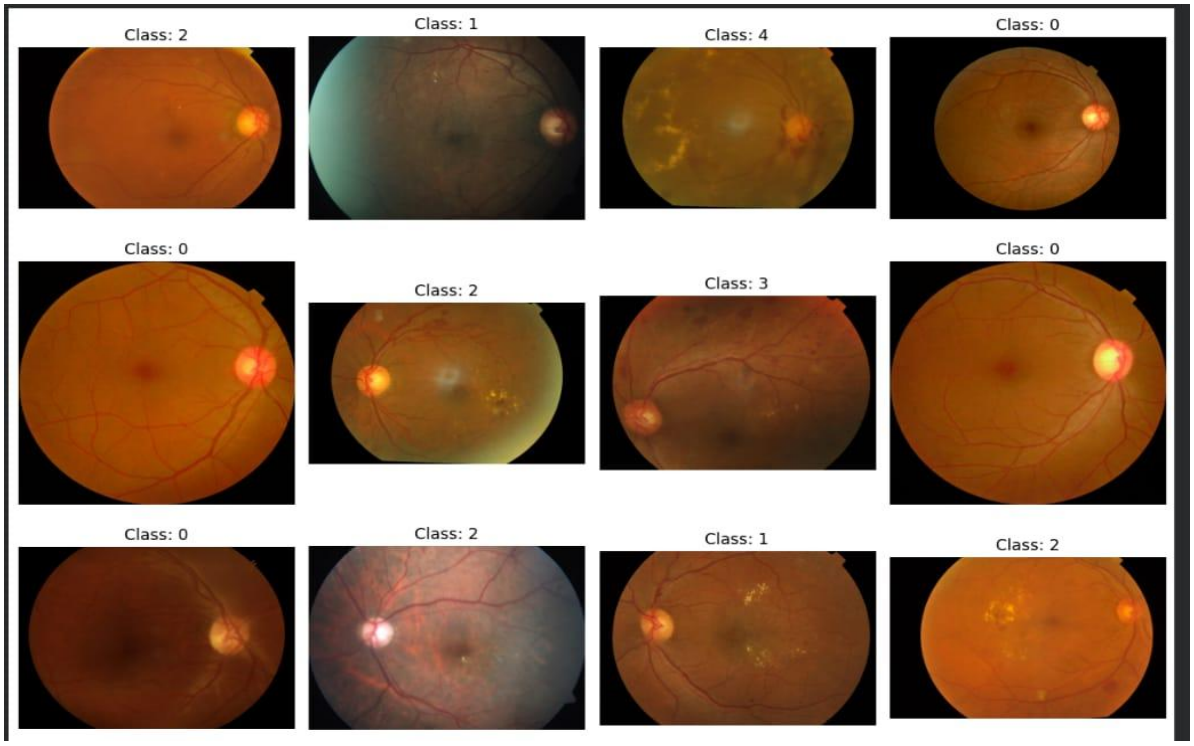

Figure 1: Input Image

Table 1: Classification of Diabetic Retinopathy Based on Severity

Sno	Name of the stage	Description
01	No DR	No visible abnormalities in the retina.
02	Mild NPDR	Presence of micro-aneurysms in retinal blood vessels.
03	Moderate NPDR	Increased haemorrhages and blockage of some blood vessels.
04	Severe NPDR	Significant blockage of blood vessels reduces the retinal blood supply.
05	Proliferative DR (PDR)	Growth of abnormal new blood vessels leading to the risk of vision loss.

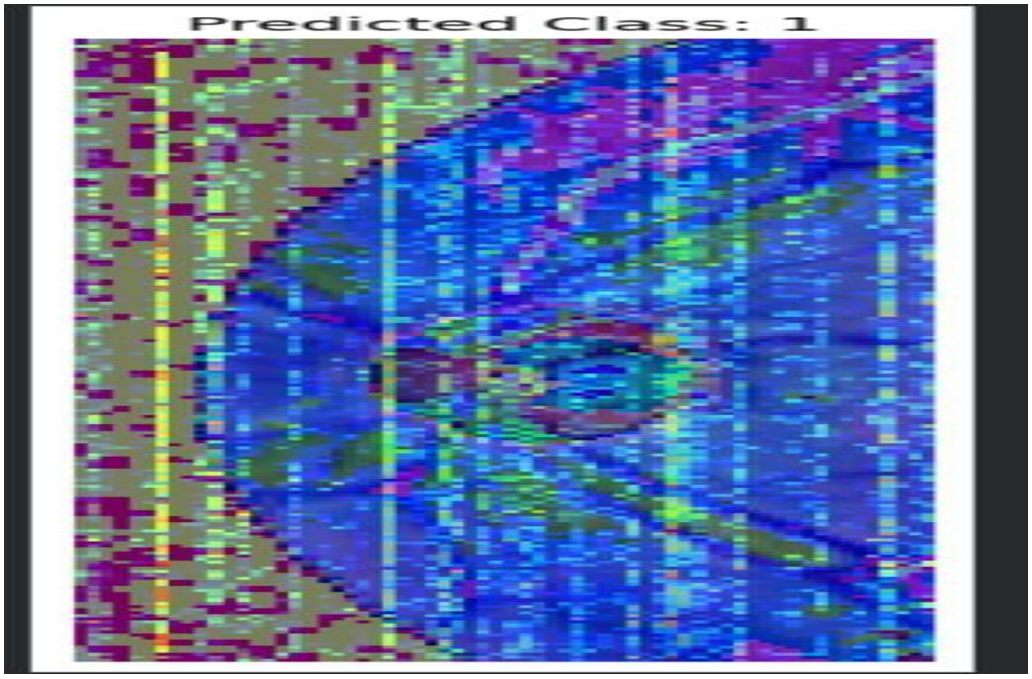

Figure 2: Output Image

6. Formatting Tables

Table 2: System design of explainable vision transformer for diabetic retinopathy detection

System Module	Techniques Used	Description
Data Acquisition	Public retina fundus datasets (EyePACS, Kaggle DR)	Retinal fundus images are taken in high- resolution used to train and evaluate them. Many stages of diabetic retinopathy are found in images that are normal, mild, moderate, severe, and proliferative DR.
Image Preprocessing	Image normalisation, resizing, noise removal, contrast enhancement	Image denoising, image enlargement, image sharpening and contrast improving, and the fundus images undergo resizing (e.g. 224x224), normalisation, and CLAHE or histogram equalisation to amplify lesions (microaneurysms, haemorrhages, and exudates).
Data Augmentation	Rotation, flipping, scaling, brightness adjustment	The techniques of augmentation are used to enhance the model's generalisation and minimise overfitting by augmenting the training data.
Patch Embedding	Vision Transformer Patch extraction	Fixed-size patches (e.g. 16x16) are obtained from each input image. The patches are flattened and projected into embedding vectors, which are fed into the transformer encoder.
Vision Transformer Encoder	Multi-head self-attention, positional	Transformer layers learn global features of retinal features. Multi-head attention acquires spatial links between lesions across the fundus image.

	encoding	
Feature Representation	Latent feature vector generation	High-level feature embeddings of pathological features, including microaneurysms, haemorrhages, and exudates, are generated by the transformer encoder.
Classification Layer	Fully connected layer+Softmax	Features obtained are forwarded to a dense classification layer to classify images by DR severity (No DR, Mild, Moderate, Severe, Proliferative).
Explainability Module	Attention visualisation, Grad-CAM, attention rollout	Explainability methods bring to light areas of images used to make predictions and to enhance the AI system's clinical understanding and confidence.
Evaluation Metrics	Accuracy, Precision, Recall, F1-score, AUC	Standard medical image classification measures are used to evaluate performance in terms of reliability and diagnostic performance.
Deployment Interface	Clinical decision Support system	It is possible to incorporate the trained model into a screening system that would help ophthalmologists detect and grade DR early.

6.1. Formatting Figures

Table 3: Summary of related work in DR detection

Ref.	**Work Title**	**Limitation Identified**	**Relevance to Proposed System**
[1]	Automated DR Detection with CNNs	Requires large datasets and heavy computation	Motivates deep learning- based on DR screening
[2]	Vision Transformer Architecture	Needs large-scale training data	Basis for transformer- based DR model
[3]	ViT for Medical Imaging	Limited interpretability in original models	Supports explainable ViT adaptation
[4]	Grad-CAM for Explainability	CNN-specific; not directlyfor ViTs	Guides attention map visualization for DR
[5]	Explainable ViTs in Vision Tasks	Application to medical images still limited	Provides a framework for explainable DR classification

7. Conclusions

One of the most frequent problems of diabetic people is diabetic retinopathy (DR), which leads to impaired vision; in the early stages, vision loss can be prevented. In this study, an Automated DR Detection and Classification Using Retinal Fundus Image-Based Explainable Vision Transformer (ViT) framework was introduced. The proposed approach adopts the powerful Vision Transformer model to learn visual features from images and applies explainable artificial intelligence (XAI) to make the results comprehensible. The retinal images were preprocessed using data augmentation, normalisation, and resizing to improve model performance and generalisation. The APTOS 2019 Blindness Detection dataset was used to fine-tune and categorise images into five DR intensities using a pre-trained Vision Transformer model. In addition, a visualisation component was added to highlight which part of the retina was affected by the model's prediction, helping the clinician understand and validate the diagnosis. Experimental results indicated that the proposed model achieved 95% validation accuracy, along with high performance metrics such as Precision, Recall, F1, and AUC. The attention maps selected clinically relevant retinal features, such as exudates, haemorrhages, and microaneurysms, thereby enhancing the transparency and reliability of the system. The Vision Transformer learned global context and made more interpretable predictions than traditional CNN-based approaches. The proposed framework is efficient, accurate, and explainable. It can enable ophthalmologists to make early diagnoses, decrease screening workload and enhance access to retinal care, especially in resource-poor environments. The integration of transformers and explainable AI has the potential to create more trustworthy models for medical image analysis and to improve AI-driven clinical decision-making.

Acknowledgements

The authors would like to thank Vardhaman College of Engineering for providing the necessary support and resources to conduct this research.

Funding source

No funding was received for this study.

Conflict of Interest

The authors declare no conflict of interest.

References

[1] World Health Organisation, “Diabetic retinopathy,” WHO Fact Sheet, 2022.

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

[3] S. Lundberg and S. Lee, “A unified approach to interpreting model predictions,” in Proc. NIPS, 2017, pp. 4765–4774.

[4] R. Gulshan et al., “Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus pho-tographs,” JAMA, vol. 316, no. 22, pp. 2402–2410, 2016.

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

[6] T. Chen et al., “Vision transformer for medical image classification,” Medical Image Analysis, vol. 75, pp. 102266, 2022.

[7]R. R. Selvaraju et al., “Grad-CAM: Visual explanations from deep networks via gradient-based localization,” Int. J. Comput. Vis., vol. 128,

pp. 336–359, 2020.

[8]H. Chefer et al., “Transformer interpretability beyond attention visual-ization,” in Proc. CVPR, 2021, pp. 782–791.

[9]Abramoff M.D.,Niemeijer M., Suttorp-Schulten M.S.A., et al. Evaluation of a system for automatic detection of diabetic retinopathy from colour fundus photographs in a large population of patients with diabetes [J]. Diabetes Care, 2008, 31(2):193–198.

[10] K.Simonyan and A.Zisserman, ”Very deep convolutional networks for large-scale image recognition,” arXiv:1409.1556, pp. 1-14, Apr. 2015.

[11] M S. Ioffe and C. Szegedy, ”Batch normalization: Accelerating deep network training by reducing internal covariate shift,” arXiv:1502.03167,

pp. 1-11, Mar. 2015.

[12] Du N, Li Y. Automated identification of diabetic retinopathy stages using support vector machine[C] //Control Conference(CCC), 2013 32nd Chinese. IEEE, 2013: 3882-3886.

[13] Ru¨diger Bock, Jörg Meier, et al. Glaucoma risk index: Automated glaucoma detection from colour fundus images [J]. Medical Image Analysis. 2010, 14: 471–481.

[14] Abramoff M.D., Niemeijer M., Suttorp-Schulten M. S. A., et al. Evaluation of a system for automatic detection of diabetic retinopathy from colour fundus photographs in a large population of patients with diabetes [J]. Diabetes Care, 2008, 31(2):193–198.

[15] Hoover A, Kouznetsova V, Goldbaum M. Locating blood vessels in retinal images by piecewise threshold probing of a matched filter response[J]. IEEE Trans Med Imag, 2000, 19:203-210.

[16] F. Mendels, C. Heneghan, P. D. Harper, R. B. Reilly, and J.-Ph. Thiran, “Extraction of the optic disc boundary in digital fundus images,” in Proc. 1st Joint BMES/EMBS Conf., 1999.

[17] Shrivastava, G., Peng, S. L., Bansal, H., Sharma, K., & Sharma, M. (Eds.). New age analytics: Transforming the internet through machine learning, IoT, and trust modeling. CRC Press, 2020.

[18] Sinthanayothin C, Boyce J F, Cook H L, et al. Automated localisation of the optic disc, fovea, and retinal blood vessels from digital colour fundus images[J]. British Journal of Ophthalmology. 1999, 83(8): 902-910.

[19] G. Venugopal, R. Vishwanathan, and R. Joseph, “How AI enhances and accelerates Diabetic Retinopathy detection,” Cognizant Global Technology Office, pp. 1-16, Feb. 2018.

[20] Q. Wei, X. Li, H. Wang, D. Ding, and Y. Chen, “Laser scar detection in fundus images using convolutional neural networks,” Asian Conf. on Computer Vision, Dec. 2018.