IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 4, April 2026

DOI: 10.17148/IJARCCE.2026.15478

Diabetic Retinopathy Classification Using Vision Transformer: A Strategy for Small Dataset Challenges

Mr K. Appala Raju[1], M. Pravallika[2], M. Bhumika[3], M.V.K.S. Harika[4]

Professor, Dept. of ECE, Andhra Loyola Institute of Engineering and Technology, Vijayawada, AP, India[1]

Dept. of ECE, Andhra Loyola Institute of Engineering and Technology, Vijayawada, AP, India [2-4]

Abstract: Early Detection of diabetic retinopathy, a complication of Vision loss in advance stages of diabetes, is essential to avoid permanent vision impairment. However, the automatic detection of diabetic retinopathy through medical image processing requires a large number of training data to build a model with good performance. This poses a challenge when working with small datasets as these models need large datasets to perform well on unseen data. Conventional Neural networks(CNNs), often fall short in capturing long-range dependencies, global pathological features across high resolution retinal images, leading to suboptimal performance in early-stage diagnosis. To address these limitations, this study proposes a Vision Transformer (ViT) model, designed to elevate DR severity classification (ranging from NO DR to Severe DR) by leveraging the self-attention mechanisms of transformer architectures. Vision Transformer(ViT) is a Deep learning architecture that generally requires large datasets for effective training. However, in this work, a smaller dataset is used because large medical datasets are difficult to access due to privacy and datasharing restrictions. The proposed approach utilizes a hierarchical structure where retinal fundus images from public (Kaggle) dataset APTOS 2109 dataset and a private (FGADR Website) dataset FGADR are divided into non-overlapping patches, embedded, and enriched with positional information. The proposal model achieves accuracy rates on threeclassification of 90% on FGADR and 86% on APTOS 2019 dataset. The model exhibits high performance, achieving a quadratic Weighted kappa (QWK) score of 0.93 on FGADR and 0.86 on APTOS. The proposed model demonstrates the good results to perform multi-class classification of DR using limited number of images.

Keywords: Diabetic Retinopathy (DR), Blindness/ Vision Loss Detection, Disease Severity Classification, Convolutional Neural Networks (CNNs), Vision Transformer (ViT).

I. INTRODUCTION

Diabetic retinopathy is a retinal vascular disease that develops in the advanced stages of diabetes. In diabetes, patients suffer from high blood sugar levels, which can cause severe damage to different organs, including the retina. High blood sugar levels damages the tiny vessels in the retina, leading to blood and small fluid leaks into the eye, resulting in diabetic retinopathy. Initially, it starts with small blood leaks, blocked and swelling blood vessels, fatty deposits in the retina and progressing to more severe changes. These symptoms lead to blurry vision, floating sots in the vision and eventually permanent vision impairment. It is hard to notice the small changes in vision in the early stages. Previously, using CNN (Convolutional neural networks) for finding the retinopathy classification. In this research, using Vision Transformer (ViT) to classify accurate Diabetic retinopathy classification by using two datasets APTOS 2019 & FGADR.

The first stage of Vision Transformer architecture is the input image and patch embedding stage. In this step, the input image is divided into small fixed-size patches, for example 16 x 16 pixels. Each patch is flattened into a vector and then transformed into an embedding using a linear projection layer. These embeddings represent the visual information of each image patch. Since the transformer model does not inherently understand spatial positions, positional encoding is added to each patch + embedding so that the model can recognize the location of each patch within the image. After patch embedding, the next stage is the transformer encoder block. This block contains multiple layers that include multi-head selfattention and feed-forward and neural networks. In the self-attention mechanism, each image patch interacts with all other patches in the image. This allows the model to capture both local and global relationships within the image.

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 4, April 2026

DOI: 10.17148/IJARCCE.2026.15478

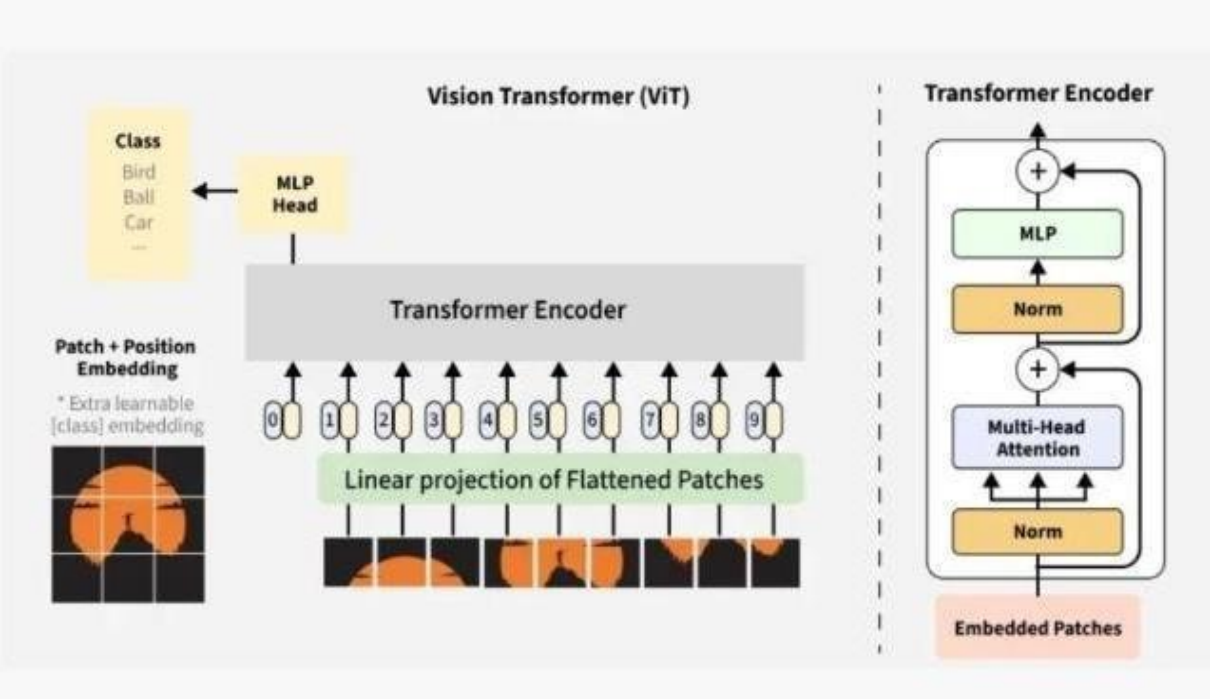

Fig 1: Vision Transformer Architecture

For example, in retinal images, the model can analyze relationships between vessels, lesions, and other retinal structures across different regions.The vision Transformer helps improve the accuracy and reliablility of automated diagnostic system. This makes it a powerful tool for supporting doctors in early disease detection and improving patient care.

II. LITERATURE SURVEY

Diabetic retinopathy is one of the most common complications of diabetese and a major cause of blindness worldwide. Early detection and timely treatment are essential to prevent severe vision loss. Traditionally, opthmologists detect diabetic retinopathy by manually examining retinal fundus images. However, manual screeinng is time consuming and requires expert knowledge. With the advancement of artificial intelligence and medical imaging technologies, automated computer-aided diagnosis system have been developed to improve the accuracy and effeciency of retinal disease detection. Deep learning models, pariticularly convolution-based architectures, have been widely used for automated diabetic retinopathy detection. Several CNN-based systems were developed using large public datasets such as EyePACS, Messidor, and APTOS. These models achieved high classification accuracy for identifying different stages of diabetic retinopathy. However, CNN architectures mainly rely on convolutional filters that focus on local image features, which limits their ability to capture global contextual relationships across the retina. Researchers found that subtle lesions distributed across different regions of the retina could be missed due to this limitation. To address these limitations, researchers have recently introduced transformer-based architecture, especially the vision Transformer, which uses self-attention mechanism to mode; global relationships between image regions. This has significantly improved the performance of automated diabetic retinopathy screening systems. The Vision Transformer (ViT) was introduced as a powerful alternative to convolution-based architectures for computer vision tasks. Unlike CNNs, which process images using convolutional filters, Vision Transformers divide images into small patches and process them as sequences using transformer encoders. The self-attention mechanism allows the model to learn relationships between different regions of the image simultaneously.

III. METHODOLOGY

Methodology involves following stages:

A. Data Collection:

Collecting various Retinal fundus images from the datasets. Those datasets are, one is publicly available dataset APTOS 2109 Blindness detection from kaggle and another dataset is private dataset. Collect it from FGADR website or mail to yizhou.szcn@gmail.com for the dataset. The FGADR dataset contains 1852 images and APTOS contains 3,662 images. In this research,use 100 images per class for three clasification from both datasets.

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 4, April 2026

DOI: 10.17148/IJARCCE.2026.15478

Table 1: Summary of Experimental Datasets

Dataset	source	No.of images	Annotation type
FGADR	FGADR website	1852	Pixel level Lesions+ Grades
APTOS	Kaggle	3662	5-class severity grade

Fig 2:Retinal fundus images

B. Image Preprocessing:

Raw fundus images undergo a multi-stage preprocessing pipeline to enhance diagnostic features:

Resizing: All images are resized to 224 x 224 pixels to match the ResNet-50 input dimensions.

Normalization: Pixel values are normalized to [0,1] using ImageNet mean and standard deviation for compatibility with pretrained ResNet weights.

C. Dataset splitting:

Split the data for training and validation. Based on the program it split the images for training and validation. There is some calculation behind the program. Take 100 images per class. 300 images from each dataset for three classifications.

Table 2: Dataset Overview

Dataset	No_DR	Mild/Moderate_DR	Severe_DR
FGADR	100	100	100
APTOS	100	100	100

D.Vision Transformer (ViT) Model:

Feature Extractions from images:

By using Vision Transformer Architecture capture the global relationships of an image. It observes each part of an image and find out the small lesions easily. Feature extraction means identifying and extracting important information (patterns) from data so that a model can understand and make predictions. Shapes, Edges, Colors, patterns are called features. From a retinal image feature extraction finds: Microaneurysms (tiny red dots), Hemorrhages (bold spots), Exudates (yellow patches), Blood vessel patterns. These features help the model whether the eye is No_DR, Mild/Moderate_DR, severe_DR.

E.Model Training:

The deep learning model is trained using the labeled retinal images. During training, the network learns to associate visual patterns with diabetic retinopathy severity levels.

Important training parameters include:

- Batch size – number of images processed at once
- Epochs – number of complete training cycles
- Optimizer – algorithms such as Adam or SGD

The training process minimizes a loss function, usually categorical cross-entropy, to improve prediction accuracy over time.

F.Model Evaluation:

After Training, the model is evaluated using validation and test datasets. This step measures how well the system can detect diabetic retinopathy stages.

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 4, April 2026

DOI: 10.17148/IJARCCE.2026.15478

Common evaluation metrics include:

- Accuracy – percentage of correct predictions
- Precision – reliability of positive predictions
- Recall (sensitivity) – ability to detect disease cases
- F1 Score – balanced between to precision and recall
- Quadratic weighted kappa (QWK)

These metrics help determine whether the model is suitable for real-world medical applications.

G.DR classification:

Classification is the final process. After Training and validation and testing, the vision transformer model gives the prediction. It can classify the image is No_DR, Mild/Moderate_DR, Severe_DR.

IV. EXPERMENTAL RESULT

For this Experiment, Google colab notebook is used. Type the program in it. Here is the simple explanation of the program. Initially create the dataset or download the dataset and upload it in the drive. Create a Main folder and move 100 images per class in the different files in the main folder and upload it into Google drive. Connect Google drive to your notebook and install the libraries required. select the data path where the folder is in the drive.

Next preprocessing the images and use the vision transformer for finding the features. Train the images using 10 epochs for FGADR and 30 epochs for APTOS. And observe the confusion matrix, accuracy graph, and DR classification.

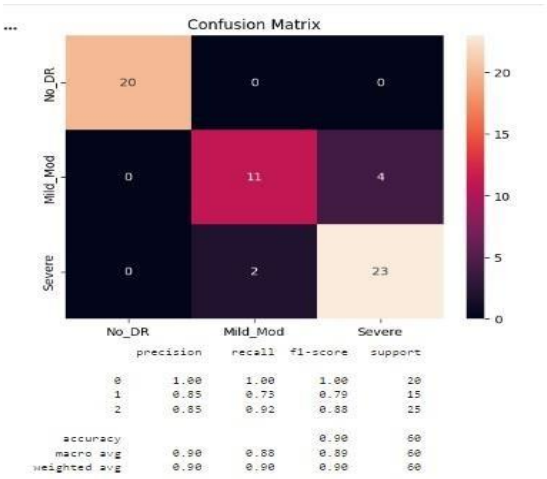

Fig 3: Confusion Matrix of FGADR

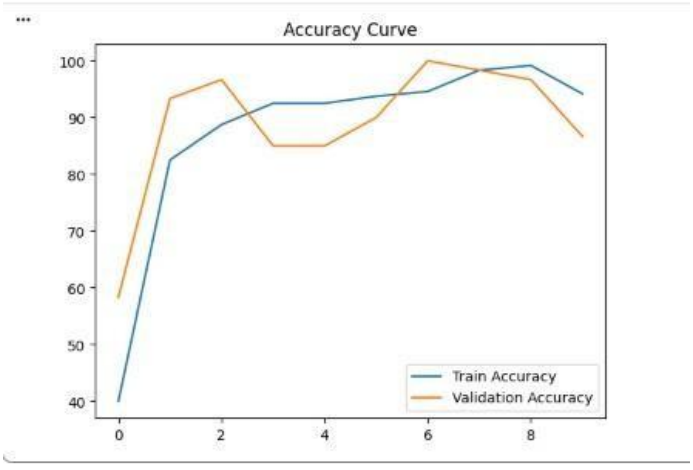

Fig 4: Accuracy Curve of FGADR

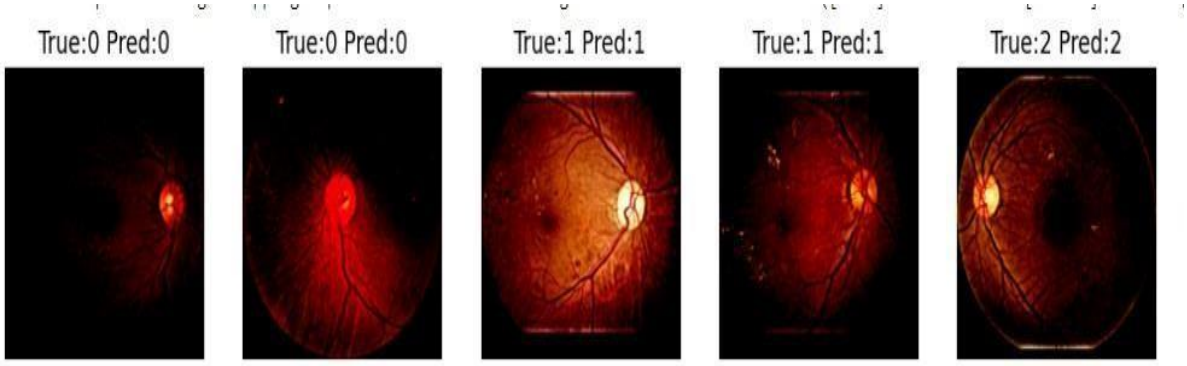

Fig 5: Classification of FGADR

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 4, April 2026

DOI: 10.17148/IJARCCE.2026.15478

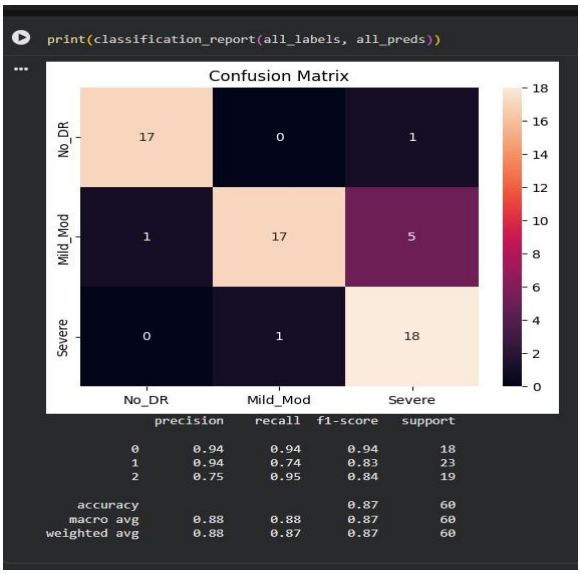

Fig 6: Confusion Matrix of APTOS

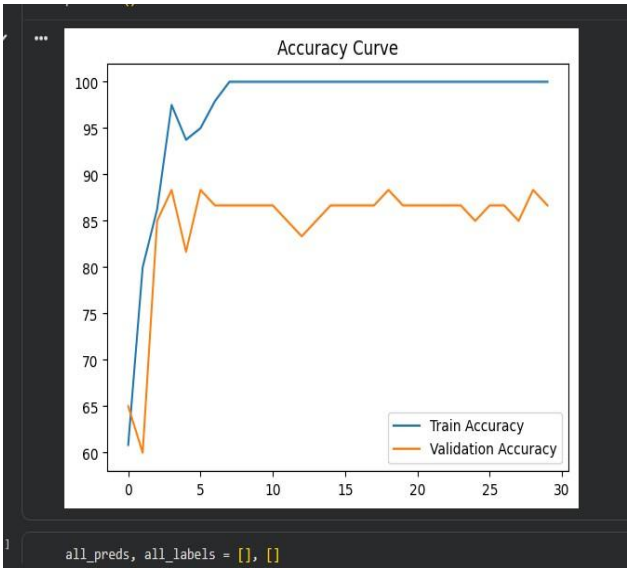

Fig 7: Accuracy curve of APTOS

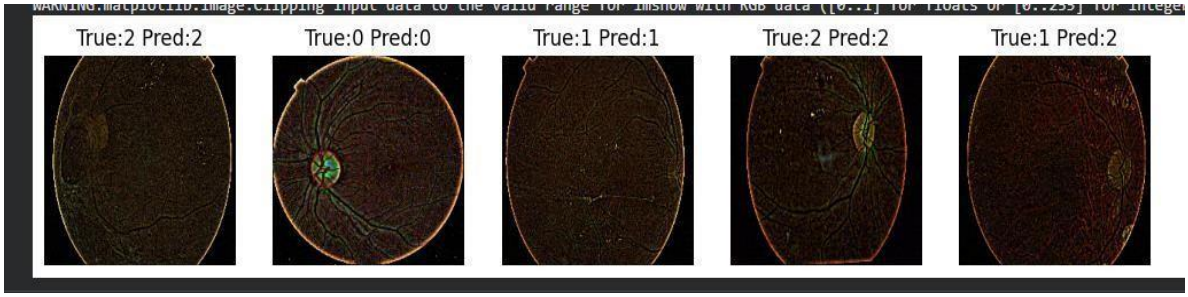

Fig 8: Classification Of APTOS

The classification figure shows True value and prediction value. True value is the original image from the raw data and the prediction value is generated by the vision Transformer / Analyzed by the vision Transformer. If the true value and predicted value is same means the vision Transformer works accurately. The Confusion matrix figure explains that how many images it can be trained accurately and how many can't be trained. For example, in the APTOS confusion matrix observe row No_DR. It has a value 17 in first column and 0 in 2nd column and 1 in 3rd column. Those values represent as, 17 images trained as No_DR and 0 images trained as mild/moderate_DR and 1 image is trained as severe_DR. In the confusion matrix there are precision, recall, f1-score, support. Precision defines how many predicted values are actually correct. Recall defines how many actual values are correctly detected. f1score is the balance between precision and recall. Support is the total number of actual samples for each class. https://github.com/2005-maddali/DR_Detection_Vision_Transformer/tree/main - code link.

V. CONCLUSION

Vision Transformer models can learn more comprehensive representations of retinal diseases, allowing them to perform multiple tasks such as classification, detection, and disease severity assessment. This integration supports the development of accurate computer-aided diagnosis systems capable of assisting ophthalmologists in clinical decision-making. Vision Transformer applying self-attention mechanism, enabling it to capture both local features and global relationships across the retina. It helps in early disease detection, large-scale screening, and prevention of vision loss.

REFERENCES

[1]. J. Wu, R. Hu, Z. Xiao, J. Chen, and J. Liu, "Vision Transformer-Based Recognition of Diabetic Retinopathy Grade," Medical Physics, 2021.

[2]. E. Cutur and N. Inan,"Multi-Class Classification of Retinal Diseases Using Transfer Learning Vision Transformers," Journal of Imaging Informatics in Medicine, 2025.

DOI: 10.17148/IJARCCE.2026.15478

[3]. Y. Tewari, N. Parihar, K. Rautela, et al.,"Diabetic Retinopathy Detection and Analysis with CNN and Vision Transformer,"2025.

[4]. K. V. Shanthala and N. C. Kundur, "RetinoFusionNet: Vision Transformer Framework for Diabetic Retinopathy Detection,"2026

[5]. M. Khater, "Diabetic Retinopathy Detection Model Using Hybrid of U-Net and Vision Transformer Algorithms," INTI Journal, 2024.

Authors

Mr K.Appala Raju, professor, Department of Electronics and Communication Engineering, Andhra Loyola Institute Of Engineering And Technology, Vijayawada, AP.

M.V.K.S.Harika, Student, Department of Electronics and Communication Engineering, Andhra Loyola Institute Of Engineering And Technology, Vijayawada, AP.

M.Pravallika, Student, Department of Electronics and Communication Engineering, Andhra Loyola Institute Of Engineering And Technology, Vijayawada, AP.

M.Bhumika, Student, Department of Electronics and Communication Engineering, Andhra Loyola Institute Of Engineering And Technology, Vijayawada, AP.

