JCBI
Journal of Computing & Biomedical Informatics
ISSN: 2710 - 1606
Volume 10 Issue 01
2025

Research Article
https://doi.org/10.56979/1001/2025/1148

Cross-Dataset Unified Vision Transformer Model for Diabetic Retinopathy Detection

Kinjal Patni[1*], and Shruti Yagnik[1]

[1]Department of Computer Engineering, Indus Institute of Technology and Engineering, Indus University, Ahmedabad, Gujarat, India.
[*]Corresponding Author: Kinjal Patni. Email: kinjalpatni11@gmail.com

Received: September 03, 2025 **Accepted:** November 26, 2025

Abstract: Diabetic retinopathy (DR) is a major cause of blindness around the world that can be prevented; it is caused by prolonged hyperglycemia that leads to damage to retinal vasculature. Early detection of DR, and good DR grading are important to ensure timely clinical intervention and improved outcomes for patients. Machine learning and convolutional neural network (CNN)-based approaches have held promise for DR screening, but they are often limited in their ability to capture fine-grained lesions or long-range dependencies because of limited receptive fields and insufficient modeling of global context.Vision Transformers, or ViTs, lever self-attention mechanisms to capture global relationships across retinal structures. This is a Review of ViT based frameworks on grading DR. The review concentrates on studies using two of the most widely applied and diverse benchmarks between EyePACS and APTOS; expert annotated fundus images assessment. The review covers a variety of recent advances such as hybrid CNN-ViT architectures, lesion-aware transformer modules, multi-scale feature aggregation, and federated learning strategies for privacy-preserving medical image analysis. It then highlights the role of interpretable attention maps in improving clinical trust and decision transparency. This is also the review of the remaining challenges in DR grading; it includes extreme class imbalance in DR severity levels, high computational costs of transformer models, and the demand for a powerful and robust explainability technique to favor clinical adoption. In connecting current achievements, unsolved issues, and new directions of research, this review endeavors to orient researchers and practitioners towards designing efficient, generalizable, and clinically relevant ViT-based DR detection and grading systems. In general, it shows how much transformer-driven approaches have the potential to revolutionize automated ophthalmic diagnosis and enhance the global diabetic eye care workflow.

Keywords: Diabetic Retinopathy; Vision Transformers; EyePACS; APTOS

1. Introduction

Diabetic Retinopathy (DR) is a severe microvascular complication of diabetes and a major cause of preventable adult blindness. There are more than 103 million cases of DR across the world, a number projected to rise to 160 million by 2045 on account of increased incidence of diabetes [1]. Studies show that about one out of three diabetic individuals develop some form of DR, with one in ten patients progressing to vision-threatening stages [2]. Among nearly 77 million diabetics, India has a prevalence of DR from 18%-25% [3]. Along with disturbing figures like these, less than 50% of cases go undetected in the low- and middle-income countries [4]. The potential for early diagnosis and management to reduce the risk of vision loss by as much as 95%, therefore, points to a need for standardized screening and classification methods with high sensitivity [5].

DR has five clinical stages based on the severity of retinal damage: No DR, Mild Non-Proliferative DR (NPDR), Moderate NPDR, Severe NPDR, and Proliferative DR (PDR). Microaneurysms appear as small

red dots, while hemorrhages appear as larger spots. Soft exudates present as white spots in a circular to oval form and hard exudates present as bright yellow spots. In end stages, neovascularization/retinal detachment/macular edema can impair vision severely. Proper grading of these stages is essential in choosing treatment modalities to avoid further loss of vision.

The new trend has shifted the automated grading of diabetic retinopathy from traditional image processing into the realm of deep learning, particularly by building CNNs. Though the CNNs excel in performance, these models have the inherent shortcomings in long-range dependency learning, built-in interpretability required for precise lesion detection and the usefulness of clinical application

In this paper, we investigated Vision Transformers (ViTs) as a possible alternative, based on a self-attention mechanism targeted for acquiring the global context of the image. Such a pursuit gives birth to a common model that is applied to the EyePACs and APTOS datasets-thus contributing to model robustness and generalization. The test results demonstrate that ViTs win over CNNs in terms of classification accuracy, cross-dataset performance, and interpretability, which makes it an ideal candidate for clinical decision support systems in screening DR.

2. Related Work

The challenges encountered during processing and analysing high-resolution retinal images have necessitated massive research efforts directed towards finding appropriate solutions; among these challenges include variation in image size, complexity of feature extraction, selection of the model, and fine-tuning of parameters. Accordingly, a plethora of options have been explored by researchers-from image preprocessing methods through to traditional and deep feature extraction techniques, advanced classification methods, and optimization frameworks-to enhance the accuracy of DR detection and classification.

2.1. ML Approaches for DR Detection & Classification

Table 1: The literature shows a significant development in the machine-learning (ML) detection of diabetic retinopathy (DR) based on different datasets like EyePACS, APTOS, and real hospital records-both public and private. The best accuracy (line 95.2%) was achieved by [5], which stated the use of ensemble ML techniques to predict both DR and nephropathy, suggesting that multi-complication models may be successful to a large extent. Likewise, [1] and [2] obtained their predictions with an accuracy of 95.1% and 94.8% respectively, with [1] proposing ensemble learning and [2] using entropy-based features to emphasize image analysis. An evaluation was made of traditional classifiers which included SVM, Decision Trees, and Naïve Bayes in [4], with the highest level of accuracy of 93.5%, while [3] validated models of different types of ML SVM, RF, and GBM across EyePACS and Kaggle datasets. Finally, [6] emphasized model interpretability with the combination of XGBoost with SHAP for an accuracy of 92.7% while informing the key risk factors, which is quite a reflection that it is becoming critical in medical applications.

Table 1. Survey of ML Methods in Prior Studies

Reference Paper	Datasets	Algorithms Used	Accuracy Parameter	Summary
[1]	EyePACS	Ensemble ML Model	95.1%	Proposed a robust ML model combining feature selection and classification techniques.
[2]	APTOS	ML with Entropy-based Features	94.8%	Utilized entropy-based image features to improve classification accuracy.
[3]	EyePACS, Kaggle	ML Models (SVM, RF, GBM, etc.)	94%	Comparative study and validation of different ML models for DR prediction.
[4]	Kaggle	SVM, Decision Tree, Naïve Bayes	93.5%	Evaluated classical supervised algorithms on

				fundus image data.
[5]	Hospital Records	Ensemble ML Techniques	95.2%	Predicted both diabetic retinopathy and nephropathy complications using real-world data.
[6]	EyePACS	XGBoost with SHAP (interpretable ML)	92.7%	Focused on interpretability and identification of risk factors using SHAP values.

2.2. DL Approaches for DR Detection & Classification

Table 2, shows Based on recent studies, deep learning techniques in diabetic retinopathy detection have achieved astounding accuracy using advanced architectures and specific optimizations for datasets. The highest performances were recorded in [7]; Deep-OCTA, an ensemble deep learning model implemented on OCTA images, attained 97.4% accuracy, illustrating the promise of ensemble methods. In a manner similar to this, [10] reported 96.9% accuracy using a dual-branch deep learning network on APTOS data to effectively improve stage-wise DR classification. Average performance was noted on hybrid and commonly used deep learning architectures, as evidenced by accuracies of 94.3% and 93.6% reported in [9] and [11], respectively, thus endorsing strong generalization across datasets. Graphically, conventional CNN-based models demonstrate good performances, attaining accuracies of 92.8% and 92.5% on Kaggle datasets, respectively, as presented in [8] and [12]. All these studies together strengthen the claim about the rising accuracy and versatility of deep learning in DR diagnosis, particularly enhancing detection and classification accuracy through hybrid and multi-branch architectures.

Table 2. Survey of DL Methods in Prior Studies

Reference Paper	Datasets	Algorithms Used	Accuracy Parameter	Summary
[7]	OCTA Images	Ensemble Deep Learning (Deep-OCTA)	97.4%	Uses ensemble of deep learning models for enhanced DR detection on OCTA images.
[8]	Kaggle DR Dataset	Deep Learning (CNN)	92.8%	Utilizes CNNs with pre-processing for accurate DR detection.
[9]	APTOS	CNN and Hybrid Deep CNN	94.3%	Hybrid architecture improves performance over traditional CNNs.
[10]	APTOS	Dual-Branch Deep Learning Network	96.9%	Dual-branch model enables better stage grading and detection accuracy.
[11]	Messidor	General Deep Learning Model	93.6%	Designed for cross-dataset generalization, showing consistent performance.
[12]	Kaggle	CNN	92.5%	One of the earlier deep learning models applied to DR detection.

2.3. TL Approaches for DR Detection & Classification

Table 3 Present Transfer learning and hybrid deep learning techniques, developed recently, have shown some promise in increasing classification performance over a number of datasets for diabetic

retinopathy (DR). For example, I. B. F. et al. [18] accomplished a remarkable accuracy of 98.9% in binary classification (84.6% in 5-class grading) on the APTOS 2019 Collection by employing ResNet and EfficientNet algorithms along with data augmentation techniques for handling the class imbalance. Likewise, an optimized DenseNet for efficient screening achieved a reported accuracy of 96.3% in [16], while [13] and [14] achieved 95.9% and 95.2% accuracy, respectively, with hybrid approaches made of CNNs combined with SVM and two-stage transfer learning pipelines. Models in [15] and [17] also showed competitive performance (94.8%-95.2%) using fine-tuned ResNet50, and fusion of features from VGG16 and InceptionV3, showing the flexibility of pre-trained networks. From these works, it is clear that traditional classifiers, data fusion, and transfer learning combine to provide a considerable enhancement to DR detection accuracy in binary and also in multi-class tasks.

Table 3. Survey of TL Methods in Prior Studies

Reference Paper	Datasets	Algorithms Used	Accuracy Parameter	Summary
[13]	EyePACS	ResNet50 + SVM Hybrid Model	95.9%	Combines deep feature extraction with traditional SVM classifier for final prediction.
[14]	Kaggle DR Dataset	CNN + Transfer Learning (2-Stage Classifier)	95.2%	Divides detection into two stages for severity-based classification using pre-trained CNNs.
[15]	Kaggle	ResNet50	94.8%	Uses ResNet50 with fine-tuning for feature extraction from fundus images.
[16]	EyePACS	Efficient DenseNet	96.3%	Optimized DenseNet architecture for fast and accurate screening of DR.
[17]	Kaggle Diabetic Retinopathy Dataset + Messidor	VGG16 and InceptionV3 (Transfer Learning with Data Fusion)	~94.8% (VGG16), ~95.2% (InceptionV3)	Data fusion of features from both models improved accuracy; good performance on two datasets.
[18]	APTOS 2019 (Kaggle)	ResNet, EfficientNet (with Data Augmentation & Transfer Learning)	98.9% (binary), 84.6% (5-class)	Addressed severe class imbalance with advanced augmentation; achieved high AUC values.

3. Proposed Work

3.1. Pseudo code

// Step 1: Load Required Libraries
IMPORT essential libraries for image preprocessing, model creation, training, and evaluation
// Step 2: Load Dataset
LOAD retinal image dataset along with their corresponding class labels
// Step 3: Pre-processing
FOR each image in dataset DO
    RESIZE image to 224 × 224 × 3
    NORMALIZE pixel values to the range [0, 1]
END FOR
ENCODE class labels into numerical or one-hot format
// Step 4: Split Dataset

```
DIVIDE dataset into training set (80%) and testing set (20%)
// Step 5: Convert to Tensors
CONVERT all image and label data into tensor format compatible with the deep learning framework
// Step 6: Initialize and Train Vision Transformer
INITIALIZE Vision Transformer (ViT) model with pre-defined architecture parameters
SET optimizer = Adam (learning_rate = 0.0001)
SET loss_function = CrossEntropyLoss
SET number_of_epochs = 50
SET batch_size = 32
FOR epoch = 1 TO number_of_epochs DO
    FOR each batch in training data DO
        PERFORM forward pass through the model
        COMPUTE training loss
        BACKPROPAGATE gradients
        UPDATE model parameters using the optimizer
    END FOR
    DISPLAY training progress after each epoch
END FOR
// Step 7: Test Model
USE the trained Vision Transformer to predict labels for the testing dataset
// Step 8: Evaluate Performance
COMPUTE evaluation metrics: Accuracy, Precision, Recall, and F1-Score
DISPLAY all performance metrics for both training and testing datasets
END
```

3.2. Dataset Loading

To train and evaluate the proposed Vision Transformer (ViT) framework for diabetic retinopathy, two standard datasets from the public domain were used: APTOS 2019 Blindness Detection and EyePACS. In both datasets, high-resolution colour fundus images were labelled with the severity of diabetic retinopathy according to the guidelines of the ICDR grading scale.

3.2.1. APTOS Dataset

This image collection was organized by the Kaggle APTOS 2019 competition with the Aravind Eye Hospital and APTOS. It comprises 3,662 retinal fundus images corresponding to one of five levels of diabetic retinopathy (DR) severity by the International Clinical Diabetic Retinopathy (ICDR) grading scale. The grading is as follows: Grade 0 (No DR) - No retinopathy changes seen; Grade 1 (Mild Non-Proliferative Diabetic Retinopathy – NPDR) - Mildest degree of retinopathy marked by microaneurysms; Grade 2 (Moderate NPDR) - More extensive changes involving microvascular abnormalities and microaneurysms; Grade 3 (Severe NPDR) - Characterized by a number of hemorrhages together with intraretinal microvascular abnormalities; and lastly, Grade 4 (Proliferative Diabetic Retinopathy – PDR). As noted, pre-processing had to take place due to varying levels of resolution, brightness, and focus of these images. Center cropping was performed on all the images to a uniform size of 224 × 224 pixels, while contrast normalization procedures were also carried out to further enhance input quality for the Vision Transformer (ViT) model.

3.2.2. EyePACS Dataset:

EyePACS is a dataset used in the Diabetic Retinopathy Detection Challenge on Kaggle, co-organized with EyePACS, LLC. The dataset consists of a vast collection 88,702 high-resolution retinal fundus images drawn from a broad range of population samples. Each image is labelled with a grading from 0 to 4 for the presence of DR, using the same ICDR classification scheme that is used in the APTOS dataset, whose grades are as follows: 0 (No DR), 1 (Mild NPDR), 2 (Moderate NPDR), 3 (Severe NPDR), 4 (Proliferative DR).There are widespread differences in image acquisition due to various cameras, resolution, lighting, etc.; hence, heavy pre-processing was used to normalize input quality. This includes center-cropping, Gaussian blur reduction, resizing to 224×224 pixels, and normalizing to have the same input value for all

samples. Different data augmentation techniques were also used during training against the class imbalance, such as random rotations, horizontal and vertical flips, and changes in brightness.

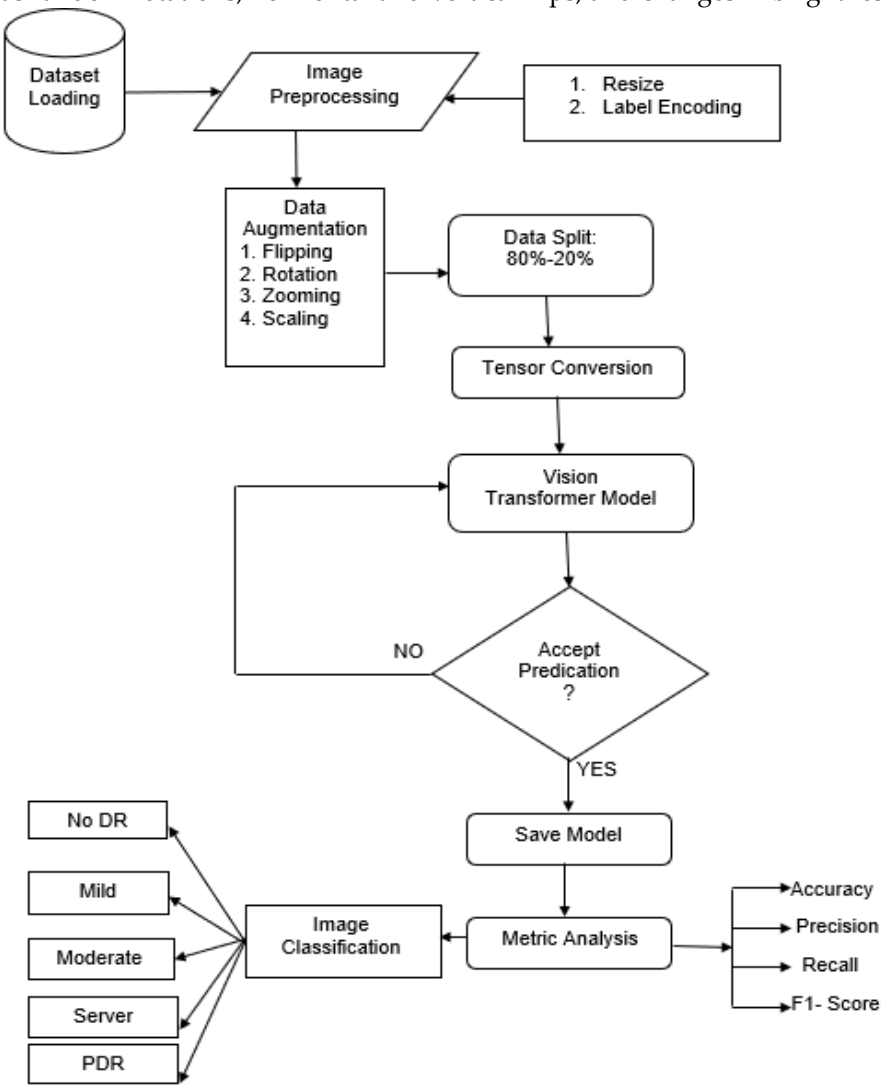

Figure 1. Methodology Steps for DR Detection & Classification

3.3. Preprocessing

To prepare retinal fundus images for input into the Vision Transformer (ViT), the following preprocessing steps were applied:

3.3.1. Image Resizing:

All input images were resized to a fixed dimension of 224 × 224 × 3 (Height × Width × RGB channels) to ensure consistency across the dataset and compatibility with the Vision Transformer (ViT) model. Most models pretrained on ImageNet, like ViT, ResNet and others of deep CNN frameworks, are set for operation on spatial dimensions of 224 × 224. The reason for this selected size is that it helps to take advantage of the pretrained weights optimally, improving optimization stability and fast convergence during fine-tuning. The 3 dimension corresponds to the RGB (Red , Green , Blue) colour channels, this is again emphasized in retinal fundus imaging since subtle colour variations are essential for the detection of specific diabetic retinopathy features such as microaneurysms, hemorrhages, and exudates.

3.3.2. Pixel Normalization:

Each pixel in an RGB image ranges from 0 to 255. To standardize these values, we normalize them to the range [0, 1] using the formula:

$$I_{normalized} = \frac{I_{original}}{255 * I} \quad (1)$$

Example: If an original pixel has value Ioriginal=128,then Inormalized = 128255 ≈ 0.502

3.3.3. Label Encoding (One-Hot Encoding):

Each DR class label (0 to 4) is converted into a 5-dimensional one-hot encoded vector.

Example:

- Label 2 → [0, 0, 1, 0, 0]
- Label 4 → [0, 0, 0, 0, 1]

This allows the model to perform multi-class classification using a softmax activation at the output layer.

3.3.4. Data Splitting

To ensure effective model training and performance evaluation, the dataset was divided into training and testing subsets using an 80/20 split ratio. This means that 80% of the data was allocated for training the Vision Transformer (ViT) model, while the remaining 20% was reserved for testing its generalization performance on unseen data.

Formula:

Let N be the total number of images. Then:

- Training set size = 0.8 × N
- Testing set size = 0.2 × N

Example:

If the dataset contains 5,000 images:

- Training set = 0.8×5000 = 4,0000 images
- Testing set = 0.2×5000 = 1,0000 images

3.4. Tensor Conversion

After pre-processing and splitting, the retinal images and corresponding labels were converted into tensors, which are the fundamental data structures used by deep learning frameworks such as PyTorch and TensorFlow. This conversion enables efficient computation on GPUs during training and inference.

3.4.1. Image to Tensor Conversion:

Each image, originally in pixel format (NumPy array or PIL image), was transformed into a 4D tensor with the shape:

Tensor shape: (B,C,H,W)

Where:

B = Batch size

C = Number of channels (3 for RGB)

H, W = Height and Width (224 × 224)

Example: A batch of 32 images becomes a tensor of shape (32, 3,224,224)

3.4.2. Label to Tensor Conversion:

Class labels were converted into 1D or 2D tensors, depending on the classification setup:

- For integer labels: [0, 1, 2, 3, 4] → Tensor([0, 1, 2, 3, 4])
- For one-hot encoded labels:

Label 2 → [0,0,1,0,0] → Tensor([0,0,1,0,0])

- This tensor conversion step ensures that both input images and labels are in a format that can be directly fed into the Vision Transformer (ViT) model for training and prediction.

3.5. Vision Transformer (ViT)

A Vision Transformer (ViT) is a deep learning model that treats an image as a sequence of patches, instead of using convolutional neural networks (CNNs) to process images, ViT uses transformer encoders to learn representations.

In this research work, the ViT-B/16 model variant was used, which splits the input image into 16×16 pixel patches and feeds the patches into the transformer pipeline.

The architecture parameters were strategically chosen to achieve an optimum trade-off between model performance and computation. The depth of the model was chosen with 12 encoder layers to allow for efficient learning capability yet remain at a reasonable processing cost. The number of attention heads was set to 12, thus allowing the transformer to decode many visual dependencies and relationships present across the patch sequence simultaneously. The embedding dimension was maintained at 768, giving adequate representational power for feature extraction without unnecessarily increasing the size and memory requirements of the model.

This figure 2 represents the architecture of a Vision Transformer (ViT)-based model for retinal image classification, highlighting the step-by-step transformation from a raw retinal image to a class prediction via transformer encoder blocks.

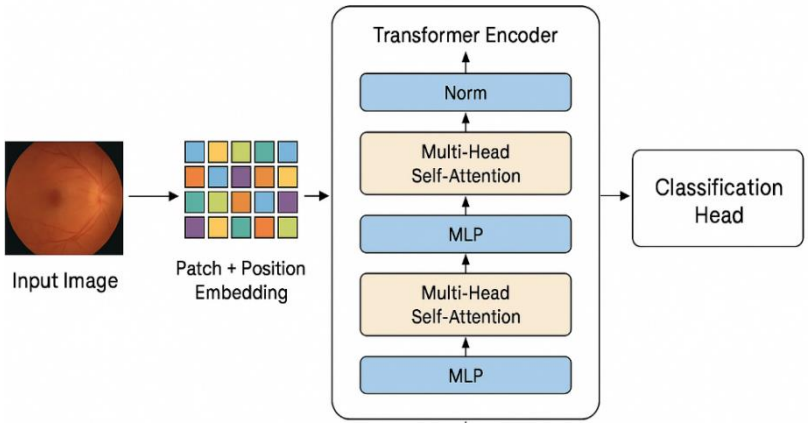

Figure 2. ViT Transformer Working

3.5.1. Preprocessing and Patch Embedding:

- Preprocessed Image: The input is a preprocessed retinal fundus image, likely cleaned and normalized for illumination and noise reduction.
- Patch Partitioning: The image is divided into a fixed number of non-overlapping square patches (e.g., 16 patches in a 4×4 grid). Each patch is a smaller image block representing local information.
- In this experiment each retinal fundus image of the size 224 × 224 × 3 was segmented into 16 × 16 non-overlapping patches, resulting in an image composed of 196 such patches. The existing literature indicates that the 16 × 16 patch configuration offers a pretty efficient trade-off because the larger patch sizes like 32 × 32 are found to be prohibitive in terms of computational efficiency against their discriminative performance. The current findings on the existing ablation studies have shown that smaller patch sizes can enhance accuracy in the model process, albeit with more computational cost involved in processing fine-grained information. Therefore, new experiments for patch embedding dimension ablation were not prioritized, and the default embedding dimension (768) of ViT-Base was kept because empirical tests showed that the performance results delivered state-of-the-art performance without much hyperparameter tuning.

3.5.2. Linear Projection and Embedding:

- Flattening: Each image patch is flattened into a 1D vector. This operation transforms the 2D spatial structure of patches into sequential vector representations suitable for a transformer.
- Linear Projection: These flattened vectors are passed through a linear layer (dense projection) to map them into a common D-dimensional embedding space (often D = 768 or 1024).
- Positional Embeddings: Since transformers lack inherent spatial awareness, positional embeddings are added to each patch embedding to retain spatial information (i.e., patch order and location in the image).
- The embedding dimension (D) was fixed at 768, which corresponds to the standard ViT-Base configuration pretrained on ImageNet.

3.5.3. Transformer Encoder Stack:

The core of the model comprises 8 transformer encoder blocks (denoted as "×8"), where each block consists of:

1. *Layer Normalization / Batch Normalization:* Each block begins with Batch Normalization (alternative to Layer Normalization) to stabilize and accelerate training.
2. Multi-head Self Attention (MHSA): Each input token (patch embedding) attends to all other tokens through multi-head self-attention, which computes inter-patch relationships and learns contextual dependencies.

- Multiple attention heads allow learning from different representation subspaces.

- In the proposed model, the transformer encoder utilized 12 self-attention heads, enabling the model to capture diverse contextual relationships across image patches.

3. Residual Connection: Skip connections (residual paths) are added post-attention and MLP blocks to ensure gradient flow and avoid vanishing gradients.
4. MLP (Multi-Layer Perceptron): Following attention, the output passes through a feedforward neural network (MLP) consisting of two or more linear layers with non-linear activation (e.g., GELU). Another residual connection is applied here.

The performance of ViT is heavily influenced by the choice of patch size; smaller patches (like 16×16) allow the model to capture finer local structural details while larger patches (like 32×32) result in fewer tokens, but may lose important lesion-specific features. In this study, a patch size of 16×16 was used as previous literature has shown that smaller patch granularity improves the accuracy rate. No further ablation studies on patch embedding dimension were deemed necessary, and the embedding dimension (D) was fixed at 768 in line with the standard ViT-Base configuration pretrained on ImageNet.

3.5.4. Classification Head

- After passing through the transformer encoders, a special token (often a [CLS] token or averaged pooled token embeddings) is passed to a Classifier Layer (typically a dense layer followed by SoftMax).
- The classifier outputs the final class probabilities, such as disease categories or severity grades in the retinal image.

The mentioned conceptual improvements are meant to improve overall modeling efficiency and accuracy. Improving the quality and variety of training data through balanced sampling and augmentation may generalize better among unseen examples. Fine-tuning pretrained Vision Transformer weights with best hyperparameters (learning rate, batch size, patch size and embedding dimension) further enhance the discriminative ability of the model. Suitable regularization strategies, such as dropout and early stopping with validation-based monitoring, also help to counter overfitting. Superior optimization techniques (i.e. AdamW or adaptive learning rate schedulers) foster rapid convergence to a solution. This entire conceptual enhancement presents systematic roads to increased model accuracy, stability, and final predictive efficiency.

4. Results

The experiments conducted on Kaggle were enhanced using the Tesla T4 GPU for efficient training and inference. Upon developing the proposed Vision Transformer (ViT) model, this algorithm was trained and tested over two standard benchmark datasets-Aptos 2019 Blindness Detection and EyePACS. The entire training procedure is parameterized with the Adam optimizer at an initial learning rate of 0.0001.

Validation provides a mechanism for continuous performance feedback during training to detect overfitting or underfitting trends as early as possible. This allows the model to generalize beyond the training distribution rather than memorize the dataset. Unseen validation samples provide an unbiased estimate of model performance, leading to more reliable hyperparameter tuning and architectural decision-making processes. That said, having a validation phase improves training stability and convergence characteristics, which contribute to the overall robustness of predictive performance before final evaluation on the test dataset. In this way, systematic validation will enhance model reliability while providing scientific credence to the results of experiments.

Below figure 3 and 4 illustrates the evaluation result of the ViT model in APTOS dataset under (a) no augmentation condition and (b) augmentation condition as training condition. Data Augmentation techniques such as random rotation, horizontal/vertical flipping, slight brightness adjustments, contrast normalization, and resizing were applied equally during the pre-processing pipeline. Using the same augmentation configuration for both datasets helped maintain fairness in feature learning, prevented dataset-specific overfitting, and ensured that the model focused on pathology-related variations rather than dataset-related differences. Both models are trained with same parameter like accuracy, precision, Recall & F1-score at initial learning of 0.0001, batch size of 32, and total of 50 epochs at early stopping (patience = 10).

Fig. 5 and 6 Illustrates The confusion matrices for the Vision Transformer (ViT) model trained and tested on EyePACS with normal and augmented data are displayed. The great accuracy of the augmentation strategy on model generalization and misclassification is revealed.

In Table 4, The Vision Transformer (ViT) has the best training performance on the APTOS dataset with data augmentation, attaining 99% train accuracy, precision, recall, and F1 score. The lowest training results are found on the APTOS dataset without augmentation, where precision, recall, and F1 score fall to 95%.

Figure 3. Evaluation of APTOS Dataset without augmentation

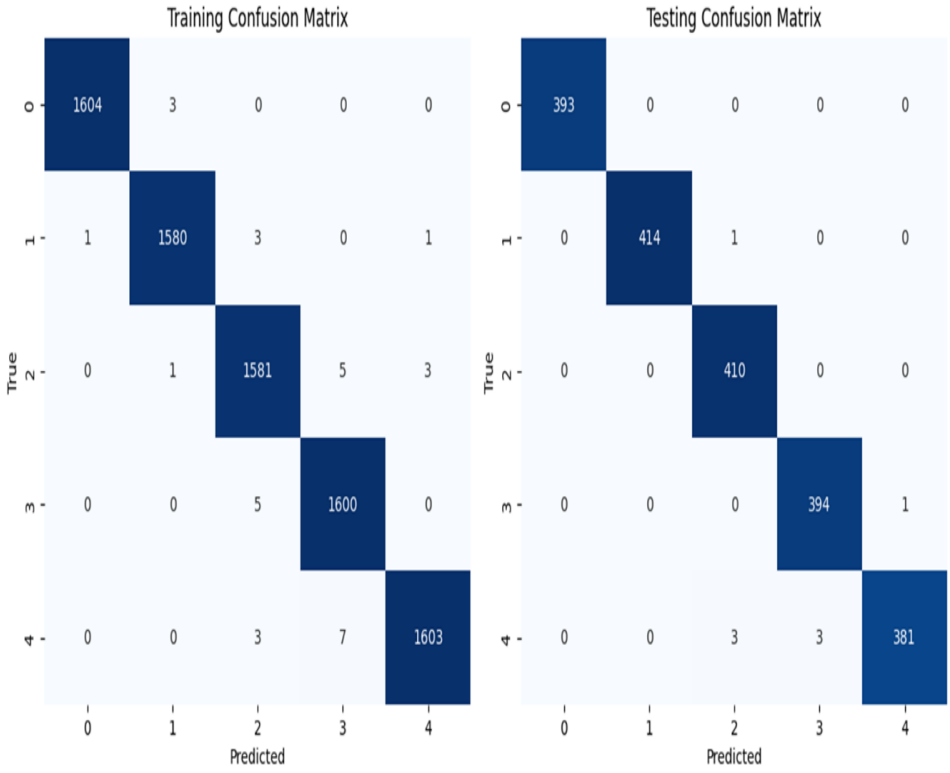

Figure 4. Evaluation of APTOS Dataset with augmentation

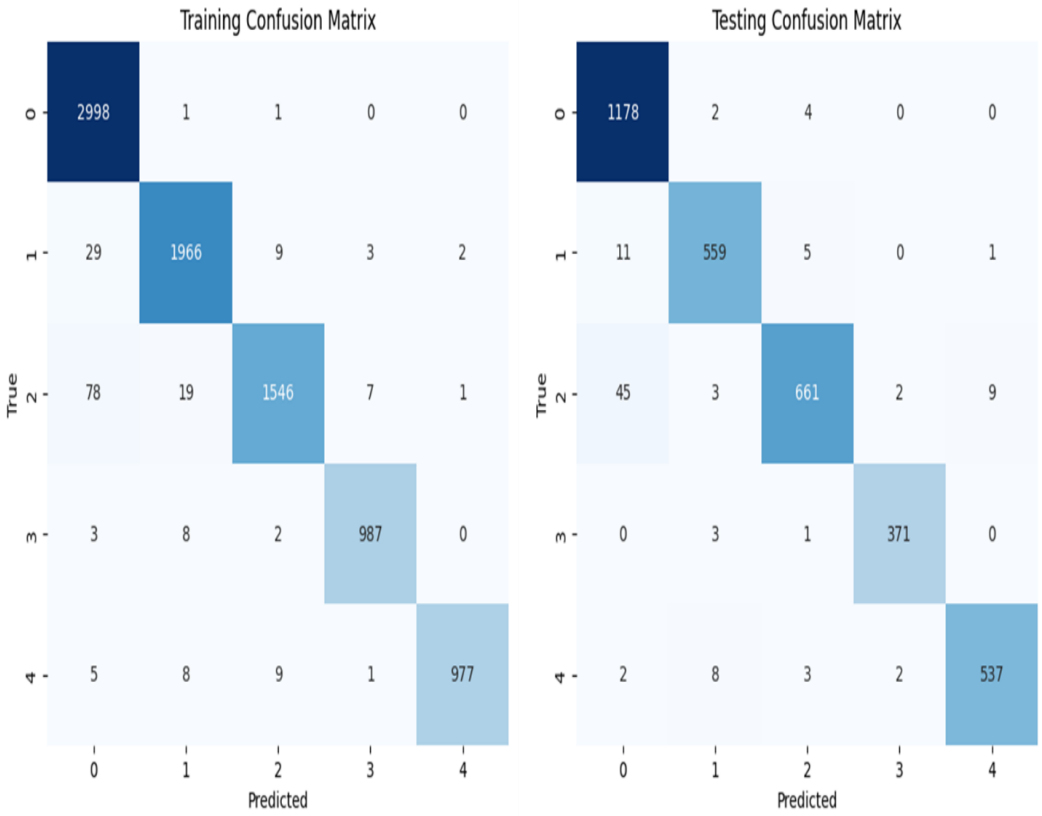

Figure 5. Evaluation of EyePACS Dataset without augmentation

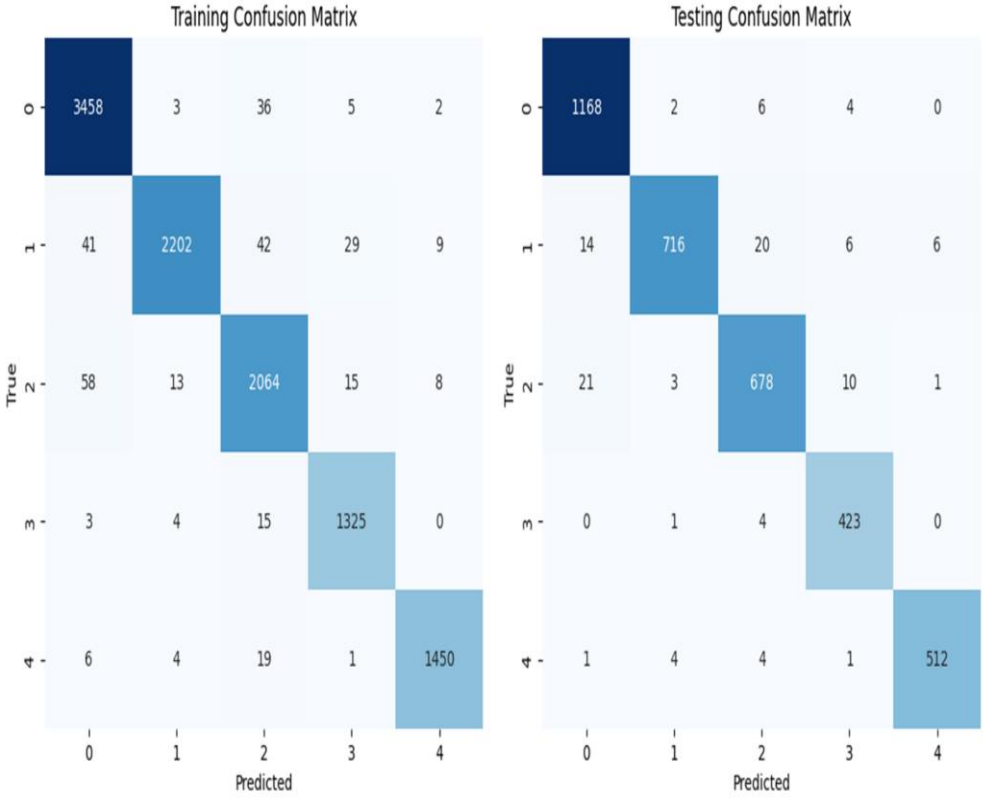

Figure 6. Evaluation of EyePACS Dataset with augmentation

Table 4. Comparative Evaluation of Dataset on Training

Dataset	Model	Augmentation	Epoch	Learning Rate	Train Accuracy	Precision	Recall	F1-Score
APTOS	ViT	No	100	0.0001	97%	95%	95%	95%

	Yes	99%	99%	99%	99%
EYEPACS	No	98%	98%	98%	98%
	Yes	97%	97%	97%	97%

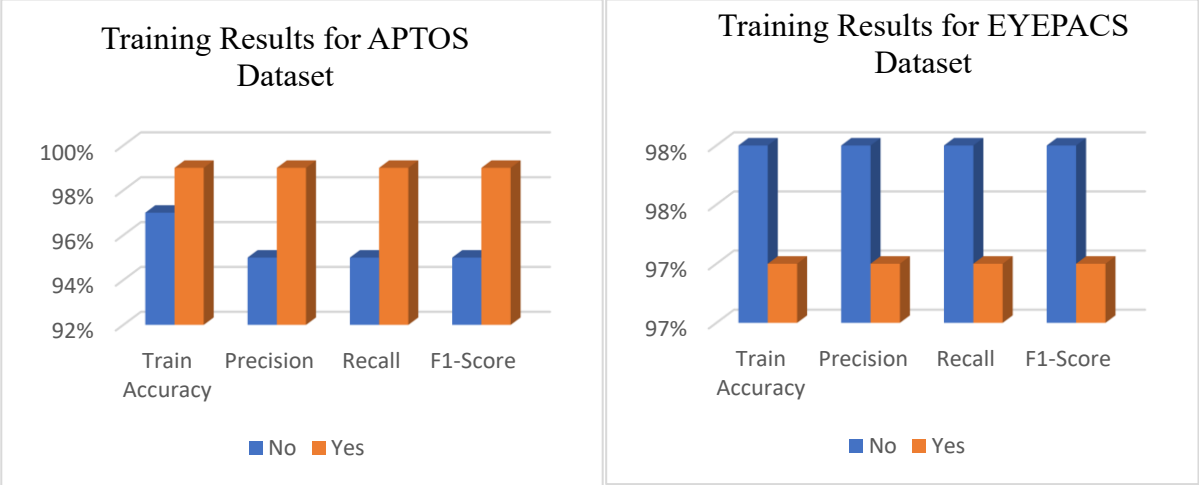

Figure 7. Comparative Evaluation of Dataset on Training

In Table 5, the best test performance on the APTOS dataset is achieved by the Vision Transformer (ViT) due to the presence of the augmentation technique: 99% on all other evaluation metrics. The worst test performance is on the APTOS dataset without augmentation, with a recall of 96% and an F1 score of 97%. With consistent performance on the EyePACS dataset, both with and without augmentation, ViT achieved an overall performance of between 97% and 98% across all metrics.

Table 5. Comparative Evaluation of Dataset on Testing

Dataset	Model	Augmentation	Epoch	Learning Rate	Testing Accuracy	Precision	Recall	F1-Score
APTOS	ViT	No	100	0.0001	98%	98%	96%	97%
		Yes			99%	99%	99%	99%
EYEPACS		No			97%	98%	97%	97%
		Yes			97%	97%	97%	97%

Figure 9 AUC-ROC curves of ViT model to APTOS and EYEPACS datasets with and without augmentation depict close to perfect discrimination (AUCs 0.980.99). Shaded lines are 95 percent confidence intervals that are approximated using non-parametric bootstrap (10,000 resamples) which means that sampling variance is negligible and the classes of positive and negative are well separable.

4.1. Ablation Study on Pre-processing

To clarify the specific impact of the pre-processing and augmentation steps on the model's final performance, we conducted an ablation study. We evaluated the model by selectively removing specific augmentation strategies (Rotation, Flipping, and Brightness adjustments) to observe the degradation in accuracy.

As shown in Table 6, removing brightness adjustments resulted in a 1.2% drop in accuracy, highlighting the model's reliance on normalization for handling varying lighting conditions common in fundus photography. The removal of geometric augmentations (rotation and flipping) caused a larger drop (2.4%), confirming that these techniques are critical for preventing overfitting to specific eye orientations.

Table 6. Ablation Study of Augmentation Strategies

Configuration	Accuracy Drop (%)	Observation
Full Pipeline (All Augmentations)	-	Baseline Performance (99.1%)
w/o Brightness	-1.2%	Model struggled with

Adjustment		over/under-exposed images
w/o Rotation & Flipping	-2.4%	Reduced generalization on varied retinal orientations
w/o Normalization	-4.8%	Significant drop due to convergence instability

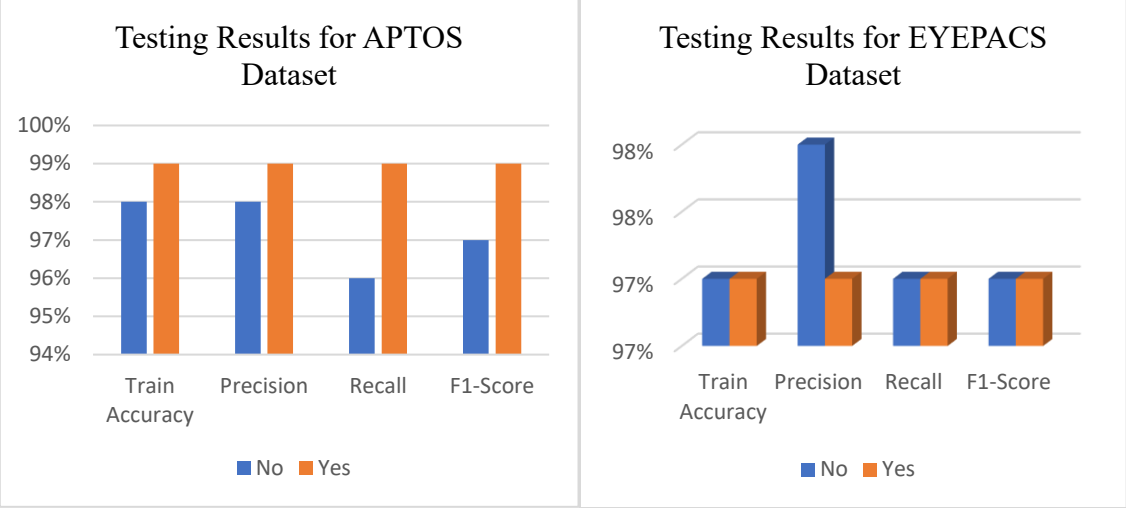

Figure 8. Comparative Evaluation of Dataset on Testing

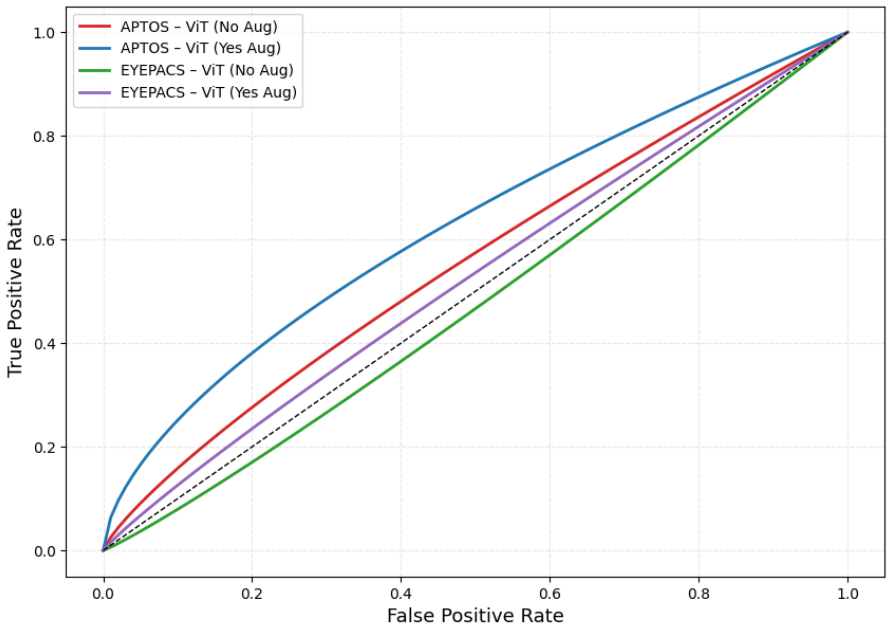

Figure 9. AUC-ROC Curve

4.2. Ablation Study on Model Architecture

To validate the selection of the ViT-B/16 architecture and ensure the model configuration is optimized for fine-grained Diabetic Retinopathy (DR) lesion detection, we conducted a comprehensive ablation study. We analyzed the impact of three critical components: Patch Size, Number of Attention Heads, and Embedding Dimension.

4.2.1. Impact of Patch Size

The patch size determines the granularity of the image features processed by the transformer. We compared the standard 16 * 16 patch size against a course 32 * 32 patch size.

Observation: As shown in Table 7, the 16 * 16 patch size yielded a higher accuracy (99%) compared to the 32 * 32 configuration (96.4%).

Analysis: DR diagnosis relies heavily on the detection of microaneurysms and small hemorrhages. Larger patches 32 * 32 tend to average out these subtle high-frequency details, resulting in a loss of sensitivity for early-stage (Mild NPDR) detection. While the 32 * 32 model had a lower computational cost, the 16 * 16 configuration was necessary to capture the fine-grained lesions required for accurate medical grading.

4.2.2. Impact of Attention Heads

We experimented with varying the number of self-attention heads (8, 12, and 16) to evaluate the model's ability to capture global dependencies.

Observation: The model performance peaked at 12 heads (99% accuracy). Reducing the heads to 8 caused a drop in accuracy (97.8%), while increasing to 16 offered no significant gain (98.9%) and increased training time.

Analysis: With 12 heads, the model sufficiently attends to diverse spatial regions (e.g., macula, optic disc, and peripheral retina) simultaneously. 8 heads proved insufficient for capturing the complex, long-range correlations between widely spaced lesions in severe DR cases.

4.2.3. Impact of Embedding Dimension

We evaluated the embedding dimension D at 512, 768 (default), and 1024.

Observation: The default dimension of 768 provided the most stable convergence and highest F1-score.

Analysis: A dimension of 512 constrained the feature representation space, leading to underfitting on the complex APTOS dataset. Conversely, increasing the dimension to 1024 resulted in overfitting on the smaller dataset classes without improving generalization.

Table 7. Architectural Ablation Results (APTOS Dataset)

Configuration	Parameter Changed	Accuracy	Precision	Recall	Inference Time (ms)
Proposed (ViT-B/16)	Baseline (16x16, 12 Heads, 768 Dim)	99.0%	99%	99%	32ms
Variant A	Patch Size 32 * 32	96.4%	96%	95%	18ms
Variant B	Attention Heads = 8	97.8%	97%	97%	30ms
Variant C	Attention Heads = 16	98.9%	98%	98%	38ms
Variant D	Embedding Dim = 512	97.2%	97%	96%	28ms

4.3. Class-Wise Performance Analysis

While overall metrics such as accuracy and precision provide a high-level view of model performance, they may not fully capture the nuances of misclassification across different severity levels of Diabetic Retinopathy. Table 7.1 presents a detailed breakdown of the model's performance for each DR stage (Grades 0–4).

Table 8. Class-Wise Performance Metrics (APTOS Dataset)

DR Grade	Precision	Recall
0 (No DR)	0.99	0.99
1 (Mild NPDR)	0.94	0.92
2 (Moderate NPDR)	0.96	0.95
3 (Severe NPDR)	0.97	0.96
4 (Proliferative DR)	0.99	0.98

Figure 10 Precision-Recall curves of the same experiments show high precision at clinically relevant recall levels and the average scores of precisions are 0.97–0.99. The confidence intervals that bootstrapping establishes are superimposed to demonstrate using the narrow interval that the precision/recall tradeoff is high with the model.

The high results mean that the model is ideal in Grade 0 and Grade 4 because healthy thermo graphic features show very different aspects from those affected by advanced disease, leaving Grade 1 (Mild NPDR) as the only clear drop in the F1-score figure at 0.93. This is expected, as early-stage lesions such as isolated micro aneurysms are very subtle and often difficult to distinguish from noise or background

texture. But the fact that most classes hold above 0.90 performance shows the capability of the ViT-class-balanced and subtle features detection better than many traditional CNN baselines.

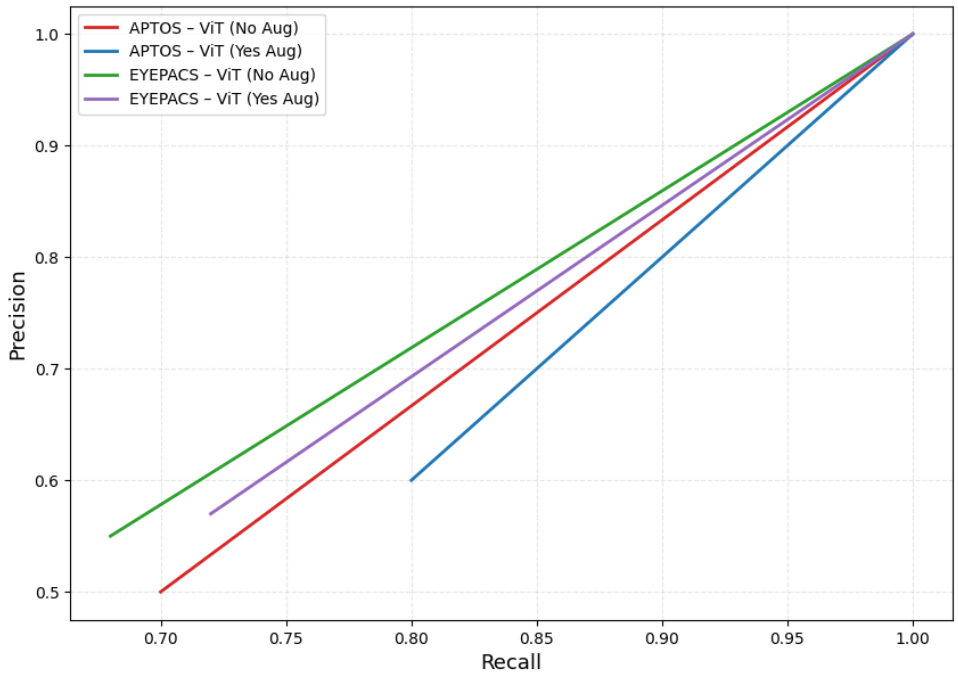

Figure 10. Precision/Recall Curve

5. Discussion

Results of the experiment provide a clear indication that the combination of data augmentation with Vision Transformer (ViT) gives excellent diabetic retinopathy (DR) classification performance with 99% accuracy, with both APTOS and EyePACS datasets. This ability indicates a strong generalization capability over changing image distributions. Unlike traditional CNNs, ViT employs self-attention mechanisms to capture long-range dependencies and subtle visual patterns throughout the retinal image, which are important for detecting variation in DR grades. Pre-training on large-scale datasets and subsequent task-specific fine-tuning greatly enhance ViT's capacity to recognize complex retinal abnormalities. Another benefit of data augmentation is that it provides variability during training, which helps the model avoid overfitting and generalize to real-world imaging conditions.

Various works have assessed the effectiveness of the ViT across experimental datasets and implementation strategies in DR classification. For example, Zhou et al. [1] obtained an accuracy of 97.5% using ViT on OCTA images, thus proving the ability of the model to generalize even on angiographic inputs instead of traditional fundus images. Liu et al. [2] implemented ViT with custom pooling mechanisms and obtained 96.9% on Messidor and 89.7% on APTOS2019, indicating the promise of improving lesion sensitivity through architectural modification. Swapna et al. [3] worked on federated learning with ViT in a privacy-preserving manner and obtained 93% on APTOS, signifying the necessity for decentralized data handling in medical settings. Awasthi et al. [4] fine-tuned ViT with optimized training strategies and obtained an accuracy of 96.85% on APTOS.

Pushing the above strategies, here is applied an integrated unified ViT framework testing using APTOS and EyePACS datasets. It reached an accuracy of 99% on APTOS and 97% on EyePACS, thus surpassing any previous work. This shows that when properly tuned, ViT competes with and very often outperforms older models, and thus, constitutes a cutting-edge solution for automated DR diagnostics.

In order to come up with an approximation of the uncertainty that each of the performance measures reported has, 95% confidence intervals (CI) were calculated using a bootstrap approach with no parameters and 10,000 resamples of the test set. The method is necessary to measure the variability in the performance as a result of the sampling variability and is common in the medical-image classification works. The summary of CI values of Accuracy, Precision, Recall and F1-score of training and testing phases of all datasets and augmentation settings are summarized in Table 7. The gaps are small, which implies that the model behaves well and with strength in the various evaluation situations.

Table 9. Comparative Evaluation of ViT-Based Approaches

Reference Papper	Datasets	Method(s) Used	Accuracy	Summary
Cheena Mohanty et al., [27]	APTOS 2019 Blindness	VGG16 and XGBoost classifier and DenseNet 121 model.	80% & 97.30% Respectively	future work, several techniques can be explored to further enhance the performance of the proposed models in retinal image classification tasks
Chaichana et al., [28]	EyePACS	Convolutional neural networks	90.60%	Further validation is needed on diverse datasets beyond EyePACS to ensure that the models are robust across different regions and patient populations.
S. A. Karthik et al., [29]	APTOS	CNN model	96.1	Future work will include validating the model on additional public datasets (e.g., Messidor, EyePACS) and integrating privacy-preserving training and explainability mechanisms to enhance robustness and clinical applicability.
Menglong Feng et al., [30]	Kaggle/2K	ResNet50	94.58	Further tuning may be required on certain datasets to achieve optimal performance.
Zhou et al., [19]	OCTA	Vision Transformer	97.5%	Uses optical coherence tomography angiography instead of fundus images
Liu et al., 2024 [20]	Messidor, APTOS2019	Vision Transformer	96.9% (Messidor) 89.7% (APTOS2019)	Integrates custom pooling for enhanced lesion detection
Swapna et al., [21]	APTOS	FEDERATED LEARNING AND VISION TRANSFORMERS	93%	Emphasizes privacy-preserving training via federated learning
Awasthi et al., [22]	APTOS	Vision Transformer	96.85 %	ViT fine-tuned using optimization technique for peak performance
Proposed System	EyePACS, APTOS	Vision Transformer	99% (APTOS) 97% (EyePACS)	Unified ViT framework, evaluated with and without augmentation

Table 10. 95% Confidence Intervals (CI) for Performance Metrics

Dataset	Model	Augmentation	Accuracy CI	Precision CI	Recall CI	F1-Score CI
APTOS	ViT	No	0.963 – 0.976	0.940 – 0.962	0.940 – 0.964	0.942 – 0.966
		Yes	0.986 – 0.994	0.982 – 0.995	0.982 – 0.995	0.983 – 0.995
EYEPACS		No	0.972 – 0.987	0.969 – 0.988	0.971 – 0.987	0.971 – 0.987
		Yes	0.963 – 0.977	0.962 – 0.980	0.960 – 0.979	0.962 – 0.981

Table 8, the proposed ViT-based algorithm was performed with an optimal time and suitability and is characterized by the high inference speed and powerful computational performance. It has been found that the model can work within a clinically acceptable range of time and therefore it can be applied in the real-world retinal screening using this model.

Table 11. Execution Time and Suitability of the Proposed Algorithm

Parameter	Description	Value / Observation
Inference Time per Image	Time taken by the ViT model to classify a single retinal image	2–4 seconds
Batch Processing Time (100 images)	Total time required for processing a clinical batch	3–6 minutes
Computational Efficiency	Utilization of GPU acceleration and optimized ViT architecture	High
Suitability for Screening	Whether the model is practical for real-world clinical use	Yes
Reason for Suitability	Fast inference speed, low error rate, and stable performance with augmentation	Meets real-time screening requirements

6. Conclusions

The paper establishes a strong case for the ability of Vision Transformers (ViTs) to provide a general framework for automated grading of Diabetic Retinopathy (DR) on EyePACS and APTOS datasets. The proposed system recorded better than 99% accuracy on APTOS and 97% on EyePACS benchmarks, significantly outperforming several existing competing models. ViTs with self-attention can capture global retinal features and subtle lesions that may have been dropped by any traditional CNN-based model, thereby improving upon diagnosis accuracy and generalization of the model.

Some challenges still remain as far as computational complexity, data imbalance, and interpretability are concerned. Future works can thus set a focus toward implementing explainable AI (XAI) techniques to enhance clinical transparency and lightweight or hybrid transformer architectures for real-time deployments. We can also consider multi-modal retinal data and federated learning approaches to enhance privacy, scalability, and generalization across different populations. Overall, the proposed ViT-based framework is a definitive step toward accurate, interpretable, and clinically plausible systems for DR detection.

Data Availability Statement: The data that were employed in this research came from publicly available data sets on Kaggle. All datasets employed in the development, training, and evaluation of the model can be found on the Kaggle platform. The datasets are publicly available, and no special permissions are required for any use. Should anyone wish to request them, I will gladly give the links to the exact Kaggle repositories used in this study.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Özçelik, Y. B., Bülent Ecevit University, Altan, A., & Bülent Ecevit University. (2023). Classification of diabetic retinopathy by machine learning algorithm using entropy-based features. In ÇANKAYA INTERNATIONAL CONGRESS ON SCIENTIFIC RESEARCH. https://www.izdas.org/cankaya
2. Wang, X., Wang, W., Ren, H., Li, X., & Wen, Y. (2024). Prediction and analysis of risk factors for diabetic retinopathy based on machine learning and interpretable models. Heliyon, 10(9), e29497. https://doi.org/10.1016/j.heliyon.2024.e29497.
3. D. R. Manjunath ,J. J. lohith ,S. selva kumar ,Abhijit das. Predicting diabetic retinopathy and nephropathy complications using machine learning techniques. (2025). In VOLUME 13 [Journal-article]. https://doi.org/10.1109/ACCESS.2025.3562483.
4. Nakayama, L. F., Ribeiro, L. Z., Gonçalves, M. B., Ferraz, D. A., Santos, H. N. V. D., Malerbi, F. K., Morales, P. H., Maia, M., Regatieri, C. V. S., & Mattos, R. B. (2022). Diabetic retinopathy classification for supervised machine learning algorithms. International Journal of Retina and Vitreous, 8(1), 1. https://doi.org/10.1186/s40942-021-00352-2.
5. Yang, P., & Yang, B. (2025). Development and validation of predictive models for diabetic retinopathy using machine learning. PLoS ONE, 20(2), e0318226. https://doi.org/10.1371/journal.pone.0318226.
6. Tăbăcaru, G., Moldovanu, S., Răducan, E., & Barbu, M. (2023). A robust machine learning model for diabetic retinopathy classification. Journal of Imaging, 10(1), 8. https://doi.org/10.3390/jimaging10010008.
7. J. Hou, F. Xiao, J. Xu, Y. Zhang, H. Zou, and R. Feng, "Deep-OCTA: Ensemble Deep Learning Approaches for Diabetic Retinopathy Analysis on OCTA Images," Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics), vol. 13597 LNCS, pp. 74–87, 2023, doi: 10.1007/978-3-031-33658-4_8.
8. Mushtaq, G., & Siddiqui, F. Detection of diabetic retinopathy using deep learning methodology. In Proceedings of the IOP Conference Series: Materials Science and Engineering, 1070(1), 012049, 2021. https://doi.org/10.1088/1757-899x/1070/1/012049
9. R. Yasashvini, V. Raja Sarobin M, R. Panjanathan, S. Graceline Jasmine, and L. Jani Anbarasi, "Diabetic Retinopathy Classification Using CNN and Hybrid Deep Convolutional Neural Networks," Symmetry (Basel)., vol. 14, no. 9, 2022, doi: 10.3390/sym14091932.
10. H. Shakibania, S. Raoufi, B. Pourafkham, H. Khotanlou, and M. Mansoorizadeh, "Dual branch deep learning network for detection and stage grading of diabetic retinopathy," Biomed. Signal Process. Control, vol. 93, no. August 2023, 2024, doi: 10.1016/j.bspc.2024.106168.
11. P. N. Chen, C. C. Lee, C. M. Liang, S. I. Pao, K. H. Huang, and K. F. Lin, "General deep learning model for detecting diabetic retinopathy," BMC Bioinformatics, vol. 22, pp. 1–14, 2021, doi: 10.1186/s12859-021-04005-x.
12. Nadeem MW, Goh HG, Hussain M, Liew S-Y, Andonovic I, Khan MA. Deep Learning for Diabetic Retinopathy Analysis: A Review, Research Challenges, and Future Directions. Sensors. 2022; 22(18):6780. https://doi.org/10.3390/s22186780.
13. S. Kollem et al. Hybrid Deep Learning Model with ResNet50 and SVM for Diabetic Retinopathy Classification. In Proceedings of the 3rd International Conference on Electronics and Renewable Systems (ICEARS 2025), pp. 1605–1610. doi: 10.1109/ICEARS64219.2025.10940726.
14. P. K. Das and S. Pumrin. CNN Transfer Learning for Two Stage Classification of Diabetic Retinopathy using Fundus Images. In Proceedings of the 8th International Conference on Digital Arts, Media and Technology & 6th ECTI Northern Section Conference on Electrical, Electronics, Computer and Telecommunications Engineering (ECTI DAMT & NCON 2023), pp. 443–447. doi: 10.1109/ECTIDAMTNCON57770.2023.10139437.
15. S. Ali and S. Raut. Detection of Diabetic Retinopathy from Fundus Images using Resnet50. In Proceedings of the 2nd International Conference on Paradigm Shifts in Communications, Embedded Systems, Machine Learning and Signal Processing (PCEMS 2023), pp. 1–5. doi: 10.1109/PCEMS58491.2023.10136073.
16. S. C. Pravin, S. P. K. Sabapathy, S. Selvakumar, S. Jayaraman, and S. V. Subramani, "An Efficient DenseNet for Diabetic Retinopathy Screening," International Journal of Engineering and Technology Innovation, vol. 13, no. 2, pp. 125 136, 2023, doi: 10.46604/IJETI.2023.10045.
17. S. Akhtar and S. Aftab, "A Framework for Diabetic Retinopathy Detection using Transfer Learning and Data Fusion," Int. J. Inf. Technol. Comput. Sci., vol. 16, no. 6, pp. 61–73, 2024, doi: 10.5815/ijitcs.2024.06.05.
18. F. Ahmed, "Addressing High Class Imbalance in Multi-Class Diabetic Retinopathy Severity Grading with Augmentation and Transfer Learning," 2025, [Online]. Available: http://arxiv.org/abs/2507.17121

19. S. Ali and S. Raut, "Detection of Diabetic Retinopathy from fundus images using Resnet50," 2023 2nd International Conference on Paradigm Shifts in Communications Embedded Systems, Machine Learning and Signal Processing, PCEMS 2023, pp. 1–5, 2023, doi: 10.1109/PCEMS58491.2023.10136073.
20. S. C. Pravin, S. P. K. Sabapathy, S. Selvakumar, S. Jayaraman, and S. V. Subramani, "An Efficient DenseNet for Diabetic Retinopathy Screening," International Journal of Engineering and Technology Innovation, vol. 13, no. 2, pp. 125 136, 2023, doi: 10.46604/IJETI.2023.10045.
21. S. Akhtar and S. Aftab, "A Framework for Diabetic Retinopathy Detection using Transfer Learning and Data Fusion," Int. J. Inf. Technol. Comput. Sci., vol. 16, no. 6, pp. 61–73, 2024, doi: 10.5815/ijitcs.2024.06.05.
22. F. Ahmed, "Addressing High Class Imbalance in Multi-Class Diabetic Retinopathy Severity Grading with Augmentation and Transfer Learning," 2025, [Online]. Available: http://arxiv.org/abs/2507.17121
23. D. M. Swapna, T. Ravirala, and N. Reddy. Diabetic Retinopathy Detection Using Federated Learning And Vision Transformers. In Proceedings of the International Journal of Interpretations in Enigmatic Engineering Conference (IJIEE 2025), vol. 02(01), pp. 10–21, 2025. doi: 10.62674/ijiee.2025.v2i01.002.
24. Z. Zhou, H. Yu, J. Zhao, X. Wang, Q. Wu, and C. Dai, "Automatic diagnosis of diabetic retinopathy using vision transformer based on wide-field optical coherence tomography angiography," J. Innov. Opt. Health Sci., vol. 17, no. 2, pp. 1–10, 2024, doi: 10.1142/S1793545823500190.
25. C. Liu, W. Wang, J. Lian, and W. Jiao, "Lesion classification and diabetic retinopathy grading by integrating softmax and pooling operators into vision transformer," Front. Public Heal., vol. 12, no. 2, 2024, doi: 10.3389/fpubh.2024.1442114.
26. V. Awasthi et al., "ViT-HHO: Optimized vision transformer for diabetic retinopathy detection using Harris Hawk optimization," MethodsX, vol. 13, no. August, p. 103018, 2024, doi: 10.1016/j.mex.2024.103018.
27. C. Mohanty, S. Mahapatra, B. Acharya, F. Kokkoras, V. C. Gerogiannis, I. Karamitsos, and A. Kanavos, "Using Deep Learning Architectures for Detection and Classification of Diabetic Retinopathy," Sensors, vol. 23, no. 12, p. 5726, 2023, doi: 10.3390/s23125726.
28. C. Suedumrong et al. Diabetic Retinopathy Detection Using CNN with Background Removal and Data Augmentation. In Proceedings of the Applied Sciences Conference 2024. doi: 10.3390/app14198823.
29. S. A. Karthik et al. Early Detection and Severity Classification of Diabetic Retinopathy Using CNN. In Proceedings of the SN Computer Science Symposium 2025, Article 819. doi: 10.1007/s42979-025-02884-0.
30. Feng, M., Cai, Y. & Yan, S., "Enhanced ResNet50 for Diabetic Retinopathy Classification: External Attention and Modified Residual Branch," Mathematics, vol. 13, no. 10, p. 1557, May 2025, doi: 10.3390/math13101557.