



HybridViT: A Dual-Path Convolutional Neural Network-Transformer Network for Five-Grade Diabetic Retinopathy Grading from Retinal Fundus Images

Huda Ahmed Albeiruti 

College of Computer Engineering, University of Technology – Iraq, Baghdad 10066, Iraq

Corresponding Author Email: Huda.A.albeiruti@uotechnology.edu.iq

Copyright: ©2026 The author. This article is published by IETA and is licensed under the CC BY 4.0 license (<http://creativecommons.org/licenses/by/4.0/>).

<https://doi.org/10.18280/isi.310924>

ABSTRACT

Received: 13 April 2026

Revised: 19 August 2026

Accepted: 1 September 2026

Available online: 30 September 2026

Keywords:

diabetic retinopathy, HybridViT, convolutional neural network-Transformer network, retinal fundus images, deep learning, medical image analysis

Diabetic retinopathy (DR) grading from retinal fundus images remains challenging due to the complex appearance of lesions, class imbalance, and the difficulty of capturing both local pathological patterns and global retinal structures. This study proposes HybridViT, a dual-path convolutional neural network (CNN)-Transformer model for five-grade DR classification. The proposed architecture integrates a ResNet50-based convolutional branch for local feature extraction with a Transformer branch for contextual representation learning, while an auxiliary CNN classifier provides complementary supervision during training. A late-fusion strategy is adopted to combine local and global features for final prediction. The model was evaluated on the APTOS 2019 retinal fundus image dataset containing five DR severity grades: No DR, Mild, Moderate, Severe, and Proliferative DR. To address class imbalance, inverse-frequency weighted sampling and focal loss were used during training, while test-time augmentation was employed during evaluation. HybridViT achieved an overall accuracy of 92.49% and weighted precision, recall, and F1-score values of 92.88%, 92.49%, and 92.50%, respectively. The model obtained particularly strong classification performance for No DR and Moderate DR categories, with F1-scores of 98.08% and 92.27%. These results indicate that the proposed CNN-Transformer design can effectively classify DR severity levels on the evaluated dataset. Further validation using independent datasets from different populations and imaging environments is required to assess its general applicability.

1. INTRODUCTION

The computer-aided diagnosis systems are progressing rapidly, since combining state-of-the-art image methods with algorithms. In particular, the application of deep learning methods leads to more automated disease diagnosis and categorization which then simplifies and saves the review time by doctors [1].

The number of people affected by diabetic retinopathy (DR) is expected to increase worldwide. DR remains a serious neurovascular complication of diabetes and a major cause of vision loss among working-age adults [2, 3].

DR presents considerable and increasing public health challenges. Although early detection is of considerable value, conventional screening involves significant clinical workload and inter-observer variability. Reducing clinical workload and inter-observer variability remains a key motivation for further development of AI-enabled automated screening for DR [4].

Despite the growth of AI-based DR systems, the issues of model generalizability, computational efficiency, and interpretability still persist and hold the systems back from use in health care environments. These constraints motivate the development of more effective DR classification techniques. In medical computer vision, convolutional neural networks (CNNs) are widely used for image classification. CNNs

efficiently develop a hierarchy-based representation that incorporates both high-level semantic characteristics and fine-grained texturing of the image, making image categorization easier. When high diagnostic accuracy is paramount, CNNs are known to perform well across a variety of classification tasks in medical imaging, with increasing importance per task utilization [5].

However, CNNs have limitations despite their widespread use. CNNs have structural limitations that hinder their ability to model global spatial dependencies compared with attention-based models. This problem is particularly severe in the context of DR classification, as the major pathological features are extremely small, indistinct and dispersed within the retina [6]. In light of these drawbacks, a new approach for computer vision has been proposed by the Vision Transformer (ViT), which can be a powerful rival of CNNs in image classification. Unlike conventional CNNs with convolutional layers, which confine the receptive field locally, ViTs employ self-attention to capture global contextual relations between distant pixels. This global modeling ability is particularly useful for DR grading since features may be present in distant pixels of the retina. An effective diagnosis tool, therefore, should be able to zoom in to detect tiny lesions as well as zoom out to examine the overarching retinal structure. However, applying standard ViTs to medical image analysis introduces clear computational

bottlenecks. The classic self-attention mechanism requires substantial memory and computational resources, primarily because its computational complexity scales quadratically with the token sequence length. This quadratic growth creates a severe bottleneck when handling high-resolution clinical scans. Thus, designing a practical DR diagnostic system requires a careful design compromise in which local feature extraction and global context merging are successfully integrated within the constraints of real hardware limits [7].

To mitigate these architectural bottlenecks, this work introduces HybridViT, a dual-pathway CNN-Transformer framework designed for five-class DR classification. The pipeline consists of a ResNet50 backbone for encoding the local features, a Transformer branch for modeling the long-range contextual information and an auxiliary convolutional pathway to stabilize and improve the categorization. Retinal features are extracted locally, whereas context is represented globally.

The Transformer branch includes a window-based attention formulation with global tokens as an architectural option. Under the current token sequence length used in this study, the attention connectivity is equivalent to full attention. Therefore, the present experiments evaluate the overall dual-path HybridViT framework.

The contribution of this study does not arise from the individual use of CNN or Transformer components alone, since such components have already been explored in medical image analysis. Instead, the proposed contribution lies in the task-specific integration of these components within a unified HybridViT framework for five-class DR grading. ResNet50 is used to preserve local retinal features, while the Transformer branch models broader contextual relationships across the extracted feature sequence. An auxiliary CNN classification pathway is retained during training to provide complementary supervision, and the CNN and Transformer features are combined during inference. Therefore, the proposed contribution is centered on the integrated dual-path design, imbalance-aware training, and five-class DR evaluation.

To address class imbalance, the optimization protocol included an inverse-frequency weighted sampling process along with a focal loss-based objective function. Test-time augmentation was also employed during inference to improve empirical prediction stability. Synthesizing these methodologies yields a diagnostic model engineered to partition fundus photographs into five distinct strata of DR severity, spanning from No DR to Proliferative DR. Consequently, the main contributions of this study are summarized below:

- Task-specific CNN-Transformer integration:** The proposed HybridViT integrates local feature extracted by the ResNet50 network with the global context modeled by the Transformer network for five-class DR grading. It effectively synergizes the localized representations of retinal images with broader contextual information within a single classification framework.

- Complementary dual-path supervision:** An auxiliary CNN classification pathway operates alongside the Transformer branch during training and provides an additional supervisory signal to the convolutional backbone. During inference, the CNN feature vector and Transformer [CLS] representation are concatenated and passed through the MLP to obtain the final classification.

- Imbalance-aware training strategy:** Inverse-frequency weighted sampling and focal loss were used during training to

reduce the effect of class imbalance, while test-time augmentation was applied during evaluation.

- Five-class DR severity classification:** The framework was evaluated across the full five-class DR grading scheme, covering No DR, Mild, Moderate, Severe, and Proliferative DR, which allows its performance and error patterns to be examined across different levels of disease severity.

- Error-oriented evaluation:** In addition to overall classification performance, confusion-matrix analysis was used to examine the distribution of errors across the five DR grades, with particular attention to confusion between neighboring severity levels.

2. RELATED WORK

Recently, ViTs have been increasingly applied to screening for DR using fundus photography [8]. While CNNs have long been prevalent in this field, recent comparative findings suggest that ViTs can surpass conventional convolutional networks when identifying referable DR cases. Inspired by this progress, there is growing research interest in applying Transformer models to the analysis of retinal imagery [9].

Following this trend, Alam et al. [10] proposed a multi-class formulation to assess the severity of DR using the APTOS 2019 benchmark dataset. Alam et al. [10] introduced self-supervised feature representations generated by a contrastive-learning scheme, namely SwAV, to examine retinal fundus images. For feature extraction, the authors utilized the ResNet50-based architecture that uses SwAV to produce a prototype-based representation. SwAV allows the model to maintain a stable representation across augmented views and do away with the need to explicitly calculate pairwise similarity. The classifiers explored by the authors attained a classification accuracy of 87.0% with precision, sensitivity, and F1-score values of 86.0%, 87.0%, and 86.0%, respectively, and an ROC area of about 83.2%.

To address the reliance on labeled data, Ren et al. [11] proposed a render-guided self-supervised DR-MAE framework for fundus image feature learning. This method alleviates problems related to annotation quality and quantity and is capable of extracting local and global features. An anatomy-aware masking method and a hybrid dimensional convolution block help preserve subtle fundus structures and anatomical features and enhance multi-scale feature extraction. On APTOS 2019, the method obtained 91.30% accuracy, 89.19% precision, 85.64% F1-score, and an AUC of 98.42%; on DDR, it achieved 86.64% accuracy, 70.61% precision, 67.93% F1-score, and an AUC of 97.13%.

Xue et al. [12] proposed an alternative line of work, which was evaluated on the APTOS 2019 dataset and proposed VMamba-m, an improved visual state-space model for DR grading. The model includes architectural refinements such as dual-branch VSS, SE channel attention, local attention, and focal loss. It was trained in PyTorch using an RTX 3080 Ti GPU, with a batch size of 32 for 150 epochs. The proposed VMamba-m achieved 79.1% accuracy, an AUC of 0.891, and a training time of 34.027 seconds per epoch. The baseline VMamba achieved 71.4% accuracy with a training time of 76.322 seconds per epoch. According to the authors' ablation analysis, the complete VMamba-m configuration improved accuracy by 12.3% and reduced the iteration time by 55.4% relative to the baseline VMamba.

From a broader comparative perspective, Zhang et al. [13]

examined five-class DR grading with APTOS 2019 as the main dataset and EyePACS as the reference dataset. They compared CNNs with ViT-Small, ViTs and SwinV2-Tiny and hybrid schemes such as LeViT-256 and CvT-13. Models were adapted with transfer learning approaches only in the fine-tuning phase and evaluated through common quantitative metrics. Grad-CAM and Attention Rollout were used to visualize attention over clinically relevant retinal regions and gain insights into model behavior. On APTOS-2019, the hybrid CvT-13 reached a QWK of 0.84, an AUC of 0.93, and an accuracy of 0.74, while SwinV2-Tiny achieved QWK: 0.80, ACC: 0.71. The authors also highlighted some limitations, especially with regard to variable performance by severity group and less accurate lesion localization in traditional ViT architectures.

Finally, Liu et al. [14] proposed a STMF-DRNet, a Swin-TransformerV2-based DR grading model. In order to extract global, local, and fine-grained lesion features from the retinal fundus image, a multi-branch structure was adopted in this framework. By integrating CBAM attention, lesion-region refinement modules, and category-attention fusion, the model

aims to improve fine-grained stage classification. Reported results show classification accuracies of 87.9% on DDR, 86.3% on EyePACS, 85.5% on APTOS-2019, and 77.0% on clinical datasets. In addition, it achieved a specificity of 94.2% and a Kappa value of 0.877, indicating strong diagnostic performance across different evaluation datasets.

In general, the past work demonstrated the efficacy of Transformer-based, self-supervised, state-space and hybrid architectures for DR grading. However, these studies vary in model architecture, training scheme, datasets and experimental settings. The current work contributes to this direction by developing a task-aware dual-path CNN-Transformer framework that combines ResNet50-based local feature learning, Transformer-based global context modeling, auxiliary CNN-based supervision, an imbalance-aware training scheme, and five-class DR severity evaluation in a single HybridViT model.

To provide a more direct comparison with previous studies, Table 1 summarizes the main datasets, model structures, classification tasks, reported results, and differences from the proposed method.

Table 1. Comparison of the proposed HybridViT with previous diabetic retinopathy (DR) classification studies

Study / Source	Comparison Details
Alam et al. [10]	Dataset: APTOS 2019 Model: SwAV with ResNet50 Task: Multi-class DR grading Results: Accuracy 87.0%; F1-score 86.0% Difference: Self-supervised representation; no dual-path CNN-Transformer classification design.
Ren et al. [11]	Dataset: APTOS 2019 and DDR Model: DR-MAE with anatomy-aware masking Task: DR grading Results: APTOS: accuracy 91.30%; F1-score 85.64%; area under the curve (AUC) 98.42% Difference: Masked-autoencoder self-supervised learning rather than dual-path CNN-Transformer classification.
Xue et al. [12]	Dataset: APTOS 2019 Model: VMamba-m with dual-branch VSS, SE channel attention, local window attention, and focal loss Task: DR grading Results: Accuracy 79.1%; AUC 0.891; training time 34.027 s/epoch (as currently reported in Related Work) Difference: Visual state-space architecture rather than the proposed ResNet50-Transformer dual-path design.
Zhang et al. [13]	Dataset: APTOS 2019 and EyePACS Model: CNN, ViT, and hybrid models, including CvT-13 Task: Five-class DR grading Results: CvT-13 on APTOS: accuracy 0.74; AUC 0.93; Quadratic Weighted Kappa (QWK) 0.84 Difference: Comparative and interpretability-oriented study using several architectures.
Liu et al. [14]	Dataset: DDR, EyePACS, APTOS 2019, and clinical data Model: STMF-DRNet with Swin-TransformerV2 and multi-branch feature extraction Task: Fine-grained DR grading Results: APTOS accuracy 85.5%; reported Kappa 0.877 Difference: Multi-branch Swin-based architecture evaluated on multiple datasets.
Proposed HybridViT	Dataset: APTOS 2019 Model: ResNet50 + Transformer + auxiliary CNN pathway Task: Five-class DR grading Results: Accuracy 92.49%; weighted F1-score 92.50% Difference: Dual-path CNN-Transformer integration with inverse-frequency weighted sampling and focal loss.

Note: CNN = convolutional neural network, ViT = Vision Transformer.

The comparison reveals that the previously published works employed various learning strategies and model architectures for DR grading: self-supervised learning, masked autoencoders, visual state-space models, and Swin-based approaches. The proposed design uses a ResNet50 backbone followed by Transformer blocks, with an additional CNN classification pathway. Among the studies reported on the APTOS 2019 dataset in the comparison table, only a few achieved classification accuracy close to that of the proposed approach. However, direct numerical comparison should be

interpreted carefully because the studies used different training settings, validation procedures, and model configurations.

3. PROPOSED METHODOLOGY

To identify DR, this study proposed a HybridViT architecture that combines the complementary advantages of convolutional and Transformer-based models. To capture detailed local information from retinal fundus images, the

framework incorporated a pre-trained ResNet-50 backbone. This architecture ensures that fine-grained structures, including microaneurysms and small lesions, are accurately preserved. These feature maps are then fed into a stack of Transformer blocks equipped with a window-based attention formulation with global tokens (sliding-window attention with global tokens; window size $w = 128$, global tokens $g = 16$).

The first and last Transformer layers use full attention, while the intermediate layers include the same window-based attention formulation. In the current configuration, the ResNet50 feature map produces 49 spatial tokens, and one [CLS] token is added, giving $N_{tok} = 50$. Since $w = 128$ is larger than N_{tok} , the windowed attention mask covers the complete token sequence and is equivalent to full attention for the present input resolution.

This design provides a structured way to model contextual relationships across the extracted feature sequence [15].

The design used a dual-path structure: the main Transformer branch modeled global retinal context, whereas the secondary CNN-based branch refined local discriminative patterns. Late fusion was performed by concatenating the Transformer [CLS] embedding and the CNN global-average-pooled vector, and then a small multilayer perceptron was applied to obtain the final logits. By combining these two complementary

views, the model integrates local and contextual representations that are useful for DR classification. Therefore, the dual-path design provides complementary local and global information for DR classification within the experimental setting used in this study.

To improve model generalization, a data augmentation scheme was applied, including traditional data augmentation during training and test-time augmentation. A weighted sampling strategy based on inverse class frequency was used during training. The sampling weight of each sample was inversely proportional to the number of training samples in its class. Therefore, samples from less frequent classes had a higher probability of being selected during training. Focal loss was used as the classification loss to give more attention to difficult samples during training.

CNNs excel at local retinal feature extraction, but Transformers are more suited for encoding contextual relations. The HybridViT architecture proposed in this work combines these heterogeneous representations via late fusion to facilitate automated DR grading; however, extensive validation is still needed to assess its performance on multiple clinical sites and specific imaging devices [16]. The overall information-processing workflow of the proposed HybridViT framework is illustrated in Figure 1.

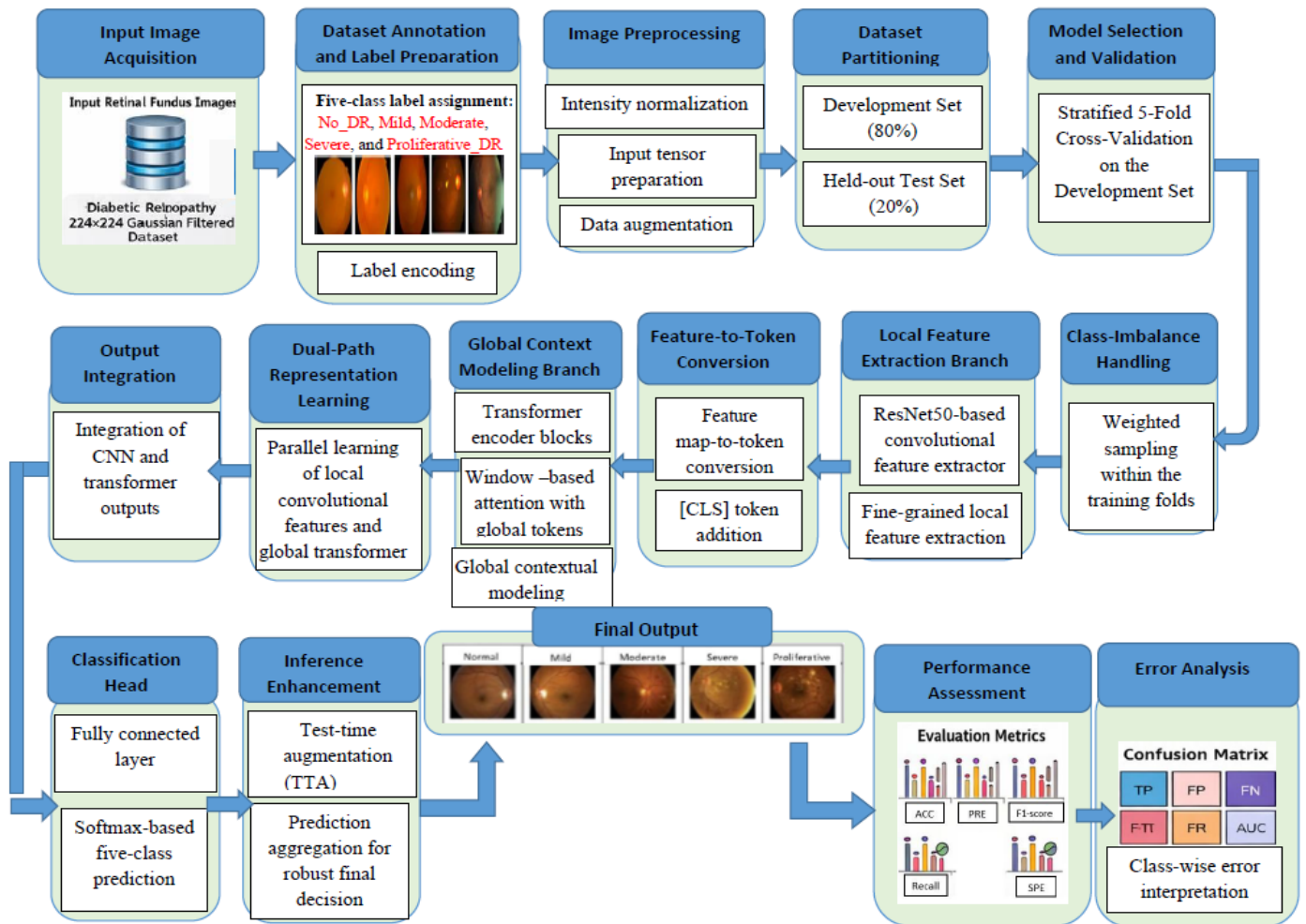


Figure 1. Overview of the proposed diabetic retinopathy (DR) information-processing workflow

3.1 Diabetic retinopathy dataset

The current work used the public dataset "Diabetic Retinopathy 224 × 224 Gaussian Filtered" produced from the

APTOS 2019 Blindness Detection competition available on Kaggle [17, 18]. It contains a total of 3,662 retinal fundus images that were Gaussian-filtered and resized to a standardized dimension of 224 × 224 pixels with the aim of

reducing noise and ensuring compatibility with common transfer-learning pipelines based on pre-trained architectures. One of five DR severity levels—No DR (1,805 photos, 49.29%), Mild (370 images, 10.10%), Moderate (999 images, 27.28%), Severe (193 images, 5.27%), and Proliferative DR (295 images, 8.06%)—is indicated on every image.

The dataset provides Gaussian-filtered retinal fundus images resized to 224×224 pixels. This standardized image size is compatible with common deep-learning architectures used in DR classification studies, including CNN-based and Transformer-based models. The images are organized according to five DR severity levels, which supports their use in supervised multi-class classification experiments [19, 20].

To reduce overfitting risk, the training pipeline used data augmentation, inverse-frequency weighted sampling, and focal loss. Data augmentation was applied only to the training subset after dataset splitting. Weighted sampling and focal loss were used to reduce the dominance of majority classes during training.

The workflow starts with retinal fundus image input, followed by image preprocessing and HybridViT-based analysis. The model produces one of the five DR severity grades as the analysis output. The workflow represents the experimental image-analysis pipeline used in this study. Practical integration with clinical information systems was not evaluated.

The class distribution shown in Figure 2 illustrates the imbalance in the DR dataset, a common characteristic of medical imaging datasets [21, 22].

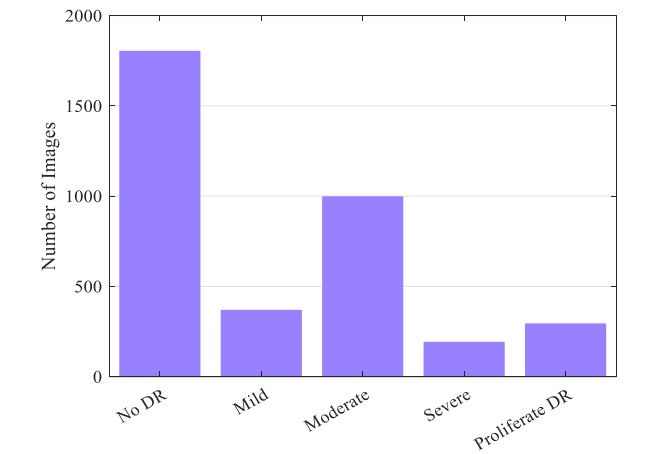


Figure 2. Diabetic retinopathy (DR) class distribution analysis

3.2 Dataset preprocessing pipeline

Figure 3 shows the RGB normalization constants used in preprocessing: mean = [0.485, 0.456, 0.406] and std = [0.229, 0.224, 0.225]. These values match the standard ImageNet normalization and were applied to align the input scale with common pre-trained backbones used in transfer learning.

Since fundus images contain color and contrast signals that are clinically significant, we kept the RGB channels. Whereas the red and blue channels give complementary information, the green channel frequently offers a better contrast between vessels and lesions. Because of this, we maintained each of the three channels and used uniform normalization throughout.

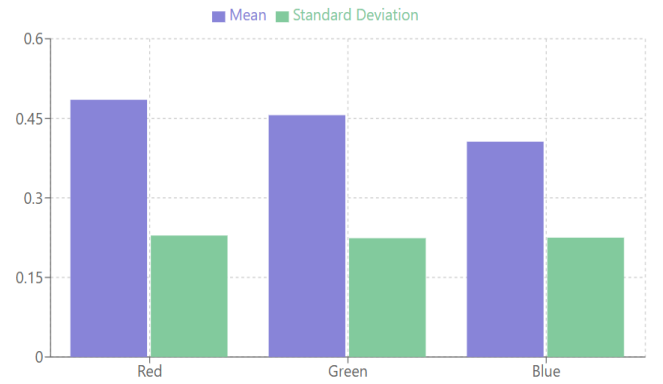


Figure 3. RGB normalization constants used for image preprocessing

The dataset contains 224×224 -pixel RGB fundus images (8-bit per channel). This consistent resolution reduces computational cost while retaining sufficient anatomical information for DR grading and comparison across different model designs.

The dataset consisted of a 20% held-out test set (Test) and an 80% development set (Dev). $K = 5$ stratified five-fold cross-validation was used throughout all model selection, hyperparameter tuning, and calibration on the development set only. The test set was used once for the final report and was not modified. Furthermore, to enhance the numerical stability of optimization and regularize input scale consistently in the three RGB channels, we utilized channel-wise normalization. This stage enables pre-trained feature extractors to work in the input range that was expected during their original training, while mitigating sensitivity to changes in light levels and device-dependent intensity differences.

Especially in transfer learning contexts, normalization is critical since a widely deployed backbone network has been pre-trained on large natural image datasets with a specified input distribution. The same normalization approach increases training stability, maintains compatibility with such pre-trained representations, and allows for fair comparisons of different model architectures.

The behavior of pixel intensities before and after normalization can be seen in Figure 4. Next, the normalization phase transforms input values and moves them closer to a uniform range to enable smoother optimization during training and provide more stable gradient updates.

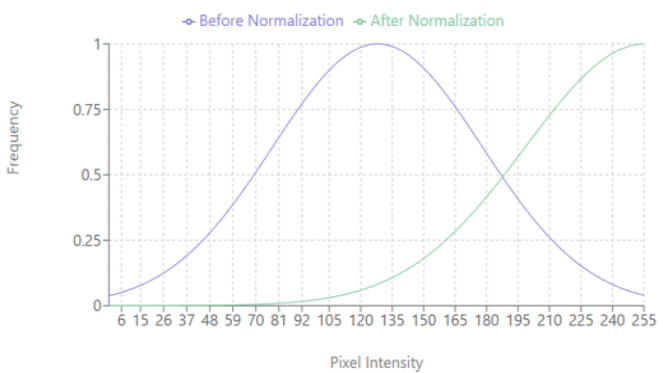


Figure 4. Intensity distribution before and after normalization

The normalization step standardizes the input intensity distribution and helps align the images with the input scale

expected by ImageNet-pretrained backbones.

The training-exclusive augmentation strategy utilized to enhance generalization is summarized in Figure 5. The pipeline is dominated by geometric transformations; Random Horizontal Flip and Random Vertical Flip were used with a probability of 0.5, which promotes invariance to orientation variations in fundus images.

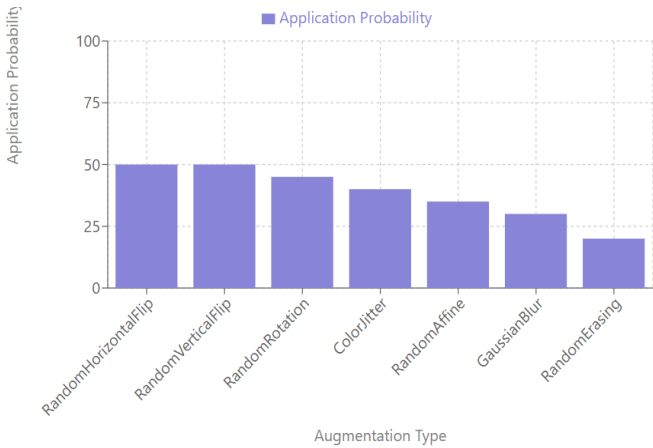


Figure 5. Data augmentation distribution

Following the stratification split, only training-only augmentations were used for training; Following the stratified split, training-time augmentations were applied only to the training subset. Validation and test images received deterministic preprocessing, with test-time augmentation applied separately during inference. As shown in Figure 5, each transform was applied with a specified probability. Specifically, rotation, color jitter, affine transformations, Gaussian blur, and random erasing were performed with their respective probabilities, and horizontal flip was used at a probability of $p = 0.5$.

Random rotation, color jitter, affine transformation, Gaussian blur, and random erasing were used as training-time perturbations. These augmentations were applied only to the training subset and should not be interpreted as external validation of acquisition robustness.

To approximate a slight loss of clarity and lower sensor noise, conservative Gaussian blur was used with a probability of $p = 0.30$, and to simulate partial occlusion while maintaining a modest perturbation and label preservation, random erasing was used with a probability of $p = 0.20$.

Similarly, with a modest erased region (2–6% of the image) and an aspect ratio between 0.3 and 3.3, random erasing ($p = 0.20$) was used. To avoid introducing additional spurious color cues, the erased region was filled with the per-channel mean. In order to increase reproducibility, a constant seed (42 in all experiments) was used for the random number generator. The basic statistics of the retinal images are shown in Table 2. The dataset contains 3,662 images and was divided into a development set of 2,930 images and a held-out test set of 732 images, corresponding approximately to an 80/20 split. Model selection and validation were conducted exclusively on the development set using stratified 5-fold cross-validation. The held-out test set was not used during training or hyperparameter tuning and was reserved strictly for final evaluation. After the model configuration was finalized, performance was assessed once on the held-out test set (732 images), and these results were reported as the final test performance. Table 3 summarizes the dataset split and the

number of classes.

Table 4 presents the class distribution of the full dataset and the held-out test set used for the final evaluation.

Because the public APTOS 2019 dataset used in this study does not provide explicit patient identifiers in the available metadata, strict patient-level splitting could not be verified. Therefore, the dataset was divided using a stratified image-level split to preserve the distribution of the five DR severity classes across the training, validation, and testing subsets. To reduce the risk of data leakage, we ensured that each image identifier appeared in only one subset. Data augmentation was applied only after the split and only to the training subset. The preprocessing parameters applied to the retinal fundus images are summarized in Table 5.

Table 2. Dataset image properties

Parameter	Value
Image resolution	224 × 224 pixels
Color channels	3 (RGB)
File format	JPG/PNG
Bit depth	8-bit per channel

Table 3. Dataset statistics

Parameter	Value
Total images	3,662
Development set size	2,930 (≈ 80%)
Held-out test set size	732 (≈ 20%)
Number of classes	5

Table 4. Class distribution of the full dataset and held-out test set

Parameter	Value	Testing Samples
No diabetic retinopathy (DR)	1,805 (49.29%)	361
Mild	370 (10.10%)	58
Moderate	999 (27.28%)	211
Severe	193 (5.27%)	49
Proliferative DR	295 (8.06%)	53
Total	3662	732

Note: Percentages are rounded to two decimal places; therefore, totals may not sum to exactly 100%.

Table 5. Preprocessing parameters

Parameter	Value
Normalization mean	[0.485, 0.456, 0.406]
Normalization standard deviation	[0.229, 0.224, 0.225]
Aspect ratio	1:1 (Square)

Class imbalance was handled using a consistent training and evaluation protocol. We used a stratified split (80% development, 20% held-out test) and performed stratified 5-fold cross-validation on the development set for model selection.

The held-out test set was not used for tuning. All hyperparameters and decision thresholds were selected using validation folds only. Inverse-frequency weighted sampling and focal loss were used during training, and the random seed was fixed at 42.

The class distribution also suggests an inherent problem of most medical image datasets: an imbalance within the five grades of severity. No DR accounts for 49.29% with 1,805 images; Moderate DR, 27.28% with 999 images; Mild

accounts for 10.10% or 370 images; Proliferative DR constitutes 8.06%, or 295 images, while Severe cases are 5.27% or 193 images. Although the dataset was consistent with typical prevalence patterns, its class imbalance during training necessitated the use of imbalance-handling techniques.

3.3 Proposed CNN-enabled Vision Transformer model

The overall workflow is presented in Figure 1, while the detailed data flow of HybridViT is described in this section. The preprocessed fundus image passes through the ResNet50 backbone and feature adaptation layer, followed by the CNN and Transformer pathways. Next, the CNN feature vector and Transformer [CLS] representation are concatenated and passed through the MLP to determine the final DR grade. We constructed a HybridViT model for grading DR, which integrates a ViT with a CNN. We adopted a ResNet50 backbone to extract local features, while the Transformer branch models broader contextual relationships.

The ResNet50 backbone was pre-trained on ImageNet and then fine-tuned on retinal fundus images, adapting the features from generic objects to retinal-specific patterns. In our configuration, ResNet50's last convolutional layer outputs a 2,048-channel feature map of size 7×7 . We then add an adaptive 1×1 convolution layer from 2,048 channels to 512 channels to prepare these features for processing by the transformer. The $7 \times 7 \times 512$ feature map is organized into a sequence of tokens (since there are 49 spatial locations, each corresponding to a token with an embedding size of 512, this yields 49 tokens). To this sequence, we append a learnable classification token (CLS), forming an overall token sequence which has a final length of 50 tokens (49 spatial tokens + CLS).

The resulting token sequence goes through a series of six Transformer blocks. Each block is made up of multi-head self-attention with eight attention heads and is succeeded by a feed-forward network (FFN) with a dimensionality expansion ratio of 4.0. This involves the use of residual connectivity and layer normalization in order to make training stable. For regularization of the model, a dropout value of 0.2 is employed.

The dual-path layout of HybridViT is a crucial component of its design. An additional CNN classifier works directly on the convolutional features in addition to the Transformer branch. During training, the auxiliary CNN loss is weighted by 0.4 in the training objective. During inference, the CNN and Transformer features are concatenated and passed through the MLP to obtain the final class prediction.

Within the Transformer blocks, a window-based attention formulation with global tokens was used. The configuration used a local window size of $W = 128$ and $g = 16$ global tokens. Because the current token sequence contains 50 tokens, the attention connectivity is equivalent to full attention.

The local window essentially encompasses the whole token sequence for the present tokenization length ($N_{\text{tok}} = 50$). Consequently, with such a specific setup, the attention pattern operates similarly to dense attention. However, the window-based formulation was retained as a design option for future settings with higher-resolution inputs and longer token sequences. Table 6 summarizes the main parameters of the convolutional backbone. The corresponding Transformer configuration and training parameters are reported in Tables 7 and 8, respectively.

Table 6. Convolutional neural network (CNN) backbone parameters

Parameter	Value
Base model	ResNet50
Pretrained	Yes (ImageNet)
Feature output dimension	2,048
Adaptive convolution output	512

Table 7. Vision Transformer (ViT) parameters

Parameter	Value
Embedding dimension	512
Number of Transformer blocks	6
Number of attention heads	8
Multilayer perceptron ratio	4.0
Dropout rate	0.2
Feature size	7×7
Number of patches	49
Token sequence length (N_{tok})	50 (49 patches + 1 [CLS])
Attention connectivity	Window-based formulation with $w = 128$ and $g = 16$; equivalent to full attention under $N_{\text{tok}} = 50$

Table 8. Training parameters

Parameter	Configuration / Value
Batch size	16
Learning rate	$2e^{-4}$ (max)
Weight decay	0.05
Number of epochs	50
Optimizer	AdamW ($\beta_1 = 0.9, \beta_2 = 0.999$)
Auxiliary loss weight	0.4
Total parameters	~47 M
Trainable parameters	~45 M
Memory footprint	~180 MB
Sampling strategy	Inverse-frequency weighted sampling
Loss function	Focal Loss
Learning rate schedule	Cosine Annealing Learning Rate ($Etd_{\min} = 1e^{-6}, T_{\max} = 50$)
Random seed	42
Initialization	ResNet50 initialized with ImageNet-pretrained weights; newly added linear layers initialized using He (Kaiming) normal initialization
Framework	PyTorch 2.5.1 + CUDA 12.1
Hardware (GPU)	NVIDIA GeForce RTX 4090 (24 GB VRAM)
Training time	Approximately 3 h 42 min for 50 epochs

The main architectural components can be summarized as follows:

- Windowed attention with global tokens is used in the Transformer blocks to control attention connectivity; for the 7×7 tokenization used here ($N_{\text{tok}} = 50$), this corresponds to dense attention, while remaining scalable to larger token sequences.
- Multi-head attention with eight heads allows the model to learn complementary feature relationships in parallel.
- Regularization includes a dropout rate of 0.2 together with layer normalization in each Transformer block.
- A multilayer perceptron expansion ratio of 4.0 provides sufficient capacity for non-linear feature transformation in the feed-forward sublayers.
- Skip connections are used throughout the network to

preserve and combine low-level and high-level representations.

Although Transformer-based models may have a high risk of overfitting when trained on limited medical datasets, the proposed HybridViT does not use a pure large-scale ViT trained directly from image patches. Instead, the Transformer branch operates on a compact feature sequence extracted by the ResNet50 backbone. In the present configuration, the CNN feature map produces 49 spatial tokens, and one [CLS] token is added, resulting in a sequence length of 50 tokens. This design reduces the length of the Transformer input sequence and allows the Transformer branch to model contextual relationships over CNN-extracted features rather than over a large number of raw image patches.

For reproducibility, the experiments were conducted using an NVIDIA GeForce RTX 4090 GPU with 24 GB VRAM and PyTorch 2.5.1 with CUDA 12.1. The random seed was fixed at 42. The ResNet50 backbone was initialized with ImageNet-pretrained weights, while the newly added linear layers were initialized using He (Kaiming) normal initialization. A cosine annealing learning rate schedule was used with $T_{\max} = 50$ and $Eta_{\min} = 1 \times 10^{-6}$. The complete training procedure for 50 epochs required approximately 3 hours and 42 minutes.

3.4 Attention connectivity specification

In the implemented HybridViT model, the ResNet50 backbone produces 49 spatial tokens from the 7×7 feature map, and one [CLS] token is added. Therefore, the total token sequence length is $N_{\text{tok}} = 50$. The Transformer branch includes a window-based attention formulation with global tokens, using a local window size of $w = 128$ and $g = 16$ global tokens.

Because w is larger than N_{tok} in the present configuration, the local window covers the complete token sequence. Therefore, the attention connectivity is dense and equivalent to full self-attention for the current input resolution. The window-based formulation is retained as a scalable architectural option for future settings with longer token sequences.

The current experiments did not test the hypothesis that sparse attention improves classification performance or reduces computational cost. This research did not perform a comparison of FLOPs, peak GPU memory usage, and inference time per image among various scenarios for different settings of the attention window.

4. RESULTS AND DISCUSSION

The proposed HybridViT architecture was evaluated for five-class DR classification on the APTOS 2019 dataset. The model achieved an overall classification accuracy of 92.49%. The highest class-wise F1-scores were obtained for No DR (98.08%) and Moderate DR (92.27%).

The evaluation included class-wise results and confusion matrix analysis. The evaluation included class-wise results and confusion matrix analysis. The following subsections discuss the findings, compare them with previous studies, and examine the strengths and limitations of the model.

The usage of the sparse attention setup has not been assessed independently in the experiments conducted. The setup in the experiments comprised 50 tokens in the token sequence, while the local window size is 128. This indicates that the whole token sequence is covered in the local window,

making the connectivity similar to full attention in this setup. Therefore, the reported results should not be interpreted as evidence that sparse attention improves performance or reduces computational cost.

It should be emphasized that the reported performance represents internal validation on the APTOS 2019 dataset only. However, as this study did not employ an independent external dataset, the results do not guarantee clinical generalization to other populations, imaging devices, or clinical contexts. Figure 6 illustrates the training and validation loss reduction across fifty epochs. The training loss decreased from roughly 0.97 at the start of training to about 0.01 by the end.

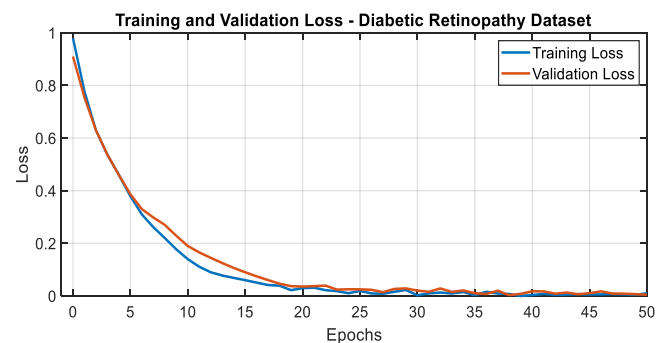


Figure 6. Training and validation loss

The validation loss dropped from 0.90 to almost zero by the final epoch, whereas the training and validation curves were very close. The trends of the two curves are consistent with stable convergence in the internal validation setting.

The largest loss reduction occurred in the first 15 epochs. By approximately epoch 15, the training loss was 0.06 and the validation loss was 0.09. Thereafter, the two curves leveled off, with relatively little divergence.

Figure 7 displays the training and validation accuracy over the course of 50 epochs. The accuracy for both training and validation increased rapidly during the first few epochs and then leveled off during the later epochs.

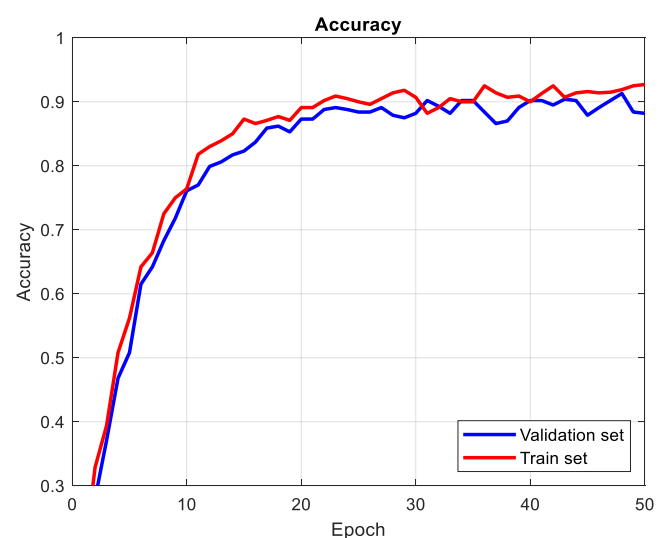


Figure 7. Training and validation accuracy

Since the study used only APTOS 2019, overfitting and dataset-specific learning cannot be fully excluded, so external validation remains necessary.

The accuracy curves increased rapidly during the early epochs and then changed more gradually in the later epochs. The close trend between the training and validation curves supports stable convergence under the internal validation setting.

The small training-validation difference shows that the methodology behaves stably in the internal validation context but does not show the ability to generalize beyond APTOS 2019. The training and validation curves followed the same trend, which indicates stability under the respective internal validation context. Ultimately, the observed stability indicates the compatibility of the training process and the architectural decisions taken in the current work within the internal validation setting.

The confusion matrix offers a class-wise perspective of the performance of HybridViT in classifying across the five grades of DR. For the Mild DR class, the model correctly classified 50 out of 58 samples, corresponding to a recall of 86.21%.

The remaining Mild DR errors were mainly predictions as No DR and Moderate DR, reflecting the difficulty of separating early DR signs from visually related neighboring grades, reflecting the visual similarity between early and intermediate DR stages.

The findings suggest that there is difficulty in distinguishing between early and intermediate stages of DR, and thus they make clinical sense given that features of DR take time to develop. For the Moderate DR class, the model correctly classified 191 out of 211 samples, corresponding to a recall of 90.52%. The main errors occurred with neighboring or visually related severity grades. The Moderate DR class had the largest number of misclassified samples, with 20 errors. Most of these errors involved confusion with neighboring or visually related grades.

Severe DR and Proliferative DR remained more challenging than the majority classes. For the Severe DR class, the model achieved a recall of 83.67%, indicating that most Severe DR samples were correctly classified, although confusion with neighboring advanced stages remained present. This indicates that both class imbalance and visual overlap among advanced-stage cases may contribute to the difficulty of late-stage separation, which is an important limitation under the current training distribution.

The confusion matrix in Figure 8 indicates that the majority of misclassifications were between neighboring DR levels; of the 55 misclassification cases, 40 (72.73%) were one severity grade away from the true class, 14 (25.45%) were two severity grades away from the true class, and one (1.82%) was three severity grades away. Interestingly, there was no direct misclassification between the No DR and Proliferative DR cases. While most of these errors were close to the correct severity level, it must not be assumed that these errors are insignificant in a clinical setting, as their significance depends on the severity grade considered positive for a given clinical decision.

The specific numerical results can be seen in Table 9, which also summarizes the overall performance for different DR categories. The performance on the class No DR is excellent, which can be observed from the high values in the precision, recall and F1-score (0.97, 0.98 and 0.98, respectively). The model therefore showed strong performance in identifying No DR cases. This performance may partly reflect the larger representation of the No DR class in the training set, although this effect was not independently evaluated.

The class-wise results show different performance among the five DR grades. The highest F1-score was obtained for No DR (98.08%), followed by Moderate DR (92.27%). Mild and Severe DR achieved F1-scores of 80.65% and 78.10%, respectively. Proliferative DR showed high precision (95.00%), but its recall was lower (71.70%), indicating that some Proliferative DR samples were predicted as other grades. These results indicate that the minority classes still have lower performance in some measures, and the class imbalance problem was not completely removed. The ablation analysis further examines the effect of weighted sampling and focal loss on the overall model performance. Table 10 compares the proposed model with previous DR grading studies.

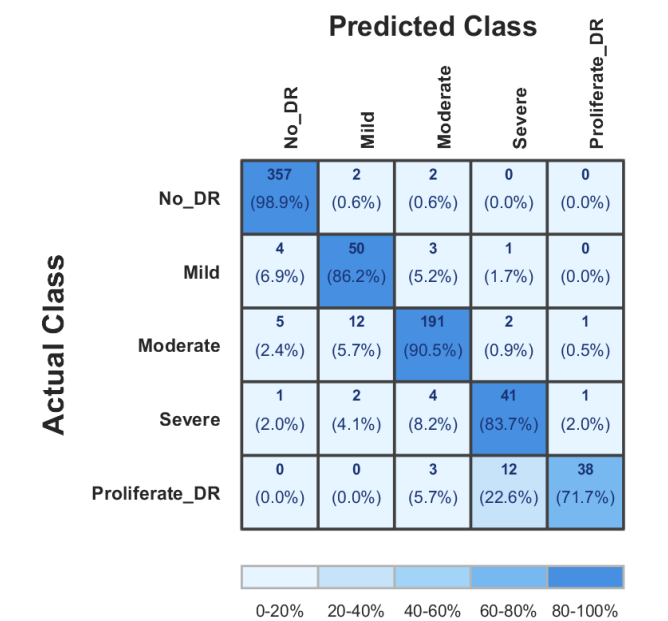


Figure 8. Confusion matrix

Table 9. Class-wise performance of HybridViT

Diabetic Retinopathy (DR) Class	Precision (%)	Recall (%)	F1-Score (%)	Support
No DR	97.28	98.89	98.08	361
Mild	75.76	86.21	80.65	58
Moderate	94.09	90.52	92.27	211
Severe	73.21	83.67	78.10	49
Proliferative DR	95.00	71.70	81.72	53

The proposed HybridViT achieved balanced overall performance on the internal APTOS 2019 test set, with an accuracy of 92.49%, weighted precision of 92.88%, weighted recall of 92.49%, and weighted F1-score of 92.50%. The weighted specificity of 97.65% also indicates a low false-positive tendency under the internal test setting. Cohen's Kappa and Quadratic Weighted Kappa (QWK) were also calculated to measure the agreement between the predicted and actual DR grades. The obtained values were 0.8856 and 0.9559, respectively. QWK is particularly relevant to the five-class DR task because the severity grades are ordinal, and it gives a larger penalty when the predicted grade is farther from the actual grade.

The model achieved better internal test performance for No DR and Moderate DR, while separation between Severe DR

and Proliferative DR remained more difficult. This difficulty may be related to overlapping retinal characteristics and subtle visual differences among advanced stages. The above results demonstrate good internal performance on the APTOS 2019 test set, and external validation is necessary before broad clinical conclusions can be drawn. In order to evaluate the core elements of the proposed HybridViT architecture, a compact ablation study was performed. The compared variants include ResNet50 alone, the Transformer branch alone, HybridViT without auxiliary CNN supervision, HybridViT without focal loss, HybridViT without weighted sampling, and HybridViT without both focal loss and weighted sampling. All variants were evaluated using the same dataset split, preprocessing

pipeline, and evaluation protocol. The purpose of this analysis was to show the relative effect of each component under the same experimental setting, not to perform a separate hyperparameter optimization for each variant. The corresponding ablation results are presented in Table 11.

The ablation results show that the full HybridViT model achieved the best overall performance among the tested variants, with an accuracy of 92.49%, a weighted F1-score of 92.50%, and a QWK of 0.9559. The ResNet50-only and Transformer-only variants achieved lower performance, which indicates that local convolutional features and Transformer-based contextual modeling provide complementary information for DR grading.

Table 10. Performance comparison of diabetic retinopathy (DR) grading models

Study	Model	Dataset (s)	Accuracy	Precision	Recall	F1-Score	Specificity	Comparative Observation
Alam et al. [10]	Self-supervised contrastive clustering built on SwAV with a ResNet50 feature extractor	APTOS 2019	87%	86%	87%	86%	/	Strong balanced performance on APTOS with a fully self-supervised design.
Ren et al. [11]	Anatomy-guided self-supervised masked autoencoder with anatomy-aware masking and hybrid dimensional convolution	APTOS 2019; DDR	91.3%	89.19%	/	85.64%	/	Reports high APTOS classification performance using a self-supervised masked-autoencoder framework.
Xue et al. [12]	Enhanced the visual state-space model with dual-branch VSS, SE attention, local window attention, and focal loss	APTOS 2019	79.1%	76.6%	53.7%	52.7%	/	Improves efficiency, but its classification accuracy is lower than the best recent methods.
Zhang et al. [13]	Comparative study of CNN, ViT, and hybrid models; the best model was CvT-13	Balanced EyePACS + APTOS-2019 test set	74%	74%	74%	74%	/	Supports the view that hybrid designs can be more reliable than plain ViT models for five-class DR grading.
Liu et al. [14]	Swin-TransformerV2-based multi-branch model with CBAM attention, lesion refinement, and category-attention fusion	DDR; EyePACS; APTOS 2019; clinical data	85.5%	/	81%	85.9%	94.2%	Shows good cross-dataset stability and strong fine-grained grading performance.
This study	HybridViT	APTOS 2019	92.49%	92.88%	92.49%	92.5%	97.65%	Competitive hybrid result with strong normal-case and moderate-stage recognition, but advanced-stage separation remains the main challenge.

Note: CNN = convolutional neural network, ViT = Vision Transformer.

*Results are reported as presented in the original studies; some studies used APTOS alone, while Zhang et al. used a balanced combined corpus from EyePACS and APTOS-2019.

Table 11. Ablation analysis of the proposed HybridViT framework

Model Variant	Accuracy	Weighted Precision	Weighted Recall	Weighted F1	QWK
ResNet50 only	88.1	88.4	88.1	88.0	0.865
Transformer branch only	89.4	89.6	89.4	89.3	0.881
HybridViT without auxiliary CNN supervision	91.1	91.3	91.1	91.0	0.932
HybridViT without focal loss only	91.0	91.1	91.0	90.9	0.928
HybridViT without weighted sampling only	90.2	90.4	90.2	90.0	0.907
HybridViT without focal loss and weighted sampling	89.0	89.2	89.0	88.8	0.872
Full HybridViT	92.49	92.88	92.49	92.50	0.9559

Note: CNN = convolutional neural network, ViT = Vision Transformer, QWK = Quadratic Weighted Kappa.

Removing the auxiliary CNN supervision reduced the accuracy from 92.49% to 91.10%, suggesting that the auxiliary pathway provides an additional but moderate contribution to feature learning. Removing focal loss or weighted sampling also reduced the performance, and removing both of them produced a larger decrease. This result supports the role of imbalance-aware training in the five-class DR classification setting.

However, these results should be interpreted as a controlled ablation analysis under the same experimental setting. A separate hyperparameter search for each variant was not performed.

5. CONCLUSIONS

The proposed HybridViT model combines a ResNet50 backbone with Transformer-based contextual processing for five-class DR classification. The model showed strong internal performance on the APTOS 2019 dataset, particularly for the No DR and Moderate DR classes.

An additional aspect of the proposed framework is the auxiliary CNN pathway used to provide additional training supervision. The system incorporates a window-based attention mechanism, which, given the configuration of 50 tokens and a 128-unit window size, acts as full attention, providing a foundation for future experiments with longer sequences. This structural approach allows for optimized feature representation by combining CNN and Transformer capabilities.

The main contribution of the outlined framework is its dual-path arrangement, which combines ResNet50-based convolutional feature extraction and Transformer-driven contextual processing. During training, an extra CNN pathway is implemented to give more supervision. The window-based attention method preserves its possible use for future studies with longer token sequences. However, in the existing 50-token layouts, it performs in a similar way to the full attention method.

A main limitation of the present study is that the model was evaluated using only the APTOS 2019 dataset, which contains 3,662 fundus images. Therefore, the reported performance represents internal validation only and should not be interpreted as evidence of clinical generalization.

Variations in image quality and possible domain differences between imaging devices were not independently evaluated. The study also did not include dedicated interpretability analysis such as Grad-CAM visualization, attention-map analysis, or lesion localization. In addition, integration with hospital information systems, interaction with clinicians, deployment on clinical computing platforms, and edge-device execution were not evaluated in the present study.

In addition, the strict patient-level splitting could not be verified because explicit patient identifiers were not available in the public APTOS 2019 metadata. Although image-level separation was applied and each image identifier was restricted to a single subset, this does not fully replace patient-disjoint validation. Future work should evaluate the proposed model using datasets that provide patient identifiers or external cohorts with confirmed patient-level separation.

The results obtained on APTOS 2019 show that the proposed model has potential for five-class DR classification. Future work will include external validation using independent datasets from different imaging settings. Other

research directions include the integration of retinal images with additional clinical data, investigation of clinician feedback mechanisms, federated learning across different institutions, dedicated interpretability analysis, and development of lightweight versions for resource-limited or edge computing platforms. These directions require separate experimental evaluation in future studies.

REFERENCES

- [1] Yaseen, S.F. (2025). A self-supervised vision transformer with token selection for breast cancer histopathology classification. *International Journal of Intelligent Engineering & Systems*, 18(11): 566-580. <https://doi.org/10.22266/ijies2025.1231.35>
- [2] Teo, Z.L., Tham, Y.C., Yu, M., et al. (2021). Global prevalence of diabetic retinopathy and projection of burden through 2045: Systematic review and meta-analysis. *Ophthalmology*, 128(11): 1580-1591. <https://doi.org/10.1016/j.ophtha.2021.04.027>
- [3] Solomon, S.D., Chew, E., Duh, E.J., et al. (2017). Diabetic retinopathy: A position statement by the American Diabetes Association. *Diabetes Care*, 40(3): 412-418. <https://doi.org/10.2337/dc16-2641>
- [4] Han, K., Wang, Y.H., Chen, H.T., et al. (2023). A survey on vision transformer. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(1): 87-110. <https://doi.org/10.1109/tpami.2022.3152247>
- [5] Yaseen, S.F., Jaleel, A., Al-Araji, A.S., Al-Raweshidy, H.S. (2024). Multi-graded brain tumor classification using YOLOv7. *Iraqi Journal of Computers, Communications, Control and Systems Engineering*, 24(3): 59-76. <https://doi.org/10.33103/uot.ijccce.24.3.5>
- [6] Meng, X.C., Wang, N., Shao, F., Li, S.T. (2022). Vision transformer for pansharpening. *IEEE Transactions on Geoscience and Remote Sensing*, 60: 1-11. <https://doi.org/10.1109/tgrs.2022.3168465>
- [7] Dosovitskiy, A., Beyer, L., Kolesnikov, A., et al. (2020). An image is worth 16×16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*. <https://doi.org/10.48550/arXiv.2010.11929>
- [8] Wu, J.H., Koseoglu, N.D., Jones, C., Liu, T.Y.A. (2023). Vision transformers: The next frontier for deep learning-based ophthalmic image analysis. *Saudi Journal of Ophthalmology*, 37(3): 173-178. https://doi.org/10.4103/sjopt.sjopt_91_23
- [9] Goh, J.H.L., Ang, E., Srinivasan, S., et al. (2024). Comparative analysis of vision transformers and conventional convolutional neural networks in detecting referable diabetic retinopathy. *Ophthalmology Science*, 4(6): 100552. <https://doi.org/10.1016/j.xops.2024.100552>
- [10] Alam, M.N.U., Bahadur, E.H., Masum, A.K.M., Noori, F.M., Uddin, M.Z. (2024). SwAV-driven diagnostics: New perspectives on grading diabetic retinopathy from retinal photography. *Frontiers in Robotics and AI*, 11: 1445565. <https://doi.org/10.3389/frobt.2024.1445565>
- [11] Ren, Y.Y., Shao, D.G., Yi, S.L. (2025). DR-MAE: Self-supervised learning for diabetic retinopathy grading based on masked autoencoder. *Journal of King Saud University Computer and Information Sciences*, 37(8): 217. <https://doi.org/10.1007/s44443-025-00159-3>
- [12] Xue, J., Wu, J.Y., Bian, Y.X., Zhang, S.Y., Du, Q.S.

- (2024). Classification of diabetic retinopathy based on efficient computational modeling. *Applied Sciences*, 14(23): 11327. <https://doi.org/10.3390/app142311327>
- [13] Zhang, W.J., Belcheva, V., Ermakova, T. (2025). Interpretable deep learning for diabetic retinopathy: A comparative study of CNN, ViT, and hybrid architectures. *Computers*, 14(5): 187. <https://doi.org/10.3390/computers14050187>
- [14] Liu, Y.Y., Yao, D.Z., Ma, Y.W., et al. (2025). STMF-DRNet: A multi-branch fine-grained classification model for diabetic retinopathy using Swin-TransformerV2. *Biomedical Signal Processing and Control*, 103: 107352. <https://doi.org/10.1016/j.bspc.2024.107352>
- [15] Nadeem, M.W., Goh, H.G., Hussain, M., Liew, S.Y., Andonovic, I., Khan, M.A. (2022). Deep learning for diabetic retinopathy analysis: A review, research challenges, and future directions. *Sensors*, 22(18): 6780. <https://doi.org/10.3390/s22186780>
- [16] Ennab, M., Mcheick, H. (2024). Enhancing interpretability and accuracy of AI models in healthcare: A comprehensive review on challenges and future directions. *Frontiers in Robotics and AI*, 11: 1444763. <https://doi.org/10.3389/frobt.2024.1444763>
- [17] Rath, S. (2022). Diabetic retinopathy 224×224 Gaussian-filtered. Kaggle Dataset. <https://www.kaggle.com/datasets/sovit Rath/diabetic-retinopathy-224x224-gaussian-filtered>
- [18] Asia Pacific Tele-Ophthalmology Society. (2019). APTOS 2019 blindness detection. Kaggle Competition. <https://www.kaggle.com/competitions/aptos2019-blindness-detection>
- [19] Mohanty, C., Mahapatra, S., Acharya, B., et al. (2023). Using deep learning architectures for detection and classification of diabetic retinopathy. *Sensors*, 23(12): 5726. <https://doi.org/10.3390/s23125726>
- [20] Pamula, U., Pulipati, V., Vijaya Suresh, G., et al. (2025). Optimizing diabetic retinopathy detection with electric fish algorithm and bilinear convolutional networks. *Scientific Reports*, 15: 14215. <https://doi.org/10.1038/s41598-025-99228-w>
- [21] Alyoubi, W.L., Shalash, W.M., Abulkhair, M.F. (2020). Diabetic retinopathy detection through deep learning techniques: A review. *Informatics in Medicine Unlocked*, 20: 100377. <https://doi.org/10.1016/j.imu.2020.100377>
- [22] Nazir, T., Irtaza, A., Javed, A., Malik, H., Hussain, D., Naqvi, R.A. (2020). Retinal image analysis for diabetes-based eye disease detection using deep learning. *Applied Sciences*, 10(18): 6185. <https://doi.org/10.3390/app10186185>