Pertanika J. Sci. & Technol. 34 (5): 3003 - 3024 (2026)

PERTANIKA JOURNALS

SCIENCE & TECHNOLOGY

Journal homepage: http://www.pertanika.upm.edu.my/

A Novel Fusion of Computer Vision Capabilities and Deep Learning Techniques for OCT Retinal Disease Diagnosis

Nagham Salim Mohammed[1]*, Omar S. Almolaa[2], and Omar Ibrahim Alsaif[3]

[1]Department of New and Renewable Energies, College of Science, University of Mosul, Mosul 41002, Nineveh, Iraq
2Department of Computer Science, Faculty of Education, University of Telafer, Telafer 41016, Nineveh, Iraq
3Department of Electronic and Communication Technology Engineering, Mosul Polytechnic College, Northern Technical University, Mosul 41002, Nineveh, Iraq

ABSTRACT

The main objective is to address the classification limitations of traditional models for diagnosing retinal macular diseases, including neovascularisation (CNV), diabetic macular oedema (DME), and age-related macular degeneration (AMD) associated with drusen. An Optical Coherence Tomography (OCT)- based computer vision framework has been proposed to classify these diseases. The models achieved this by addressing the inability of traditional models to learn spatial dependencies and the need for an inductive bias to learn fine retinal texture patterns in Vision Transformer (ViT) models. A dual-model HR50_ViTB/16 was proposed, comprising a full-depth ViTB/16 model for capturing spatial information, a pre-trained ResNet50 model for extracting local, multiscale hierarchical information to further improve performance, and a learnable merging layer to integrate local and global information extracted from both branches. The proposed framework outperformed all traditional models, achieving an overall accuracy of 96.14% and an F1 score of 0.952 for the DRUSEN class, which is the most difficult and has very limited data for classification. Its significance lies in the fact that these results demonstrate the quality of the local and global capabilities of ResNet50 networks and $ViT_{B/16}$ transformers, respectively, which can complement each other to overcome the limitations of each model.

Keywords: CNV, DME, DRUSEN, Dual-Branch, Feature Fusion, Medical Image Analysis, OCT (Optical Coherence Tomography), $ResNet_{50}$, Retinal Coherence Diseases, $Vision\ Transformer_{B/16.}$

ARTICLE INFO

Article history:
Received: 11 August 2026
Accepted: 29 September 2026
Published: 30 October 2026

DOI: https://doi.org/10.47836/pjst.34.5.19

nagham.salim@uomosul.edu.iq (Nagham Salim Mohammed)
omar.almolaa@uotelafer.edu.iq (Omar S. Almolaa)
omar.alsaif@ntu.edu.iq (Omar Ibrahim Alsaif)
* Corresponding author

INTRODUCTION

Blindness or visual impairment is a global problem affecting 295 million people with

e-ISSN: 2231-8526

© Universiti Putra Malaysia Press

moderate-to-severe vision impairment (2020) and 43.3 million people with blindness (GBD 2019 Blindness and Vision Impairment Collaborators & Vision Loss Expert Group of the Global Burden of Disease Study, 2021). The global problem is growing, particularly in low- and middle-income countries (Burton et al., 2021; Vision Loss Expert Group of the Global Burden of Disease Study & the GBD 2019 Blindness and Vision Impairment Collaborators, 2024).

Blindness-causing retinal diseases involve Choroidal Neovascularisation (CNV), Diabetic Macular Edema (DME), and DRUSEN (Ma et al., 2022). CNV is among the clinical forms of AMD in which a choroidal blood vessel lies beneath the retinal pigment epithelium, leading to damage to the photoreceptors (He et al., 2021). There could be other causes of v, such as loss, like diabetic retinopathy, which affects many individuals globally in 2020 (Teo et al., 2021). DRUSEN are the most common early sign of macular degeneration. So, lipoprotein deposits under the retinal pigment epithelium are related to age-related macular degeneration (AMD) (Li et al., 2023; Xu et al., 2021).

The Optical Coherence Tomography (OCT) is an advanced imaging method that may assist in identifying pathological markers, neovascularisation (CNV), diabetic macular oedema (DME), DRUSEN, subretinal membranes, pigmented epithelial detachment, intraretinal fluid, macular thickening, and xanthomas (Akpinar et al., 2024). However, distinguishing certain pathological patterns in OCT images remains challenging, including discriminating between CNV and DME or between early DRUSEN and normal retinal morphology (Burton et al., 2021). The difficulty underscores the need for efficient, automated OCT image classification techniques that provide repeatable, quantifiable diagnostic support (Laouarem et al., 2024).

For the automatic detection of retinal diseases, Computer vision and deep learning techniques have been shown to yield good results for detecting CNV, DME, and DRUSEN using a transfer learning-based CNN such as Inception-V3 (Hassan et al., 2023; Kermany et al., 2018). However, due to the local receptive field, conventional CNNs might be unable to capture the long-range spatial dependence, which could be problematic for recognising spatially dispersed lesions such as DRUSEN and for distinguishing the visually similar appearance of a fluid-filled appearance in CNV and DME (Dutta et al., 2023; Fang et al., 2019; Gencer & Gencer, 2025). ViT overcomes this chamodelling by modelling non-local relationships between different image patches using MHSA, enabling a better understanding of global retinal structures in OCT classification (Dosovitskiy et al., 2021; Yang et al., 2025).

To address this problem, the Vision Transformer (ViT), where the transformer architecture is used, splits the image into non-overlapping patches, and each patch is handled as a time step sequentially with a process called Multi-Head Self-Attention (MHSA) that is used to capture the relationships between distant parts of an image (Dosovitskiy et al., 2021; Mohammed et al., 2026). This feature is crucial in the context of OCT classification,

where multiple retinal areas and multiple layers can contain patterns indicative of disease; it has also been demonstrated in the OCT2017 dataset (Akça et al., 2024; Wang et al., 2023). This is especially relevant for OCT images, where small variations in the texture of the retinal layers and reflectivity may be crucial for diagnosis (Wang et al., 2025).

In this study, a new dual-branch computer vision architecture (HR5_ViTB/16) is proposed for four-class classification of retinal diseases from OCT images. Specifically, the architecture consists of a Vision Transformer (ViT-B/16) branch comprising 12 encoder blocks, 12 attention heads, an embedding dimension of 768, and a patch size of 16, yielding 196 patch tokens, in parallel with a pre-trained ResNet-50 residual feature-extraction network. The features from the ViT (CLS) token and the global average-pooled features from the ResNet-50 network are concatenated and then passed through a learnable linear fusion layer with layer-norm, enabling the joint optimisation of local and global features in an end-to-end manner. The architecture is trained and tested on the OCT2017 dataset, which consists of four classes of retinal diseases.

The novelty of the proposed framework is achieved by aligning representational spaces before concatenation, where the ResNet-50 representation (R^(2048)) is converted via a LayerNorm, GELU, and Dropout projection into R^(768), which is part of the cls_token_out of $ViT_{B/16}$ before sequencing. The result is that the next linear transformation (R^(1536)→R^(768)) learns from the true semantic connections between the two spaces, rather than processing incompatible features using statistical techniques. The use of a full sequence rather than some form of dynamic weighting (cross-attention or gating) is an intentional design decision to keep the whole sequence of representational information of both branches while retaining complementary semantic fullness between the convolutions and the transformers. Sequencing with learned linear projections yields a concise representation, and the degree of alignment between representational spaces before merging primarily determines the success of the merging approach, not the complexity of the merge layer itself.

Branch ResNet-50 its output with dimensions (2048) is dropped to dimension (768) by relying via LayerNorm → GELU → Dropout, which matches the dimensions of ViT CLS before the merge process, and thus both branches are inserted into the merge layer with the same equal representational capacity, thus avoiding the bias in information inclusion by increasing one of the branches, and the merge vector with dimension (1536) preserves the complete information for both branches.

This combination process is a (Linear 1536→768, LayerNorm, GELU, Dropout 0.15) projection of the hat model learned from the best weights of the CNN and ViT feature representations. This is not merely an integration of the two branches or an attention gate that uses only one weight. For DRUSEN, texture information is based on the spatial attention representation of the ViT branch, whereas CNV uses the ResNet branch.

The DRUSEN category is the most challenging, as it requires information about both the local texture (RPE surface details) and the global texture (multi-patch RPE ripple). Using the results of the four architectures, the superiority of the adopted architecture can be demonstrated.

LITERATURE REVIEW

Early computational methods for detecting DRUSEN were based on classic image-processing techniques used in fundus fluorescein angiography (FFA) images. A computer-aided detection (CAD) system for automatic identification and areal size calculation of DRUSEN regions in FFA images was developed by Demir et al. (2018) and served as a morphological baseline for ARMD diagnosis. $HR_{50}_ViT_{B/16}$ is a much more difficult task than HR50_FFA/16, as it does not use colour OCT FFA images but instead extracts DRUSEN from a grayscale cross-sectional OCT image, emphasising the importance of the global attention module provided by the ViT branch.

Through a survey proposed by (Liew & Agaian, 2025), the developments of different kinds of features extractors from manual and deep learning techniques have been studied with respect to accuracy, sensitivity, specificity, and computational efficiency for the detection of various diseases of eyes including glaucoma, diabetic retinopathy, and macular degeneration, as well as, cyberattacks and adversarial attack threats on OCT diagnostic systems are concerned, a more effective method has been proposed.

The proposed algorithm by Rahil et al. (2023) introduced a deep learning-based clustering model that uses CNN techniques to segment various retinal patterns in OCT images. The proposed technique's results are excellent, both quantitatively and qualitatively, compared to other state-of-the-art algorithms on the RETOUCH dataset, with an overall performance improvement of 1.8%.

In the case of OCT-based multi-class classification of retinal diseases, Zhou et al. (2023) proposed using the ViT model, which has achieved high performance in distinguishing normal images from irregular conditions such as retinal vascular macular degeneration (CNV), retinal degeneration, and diabetic macular oedema. Achieve an accuracy rate of 95.76% and a sensitivity of 95.77%, outperforming many traditional CNN models statistically.

Aslan (2023) also confirmed this with results from comparative studies of ViT and ResNet networks for image classification of skin diseases, in which ViT surpassed ResNet in accuracy (68.93% versus approximately 61%), despite its longer training time, thereby increasing its generalizability to other areas of medical imaging. In terms of baseline models, RETFound is a model pre-trained on 1.6 million unlabeled OCT images.

Du et al. (2024) applied their method to improve the classification of disease features, such as subretinal fluid, intraretinal fluid, retinal ganglion cysts, and pigment epithelial

detachment. The results of the application on RET showed similar performance to ResNet-50 in terms of accuracy and area under the ROC curve (0.75–0.80).

Raut et al. (2025) introduced a hybrid approach that uses ViT, Inception-V3, and $ResNet_{50}$. This is because results from applying standard databases have shown that using multiple architectures can improve the classification and identification of eye diseases compared with a single architecture. It is advantageous since it fastens treatments.

METHODOLOGY

Objectives

Hybrid Residual_ViT (HR_{50}_$ViT_{B/16}$), facilitating the associated improvement of global self-attention and hierarchical residual feature representations, as well as, the framework, achieved an F1-score of 0.952 on the DRUSEN data set, reflecting an improvement over the best competing hybrid model, known as Conv-ViT, which achieved an DRUSEN F1-score of 0.78 (Dutta et al., 2023), by resolving the long-standing DRUSEN classification bottleneck, by effectively capturing the spatially diffuse sub-RPE patterns across the macular domain.

Materials and Methods

Dataset OCT2017

The OCT2017 dataset was used for the experiments. This publicly available dataset contains high-resolution grayscale OCT retinal scans, and the images are labelled with four different clinical categories. These categories are choroidal neovascularisation (CNV), diabetic macular oedema (DME), DRUSEN deposits, and NORMAL retinal tissue. Images have been loaded in RGB format to ensure compatibility with the pre-trained ResNet-50 backbone, as this neural network was initialised on a three-channel image dataset, namely ImageNet. The dataset can be found at the following link: https://data.mendeley.com/datasets/rscbjbr9sj/3 (Kermany et al., 2018). Figure 1 illustrates OCT Types.

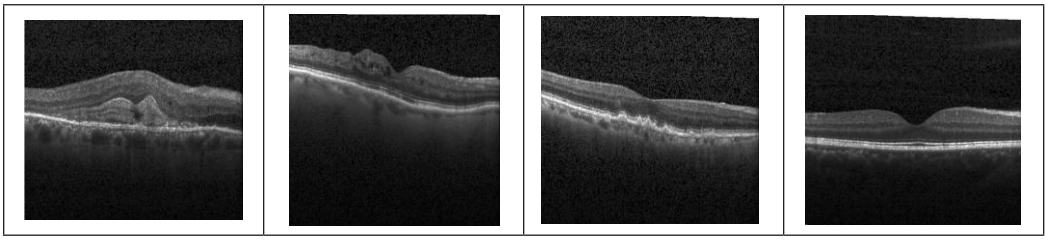


Figure 1. OCT Dataset retinal categories

Dataset Preprocessed

An asymmetric data augmentation strategy was adopted, where a range of stochastic transformations were applied to the training images, including:

Random resizing followed by random cropping to 224 × 224 pixels. Random horizontal and vertical flipping with probabilities 0.5 and 0.3, respectively; Colour jitter, modulating brightness, contrast, saturation, and hue by ±0.4, ±0.4, ±0.3, and ±0.1, respectively. Random affine transformation, including translation up to ±10% and scaling between (0.9 and 1.1). Random perspective transformation with distortion scale = 0.2 and probability = 0.3. Gaussian blur with a kernel size of 3 and sigma uniformly distributed between 0.1 and 2.0. Random erasing of rectangular regions with a probability of 0.2, scale range of [0.02, 0.15]. Subsequent normalisation of all training images using ImageNet-derived channel-wise statistics, where:

$$\mu = [0.485, 0.456, 0.406] \text{ and } \sigma = [0.229, 0.224, 0.225]$$

The validation and test images, on the other hand, were only subject to a deterministic data augmentation pipeline, in which all images were resized to 224×224 pixels and normalised.

Table 1
Class-Wise ACCOUNT Sample OCT2017 Dataset.

Class	Training	Validation	Test
CNV	31,624	**5,585**	250
DME	9,646	**1,703**	250
DRUSEN	7,323	**1,290**	250
NORMAL	22,368	**3,945**	261
TOTAL	70,961	**12,523**	1,011

In Table 1, to address the imbalance of data between categories, the training data size was set to 8616 images per category to stabilise and improve the model. Thus, a balanced training set consisting of 34464 images was obtained; for the DRUSEN category, the required size was achieved using data augmentation, while both the verification and testing sets listed in the table retained their original category distributions during model development and evaluation. The OCT2017 dataset (Kermany et al., 2018) classifies each image by disease, random patient ID, and image number. Test and training folders were predefined and created by the original developers specifically to separate patients into the two groups. In the proposed methodology, these folders are not merged; patient images used for training do not appear in the test set. From the original dataset design, the test set at the patient level remains independent.

Model Architecture

Patch Embedding Module

1. Input images → 224 × 224 × 3. The images are divided into a grid of non-overlapping patches, and each patch is linearly projected to a continuous embedding space through a convolutional projection layer. The dimensionality of the embedding is set to d=768.

2. The convolutional projection layer → P=16 kernel size and a stride → P=16. The nn.The Conv2d module in PyTorch is used for convolutional projection layers.

3. The input to the transformer model consists of a classification token [CLS] ∈ R^d and a sequence of patch tokens. The classification token [CLS] ∈ R^d.

4. The input sequence and the classification token are then linearly combined with the positional embeddings through an element-wise addition operation. The positional embeddings ∈ R^{(N+1) × d}.

Vision Transformer Encoder

Multi-head self-attention (MHSA) is applied to the input, projected into queries and keys/values with a linear layer (Linear (d, 3d)), and the projected values (H = 12) heads have (d_h = d/H = 64) dimensions. The scaled dot product attention is computed as Equation 1, below (Vaswani et al., 2017):

$$\text{Attention}(Q,K,V)=\text{Softmax}(\frac{QK^T}{\sqrt{d_k}})V \quad [1]$$

1. Dropout(p=0.1) is utilised for both attention weights and output projections and uses the stochastic depth (DropPath) in the residual branch with a linearly increasing rate from 0 to 0.1 for the 12 blocks.

2. Wise Feed-Forward Network (FFN) with two linear layers, GELU activation functions, and dropout layers, with a hidden dimension (4d = 3072), after the LayerNorm operation used (CLS) taken as a global feature representation, known as cls_token_out ∈ R^{768}. Figure 2 Attention Head (12 Layers).

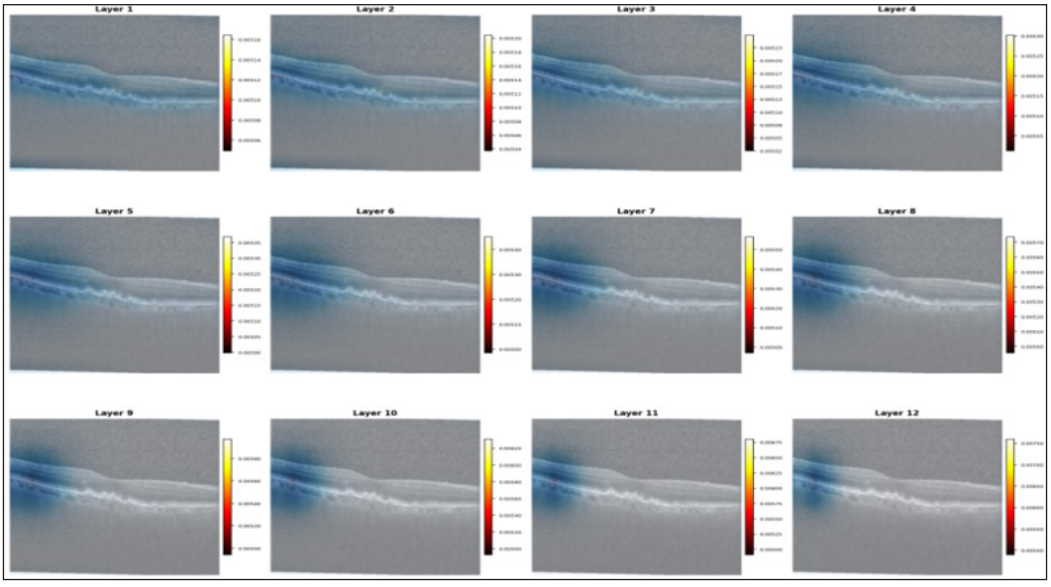


Figure 2. Attention maps for a typical DRUSEN case at the layer-wise average

Residual Feature Extractor

1. The architecture follows the sequence: 7×7 Conv (64) → Batch Normalisation → ReLU Activation → Max Pooling → Residual Block (Layer1) → Residual Block (Layer2) → Residual Block (Layer3) → Residual Block (Layer4) → Channel Depths of 256, 512, 1024, and 2048, respectively.

2. Adaptive Average Pooling (1×1) → Feature Maps ∈ R^ {2048}.

3. Layer Normalisation → GELU Activation → Dropout (0.1) to produce resnet_proj ∈ R^ {768}, which has the same dimension as the ViT branch to enable fusion.

Feature Fusion and Classification

1. The CLS token representation from the ViT branch (cls_token_out ∈ R^ {768}), ResNet-50 feature vector (resnet_proj ∈ R^ {768}), are concatenated along the feature dimension, and a joint representation is formed as combined ∈ R^ {1536}.

2. This is done in such a manner that it preserves the full information content of both branches without any prior assumption of weighting of both feature sources.

3. Passed through a Fusion Layer, which was implemented as a sequential module consisting of the following steps: a linear transformation (1536 → 768) → LayerNorm

→ GELU activation → dropout layer with a dropout rate of 0.15. The resulting fused embedding ∈ R^ {768} was then forwarded through the classification head.

4. The classification head (head), which was a two-layer MLP, consisted of the following steps: a linear transformation (768 → 512) → LayerNorm → GELU activation → dropout (p = 0.3).

5. The final linear layer (512 → 4) is applied for generating unnormalized class logits for the four OCT diagnostic classes, i.e., CNV, DME, DRUSEN, and NORMAL. During inference, class probabilities are computed by applying the softmax function to the logit outputs. Figure 3 shows the HR_{50}_$ViT_{B/16}$ Algorithm Pipeline. Table 2 illustrates the training network hyperparameter configuration.

Table 2
The Hyperparameter Configuration for Training HR_{50}_ViT_{16} in Oct

C	Value	Justification
Patch size (P)	16 × 16 px	196 tokens; standard ViT-B/16
Embedding dimension (d)	768	ViT-Base configuration
Transformer depth (L)	12	ViT-Base configuration
Attention heads (H)	12 (d_h = d/H = 64)	Head dim. = 64
MLP ratio (r)	4.0	d_ff = 3,072
Dropout rate	0.1	Regularisation
Max epochs	100	Subject to early stopping
learning rate (η)	3×10^{-4}	AdamW standard for transformers
Warm-up fraction	10% (pct_start = 0.1)	Gradient stability
Optimiser	AdamW (β_1=0.9, β_2=0.999, ε=1×10^{-8})	Standard for Transformer training
$ResNet_{50}$ branch initialization	ImageNet-pretrained	Transfer learning for low-level visual features
Software	PyTorch(2.0), scikit-learn (1.3), torchvision(0.15)	Reproducibility
LR scheduler	OneCycleLR	Faster convergence, reduces sensitivity to LR choice
Loss function	Cross-Entropy Loss	Standard for multi-class classification

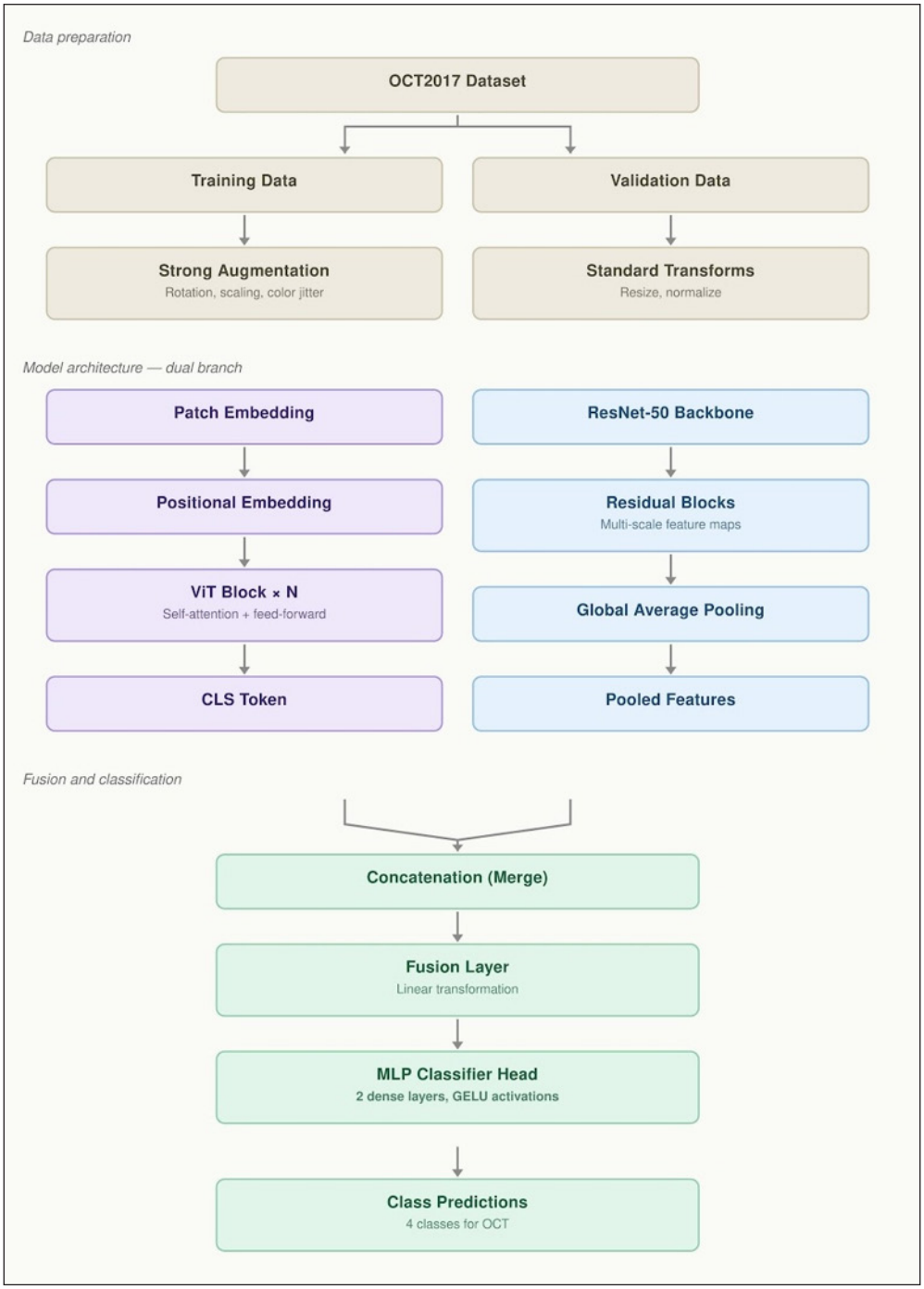


Figure 3. Overview of the proposed architecture of HR_{50}–$ViT\text{-}_{B/16}$ for OCT image classification

```
ViT_B/16 BRANCH — Global Context Extraction
═══════════════════════════════════════════
P ← PatchEmbedding(x)
   // Conv2d(kernel=16, stride=16) → 196 patches
   // Each patch projected: p_i ∈ ℝ^{768}
T ← Concat([CLS_token, p_1, …, p_196]) + PosEmbed
   // T ∈ ℝ^{197×768}
FOR i ← 1 TO 12 DO        // stochastic depth: 0 → 0.1
   T ← TransformerBlock_i(T)
      // MHSA: H=12 heads, d_h=64
      // QKV projection: Linear(768, 3×768)
      // Attention: softmax(QK^T / √d_h) · V
      // Dropout(0.1) on attention weights & projections
      // FFN: Linear(768→3072) → GELU → Linear(3072→768)
END FOR
cls_token_out ← LayerNorm(T)[ :, 0 ]    // ∈ ℝ^{768}
═══════════════════════════════════════════
② ResNet_50 BRANCH — Local Hierarchical Extraction
═══════════════════════════════════════════
f ← Conv2d(x, 64, 7×7, stride=2) → BN → ReLU → MaxPool
f ← ResidualBlock_Layer1(f)    // f ∈ ℝ^{56×56×256}
f ← ResidualBlock_Layer2(f)    // f ∈ ℝ^{28×28×512}
f ← ResidualBlock_Layer3(f)    // f ∈ ℝ^{14×14×1024}
f ← ResidualBlock_Layer4(f)    // f ∈ ℝ^{7×7×2048}
f ← AdaptiveAvgPool(f, 1×1).flatten()    // f ∈ ℝ^{2048}
resnet_proj ← Linear(f, 768)
          → LayerNorm → GELU → Dropout(0.1)
          // resnet_proj ∈ ℝ^{768}
═══════════════════════════════════════════
③ LEARNABLE FUSION LAYER
═══════════════════════════════════════════
combined ← Concat([cls_token_out, resnet_proj], dim=1)
        // combined ∈ ℝ^{1536}
        // preserves full information without prior weighting
fused ← Linear(combined, 768)
      → LayerNorm → GELU → Dropout(0.15)
      // fused ∈ ℝ^{768}
═══════════════════════════════════════════
④ CLASSIFICATION HEAD
═══════════════════════════════════════════
h      ← Linear(fused, 512) → LayerNorm → GELU → Dropout(0.3)
logits ← Linear(h, 4)                    // logits ∈ ℝ^4
IF training THEN
   loss ← CrossEntropyLoss(logits, y, label_smoothing=0.1)
ELSE
   ŷ ← argmax(Softmax(logits))
```

Note. An algorithm classification pipeline based on ResNet-50 and vision transformer for retinal disease classification from Optical Coherence Tomography (OCT) images

EVALUATION METRIC

The performance of the HR50_ViTB/16 model was measured using the following metric (Alsaif et al., 2023; Sujon et al., 2025):

Accuracy is the number of samples correctly classified, divided by the total number of samples in the data set. The accuracy is computed by using the following Equation 2.

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \quad [2]$$

Positive Precision, also known as Positive Predictive Value, is the number of cases that are positive divided by the total number of cases identified as positive. It is given by the following Equation 3:

$$\text{Precision} = \frac{TP}{TP + FP} \quad [3]$$

Recall (Sensitivity) is the measure of the capacity of the model to recognise the actual positive cases by computing the ratio of the number of actual positive cases to the total number of actual positive cases. It is given by the following Equation 4:

$$Recall = \frac{TP}{TP + FN} \quad [4]$$

F1_Score is the harmonic mean of both precision and Recall. It is the single scale that incorporates both precision and Recall. It is the most suitable scale to be used when dealing with a highly imbalanced dataset. This is because the F1_score is given by Equation 5:

$$F1\text{-}Score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \quad [5]$$

Wilson Score Interval

Equation 6 shows that the reliability of accuracy is calculated using the Wilson Score Interval for the test set. Where: $\hat{p}$ = observed proportion (accuracy), n^2 = sample size (test set size), z^2 =critical value of standard normal distribution (z = 1.96 for 95% confidence interval (Brown et al., 2001).

$$CI_{95\%} = \frac{\hat{p} + \frac{z^2}{2n} \pm z\sqrt{\frac{\hat{p}(1-\hat{p})}{n} + \frac{z^2}{4n^2}}}{1 + \frac{z^2}{n}} \quad [6]$$

Two-Proportion Z-Test

Equation 7, used to calculate the overall classification accuracy for the proposed model comparison with other baselines, $\hat{p}_1$, $\hat{p}_2$ = accuracy of Model 1 and Model 2, x_1 and

x_2 = number of correct classifications in each model, n_1,n_2= test set sizes ($n_1 = n_2 = 1{,}011$) (Agresti, 2019).

$$z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}(1-\hat{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \qquad [7]$$

Where : $\hat{p} = (x_1 + x_2) / (n_1 + n_2)$

RESULT AND DISCUSSIONS

Figure 4 shows the confusion matrix analysis. The proposed model's CNV class resulted in a Recall of 96.8%. The confusion between the two health problems can be attributed to both pathologies exhibiting characteristic radiological features of subretinal and intraretinal fluid collections, which present similar OCT intensity profiles in the macular region. The DME class resulted in 95.2% Recall, while the DRUSEN class resulted in 94.8% Recall, the lowest Recall for the proposed model. The improvement in the proposed model's DRUSEN classification performance compared to the previous CNN-based approach can be attributed to the global self-attention mechanism incorporated by the ViT branch, which captures the diffuse distribution of DRUSEN deposits over the macular area; the NORMAL class resulted in a recall of (97.7%). The feature space can be said to have been organized by the model itself because the features that have been organized by the model (the ultimate feature space, the 768 dimensional integrated feature space in the decision architecture of the system) show the accumulation of error in terms of one or two neighbouring cells (DRUSEN → DME/NORMAL; CNV/DME bidirectional), where some pairs of cells (CNV↔NORMAL, DME↔NORMAL) have close to zero confusion.

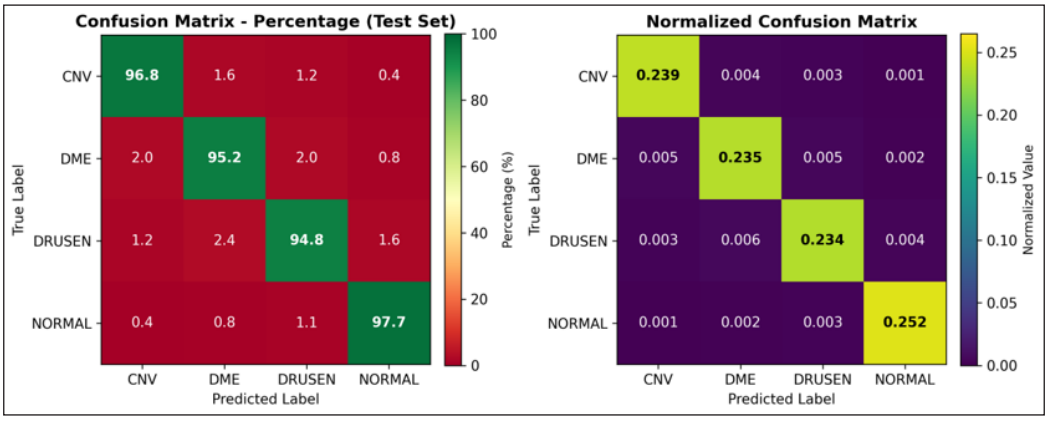
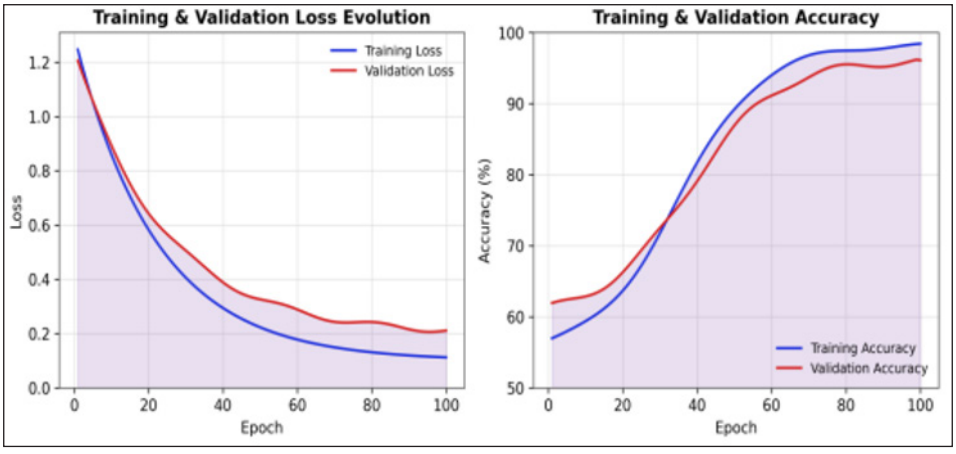


Figure 4. Confusion matrix-based performance evaluation of the proposed architecture on the OCT2017



(a) (b)

Figure 5. Complete training analysis for $HR_{50}_VIT_{B/16}$ on the OCT Dataset for a total of 100 epochs: (a)training &validation loss evolution; (b)training &validation accuracy

The training loss (Figure 5(a)) decreases monotonically from its initial value of about 1.25 to a final minimum of 0.196 at epoch 100. The validation loss has a similar history to the training loss, but over the entire epoch range. The training and validation accuracy curves in Figure 5 (b) show rapid improvement in epochs 1-40, followed by a gradual improvement from epochs 40 to 85, so the best validation accuracy is attained at epoch 100, at 96.11%.

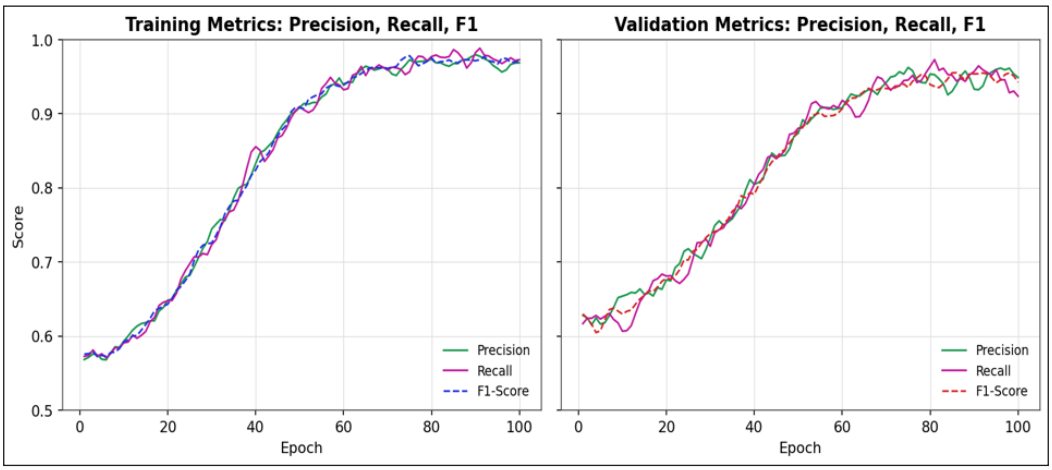


Figure 6. Training metrics (Precision, Recall, F1); Validation metrics (Precision, Recall, F1)

Figure 6 shows the epoch-wise behaviour of the three different classification metrics, Precision, Recall, and F1-Score, on the Training and validation metrics, thus offering

a detailed description of the learning process of the model apart from the accuracy measurement.

Table 3
Performance Measures of the Proposed hr_{50}–$vit\text{-}_{b/16}$ Model in October 2017.

Class	Precision	Recall	F1-score	Specificity	Acc. (%)
CNV	0.964	0.968	0.966	98.8%	96.8
DME	0.952	0.952	0.952	98.4%	95.2
DRUSEN	0.956	0.948	0.952	98.6%	94.8
NORMAL	0.977	0.977	0.975	99.1%	97.7
Weighted Avg.	**0.962**	**0.961**	**0.961**	**98.7%**	96.14

The classification report using the "one-versus-the-rest" approach is shown in Table 3. The accuracy of each class (which is calculated by (TP_c + TN_c)/N) ranges between 95.2- 94.8(DME, DRUSEN) and 97.7 (NORMAL). The Recall weighted Avg. (0.9614) corresponds numerically with overall accuracy (96.14%).

Table 4
Class-wise Comparison for Ablation Variants of the Proposed Architecture

No	Architecture	Precision	Recall	F1-Score	Acc. (%)
a	$ResNet_{50}$ Only	0.8022	0.7862	0.7941	79.41
b	$ViT_{B/16}$ Only	0.8591	0.8476	0.8533	85.33
c	Concat (No Fusion)	0.9364	0.9193	0.927	91.79
d	**HR50_$ViT_{B/16}$**	**0.9614**	**0.9614**	**0.9614**	96.14

Table 5
Class-Wise Analysis of Ablation Variants on the Oct 2017 Dataset

No	Architecture	CNV F1	DME F1	DRUSEN F1	NORMAL F1
a	$ResNet_{50}$ Only	0.93	0.74	**0.56**	0.94
b	$ViT_{B/16}$ Only	0.90	0.82	0.76	0.93
c	Concat (No Fusion)	0.94	0.90	0.88	0.95
d	**HR_{50}_$ViT_{B/16}$**	**0.966**	**0.952**	0.952	0.975

From the results in Tables 4 and 5, the ResNet50 architecture achieved F1 scores of 0.93-0.94 when applied separately to the CNV and NORMAL datasets, owing to its noticeable local features, including a new inversion of the vascular membrane structure and the retinal outline. However, there was a decrease in performance of the ResNet50 architecture for DRUSEN detection (F1: 0.56). The self-attention module of $ViT_{B/16}$ achieved significantly higher F1 scores for DRUSEN (0.76) and DME (0.82). The combination of

the $ResNet_{50}$ and $ViT_{B/16}$ features, without any learned weighting, achieves an impressive result (accuracy 91.79, DRUSEN F1 0.88) compared to each model alone. The result is $HR_{50}_ViT_{B/16}$ with a DRUSEN F1 value of 0.952.

$ResNet_{50}$ convolutional models operate on local receptive fields and can detect edge responses, intensity gradients, and texture patterns. For CNV, neovascular membranes exhibit bright, localised characteristics that can be detected in 3×3 or 5×5-pixel regions. In the case of DRUSEN, detection of small ripples in RPE surface geometric distortion is required at 50- to 200-pixel intervals across multiple retinal layers. The CNN lacks sufficient context to distinguish those small ripples from the smooth profile of the retinal pigment epithelium in healthy images, resulting in a high number of false negatives for this type.

The $ViT_{B/16}$ architecture divided image 224×224 into 196 patches of 16×16 pixels without intersection, and concurrently considers the symmetrical connections between all patches through multi-headed self-attention mechanism (12 heads, 12 layers, embed_dim = 768), so, the spatial pattern of retinal pigment epithelial ripples (yellow deposits) and fluid-filled spaces within retinal layers (diabetic macular edema) irregular patterns characterized by their spatial distribution, not by the intensity of their localized pixels. The $ViT_{B/16}$ architecture alone is limited because it lacks the fine-grained convolutional texture characteristics. The F1 value of the CNV index was reduced from 0.93 ($ResNet_{50}$) to 0.90 ($ViT_{B/16}$), representing a reduction of -0.03, since the resolution of 16×16 patches lacks the sub-patch detail of the strongly reflecting neovascular membranes; so, the global attention and local texture extraction complement each other but are not the same skill set; both of the essential architectures are insufficient alone.

The distinction between (c) and (d) is the way the two feature spaces are fused. Model (c) fuses the 768-dimensional $ResNet_{50}$ representation with the 768-dimensional ViT CLS code into a 1536-dimensional vector when it is fed to the classification head. So, the $ResNet_{50}$ and $ViT_{B/16}$ feature paths are considered equally valuable for all classes, although DRUSEN classification takes more advantage of $ViT_{B/16}$ attention features, while CNV classification takes more advantage of $ResNet_{50}$ texture features.

The combination block in $HR_{50}_ViT_{B/16}$ (Linear 1536→768, LayerNorm, GELU, Dropout) learns a projection that adaptively reweights each feature flow based on the input image content. During training, the combination block parameters capture a specific weighting for each class, learning the weighting of the two complementary feature sets specific to that class. For DRUSEN, the fusion layer increases the contribution of $ViT_{B/16}$ attention features that capture the RPE surface texture, whereas for CNV, it increases the contribution of $ResNet_{50}$ texture reactions to the highly reflective membranes. For images with DRUSEN, the gradient update provides more weight to $ViT_{B/16}$ branch features; for images with CNV, it provides more weight to $ResNet_{50}$ branch features. It is unachievable through simple clustering.

The contribution of the fusion block (c → d) alone quantitatively is DRUSEN F1 (+0.072), DME F1 (+0.052), CNV F1 (+0.026), and NORMAL F1 (+0.025). The fusion module accounts for 18.4% of the total DRUSEN F1 increase from the ResNet-50 model to the proposed architecture (0.072 / 0.392 = 0.184) and 24.5% of the total DME F1 increase (0.052 / 0.212 = 0.245).

The Wilson interval statistical analysis was applied to assess the accuracy and reliability of each model, and a two-ratio z-test was used to establish statistical significance of accuracy differences between the proposed model and each baseline on a test set of 1011 images. The results can be illustrated in Tables 6 and 7.

Table 6
Wilson Score Confidence Intervals

No	Architecture	Acc. (%)	95% CI Lower	95% CI Upper
a	$ResNet_{50}$ Only	79.41	76.81	81.79
b	$ViT_{B/16}$ Only	85.33	83.02	87.38
c	Concat (No Fusion)	91.79	89.94	93.33
d	**HR_{50}_$ViT_{B/16}$**	**96.14**	**94.77**	**97.16**

Table 6 shows the accuracies of the different architectures on the fixed test set from OCT2017 (N = 1,011), with 95% confidence intervals estimated using Wilson score confidence intervals. The observation intervals of the four models don't overlap, and the features learned in each of the proposed architectures ($ViT_{B/16}$ branch, $ResNet_{50}$ branch, Learner integration) are transferred to the following layer.

Table 7
Z-Test for Two Ratios to Compare Models

Architecture	z-statistic	p-value
d vs. a (HR_{50}_$ViT_{B/16}$ vs. $ResNet_{50}$ only)	11.48	4.8×10^{-4}
d vs. b (HR_{50}_$ViT_{B/16}$ vs. $ViT_{B/16}$ only)	8.38	5.0×10^{-6}
d vs. c (HR_{50}_$ViT_{B/16}$ vs. Concat, no fusion)	4.11	< 0.0001

Table 7 shows two-proportion z-test results comparing the overall classification accuracy of the proposed HR50_ViTB/16 model against each ablation baseline on the test set (N = 1,011 per model). The interval results for the models indicate that each model was well estimated and remained stable as the test set size increased, and the two-ratio z-test results confirm the information gained by each model in the proposed merging architecture of the framework. $p = 4.0 \times 10^{-5}$ is the quantitative statistical basis demonstrating the improvement in the framework's merging architecture. Table 8 compares the proposed methodology's performance with other studies.

Table 8
Comparative Performance of Prior Studies on the Oct 2017 dataset

Study	Architecture	Accuracy	Key Distinction from $HR_{50}_ViT_{B/16}$
Perdomo et al., 2018	Custom CNN (12-layer) Leave-one-out CV	Accuracy= **93.75%**	on DME vs Normal classification using binary classification
Bhandari et al., 2023	Lightweight CNN + with k-fold (k=10), using FGSM, LIME, and SHAP	Accuracy= 96 %, test accuracy= 94.29%, verification accuracy= 94.12%	Classification using pure CNN on OCT images (CNV, DME, DRUSEN, Normal), Novelty is adversarial robustness (FGSM) + explainability (LIME, SHAP)
Dutta et al., 2023	InceptionV3+ResNet-50+ViT	weighted F1 / Acc= 0.944 / 94.46%	triple-stream hybrid model for Conv-ViT with InceptionV3, ResNet-50 and ViT used in achieved a weighted F1-score 0.944, F1-score of DRUSEN 0.780 and 94.46% accuracy in OCT2017 dataset.
Ma et al., 2022	ConvNet + Transformer	weighted F1 / Acc=91.56%	hybrid ConvNet-Transformer architecture, Residual dense block+ parallel ConvNet and Transformer branches for adaptive re-weighting fusion. For the OCT2017 dataset, HCTNet was reach 91.56% accuracy
Yang et al., 2024	Multi-scale CNN + Transformer	97.16%	In October 2017, the process of four-class classification used HRS-Net, a hybrid multi-scale network in which ResNet50 was branched out into parallel CNN (residual + attention) and Swin Transformer streams.
$HR_{50}_ViT_{B/16}$	$ResNet_{50}$ + $ViT_{B/16}$	96.14%	CNV, DME, DRUSEN, Normal

Transformers have also been used for pixel-wise segmentation of medical images in TransUNet (Chen et al., 2024; Turan et al., 2025), and the Attention Mechanism (Valanarasu et al., 2021) was applied to gland segmentation and polyp segmentation, respectively. These two methods are relatively new applications of the Attention Mechanism in ophthalmology. They should both be adapted for the classification problem.

CONCLUSION

The ($HR_{50}_ViT_{B/16}$) approach used for the classification of OCT images for retinal pathology showed 96.14% overall accuracy. The use of the $ViT_{B/16}$ encoder and the $ResNet_{50}$ backbone in parallel enables joint consideration of global spatial dependencies, modelled by 12 layers of Multi-Head Self-Attention with 196 patch tokens, as well as hierarchical local texture features extracted via the four residual stages. The improvement in DRUSEN F1-score was 0.952. The qualitative analysis of the attention overlay showed that the ViT

branch successfully localises attention to the relevant sub-RPE and intraretinal areas in the DRUSEN and DME cases, respectively, providing evidence of the model's ability to learn diagnostically meaningful spatial representations. To further improve the existing fusion paradigm, we propose replacing the linear concatenation strategy with a new strategy based on the Multi-Head Cross-Attention (MHCA) method in the ViT transformer as Future Work. It can effectively address inter-class similarity. Training the model with multiple random seeds (at least five) to enhance statistical stability and estimate the variation in the operation process.

LIST OF ABBREVIATION

Term	Definition
CNV	Neovascularisation
DME	Diabetic Macular Oedema
AMD	Age-Related Macular Degeneration
OCT	Optical Coherence Tomography
ViT	Vision Transformer Model

REFERENCES

Agresti, A. (2019). *An introduction to categorical data analysis*. (3rd ed.). Wiley.

Akça, S., Garip, Z., Ekinci, E., & Atban, F. (2024). Automated classification of choroidal neovascularization, diabetic macular edema, and drusen from retinal OCT images using vision transformers: A comparative study. *Lasers in Medical Science, 39*, Article 140. https://doi.org/10.1007/s10103-024-04089-w

Akpinar, M. H., Sengur, A., Faust, O., Tong, L., Molinari, F., & Acharya, U. R. (2024). Artificial intelligence in retinal screening using OCT images: A review of the last decade (2013-2023). *Computer Methods and Programs in Biomedicine, 254*, Article 108253. https://doi.org/10.1016/j.cmpb.2024.108253

Alsaif, O. I., Hasan, S. Q., & Maray, A. H. (2023). Using skeleton model to recognize human gait gender. *IAES International Journal of Artificial Intelligence, 12*(2), 974-983. https://doi.org/10.11591/ijai.v12.i2.pp974-983

Aslan, M. F. (2023). Comparison of vision transformers and convolutional neural networks for skin disease classification. In *Proceedings of the International Conference on New Trends in Applied Sciences* (pp. 31-39). https://doi.org/10.58190/icontas.2023.51

Bhandari, M., Shahi, T. B., & Neupane, A. (2023). Evaluating retinal disease diagnosis with an interpretable lightweight CNN model resistant to adversarial attacks. *Journal of Imaging, 9*(10), Article 219. https://doi.org/10.3390/jimaging9100219

Brown, L. D., Cai, T. T., & DasGupta, A. (2001). Interval estimation for a binomial proportion. *Statistical Science, 16*(2), 101-133. https://doi.org/10.1214/ss/1009213286

Burton, M. J., Ramke, J., Marques, A. P., Bourne, R. R. A., Congdon, N., Jones, I., Ah Tong, B. A. M., Arunga, S., Bachani, D., Bascaran, C., Bastawrous, A., Blanchet, K., Braithwaite, T., Buchan, J. C., Cairns, J.,

Cama, A., Chagunda, M., Chuluunkhuu, C., Cooper, A., . . . Faal, H. B. (2021). The Lancet Global Health Commission on Global Eye Health: Vision beyond 2020. *The Lancet Global Health, 9*(4), e489-e551. https://doi.org/10.1016/s2214-109x(20)30488-5

Chen, J., Mei, J., Li, X., Lu, Y., Yu, Q., Wei, Q., Luo, X., Xie, Y., Adeli, E., Wang, Y., Lungren, M. P., Zhang, S., Xing, L., Lu, L., Yuille, A., & Zhou, Y.. (2024). TransUNet: Rethinking the U-Net architecture design for medical image segmentation through the lens of transformers. *Medical Image Analysis, 97*, Article 103280. https://doi.org/10.1016/j.media.2024.103280

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., & Houlsby, N. (2021). An image is worth 16×16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*. https://openreview.net/forum?id=YicbFdNTTy

Du, K., Nair, A. R., Shah, S., Gadari, A., Vupparaboina, S. C., Bollepalli, S. C., Sutharahan, S., Sahel, J.A., Jana, S., Chhablani, J., & Vupparaboina, K. K. (2024). Detection of disease features on retinal OCT scans using RETFound. *Bioengineering, 11*(12), Article 1186. https://doi.org/10.3390/bioengineering11121186

Dutta, P., Sathi, K. A., Hossain, M. A., & Dewan, M. A. A. (2023). Conv-ViT: A convolution and vision transformer-based hybrid feature extraction method for retinal disease detection. *Journal of Imaging, 9*(7), Article 140. https://doi.org/10.3390/jimaging9070140

Fang, L., Jin, Y., Huang, L., Guo, S., Zhao, G., & Chen, X. (2019). Iterative fusion convolutional neural networks for classification of optical coherence tomography images. *Journal of Visual Communication and Image Representation, 59*, 327-333. https://doi.org/10.1016/j.jvcir.2019.01.022

Gencer, G., & Gencer, K. (2025). Advanced retinal disease detection from OCT images using a hybrid squeeze and excitation enhanced model. *PLOS ONE, 20*(2), Article e0318657. https://doi.org/10.1371/journal.pone.0318657

Hassan, E., Elmougy, S., Ibraheem, M. R., Hossain, M. S., AlMutib, K., Ghoneim, A., AlQahtani, S. A., & Talaat, F. M. (2023). Enhanced deep learning model for classification of retinal optical coherence tomography images. *Sensors, 23*(12), Article 5393. https://doi.org/10.3390/s23125393

He, X., Deng, Y., Fang, L., & Peng, Q. (2021). Multi-modal retinal image classification with modality-specific attention network. *IEEE Transactions on Medical Imaging, 40*(6), 1591–1602. https://doi.org/10.1109/tmi.2021.3059956

Kermany, D. S., Goldbaum, M., Cai, W., Valentim, C. C. S., Liang, H., Baxter, S. L., McKeown, A., Yang, G., Wu, X., Yan, F., Dong, J., Prasadha, M. K., Pei, J., Ting, M. Y. L., Zhu, J., Li, C., Hewett, S., Dong, J., Ziyar, I., . . . Zhang, K. (2018). Identifying medical diagnoses and treatable diseases by image-based deep learning. *Cell, 172*(5), 1122-1131.e9. https://doi.org/10.1016/j.cell.2018.02.010

Laouarem, A., Kara-Mohamed, C., Bourennane, E.B., & Hamdi-Cherif, A. (2024). HTC-retina: A hybrid retinal diseases classification model using transformer–convolutional neural network from optical coherence tomography images. *Computers in Biology and Medicine, 178*, Article 108726. https://doi.org/10.1016/j.compbiomed.2024.108726

Li, D., Ran, A. R., Cheung, C. Y., & Prince, J. L. (2023). Deep learning in optical coherence tomography: Where are the gaps? *Clinical and Experimental Ophthalmology, 51*(8), 853-863. https://doi.org/10.1111/ceo.14258

Liew, A., & Agaian, S. (2025). Comprehensive survey of OCT-based disorders diagnosis: From feature extraction methods to robust security frameworks. *Bioengineering, 12*(9), Article 914. https://doi.org/10.3390/bioengineering12090914

Ma, Z., Xie, Q., Xie, P., Fan, F., Gao, X., & Zhu, J. (2022). HCTNet: A hybrid ConvNet-transformer network for retinal optical coherence tomography image classification. *Biosensors, 12*(7), Article 542. https://doi.org/10.3390/bios12070542

Mohammed, N. S., Almolaa, O. S., Abdullah, A. S., & Alsaif, K. I. (2026). Accurate brain tumor classification using MRI images based on a hybrid Vision Transformer and BiLSTM framework. *ITEGAM-JETIA, 12*(60), 526-535. https://doi.org/10.5935/jetia.v12i60.3889

Perdomo, O., Otalora, S., Gonzalez, F. A., Meriaudeau, F., & Muller, H. (2018). OCT-NET: A convolutional network for automatic classification of normal and diabetic macular edema using SD-OCT volumes. In *2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)* (pp. 1423-1426). IEEE. https://doi.org/10.1109/isbi.2018.8363839

Rahil, M., Anoop, B. N., Girish, G. N., Kothari, A. R., Koolagudi, S. G., & Rajan, J. (2023). A deep ensemble learning-based CNN architecture for multiclass retinal fluid segmentation in OCT images. *IEEE Access, 11*, 17241-17251. https://doi.org/10.1109/access.2023.3244922

Raut, P., Babar, S., & Mahalle, P. (2025). Fusion of vision transformer, Inception-V3 and ResNet50 for efficient eye disease detection. *Journal of Information Systems Engineering and Management, 10*(7s), 315-325. https://doi.org/10.52783/jisem.v10i7s.864

Sujon, K. M., Hassan, R., Choi, K., & Samad, M. A.. (2025). Accuracy, precision, recall, F1-score, or MCC? Empirical evidence from advanced statistics, ML, and XAI for evaluating business predictive models. *Journal of Big Data, 12*, Article 268. https://doi.org/10.1186/s40537-025-01313-4

Teo, Z. L., Tham, Y.C., Yu, M., Chee, M. L., Rim, T. H., Cheung, N., Bikbov, M. M., Wang, Y. X., Tang, Y., Lu, Y., Wong, I. Y., Ting, D. S. W., Tan, G. S. W., Jonas, J. B., Sabanayagam, C., Wong, T. Y., & Cheng, C.Y. (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

Turan, S. C., Çıplak, Z., Sarıkaş, A., & Yıldız, K. (2025). A deep learning approach to document recovery: High performance with DenoiseU-Net. *Afyon Kocatepe Üniversitesi Fen ve Mühendislik Bilimleri Dergisi, 25*(6), 1323-1335. https://doi.org/10.35414/akufemubid.1628066

Valanarasu, J. M. J., Oza, P., Hacihaliloglu, I., & Patel, V. M. (2021). Medical transformer: Gated axial-attention for medical image segmentation. In M. de Bruijne, P. C. Cattin, S. Cotin, N. Padoy, S. Speidel, Y. Zheng, & C. Essert (Eds.), *Medical image computing and computer assisted intervention – MICCAI 2021* (pp. 36-46). Springer. https://doi.org/10.1007/978-3-030-87193-2_4

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. In I. Guyon, U. von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, & R. Garnett (Eds.), *Advances in neural information processing systems* (Vol. 30, pp. 5998-6008). Curran Associates. https://proceedings.neurips.cc/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html

Vision Loss Expert Group of the Global Burden of Disease Study, & the GBD 2019 Blindness and Vision Impairment Collaborators. (2024). Global estimates on the number of people blind or visually impaired by diabetic retinopathy: A meta-analysis from 2000 to 2020. *Eye, 38*(11), 2047-2057. https://doi.org/10.1038/s41433-024-03101-5

Wang, H., Guo, X., Song, K., Sun, M., Shao, Y., Xue, S., Zhang, H., & Zhang, T. (2023). OCTFormer: An efficient hierarchical Transformer network specialized for retinal optical coherence tomography image recognition. *IEEE Transactions on Instrumentation and Measurement, 72*, 1-17. https://doi.org/10.1109/TIM.2023.3329106

Wang, S., Feng, S., Wang, Z., Ji, Z., Liu, J., Chen, W., Fu, B., Liu, R., Chen, W., Dai, Y., & Li, C. (2025). Structural-prior guided and feature-enhanced transformer with masked image modeling pretraining for retinal layers and fluid segmentation in macular edema OCT images. *Biomedical Optics Express, 16*(12), 5096-5117. https://doi.org/10.1364/boe.574769

Xu, Z., Wang, W., Yang, J., Zhao, J., Ding, D., He, F., Chen, D., Yang, Z., Li, X., Yu, W., & Chen, Y. (2021). Automated diagnoses of age-related macular degeneration and polypoidal choroidal vasculopathy using bi-modal deep convolutional neural networks. *British Journal of Ophthalmology, 105*(4), 561-566. https://doi.org/10.1136/bjophthalmol-2020-315817

Yang, H., Chen, L., Cao, J., & Wang, J. (2024). HRS-Net: A hybrid multi-scale network model based on convolution and transformers for multi-class retinal disease classification. *IEEE Access, 12*, 144219-144229. https://doi.org/10.1109/access.2024.3467286

Yang, M., Du, J., & Lv, R. (2025). CRAT: Advanced transformer-based deep learning algorithms in OCT image classification. *Biomedical Signal Processing and Control, 104*, Article 107544. https://doi.org/10.1016/j.bspc.2025.107544

Zhou, Z., Niu, C., Yu, H., Zhao, J., Wang, Y., & Dai, C. (2023). Diagnosis of retinal diseases using the vision transformer model based on optical coherence tomography images. *Proceedings of SPIE, 12601*, Article 1260102. https://doi.org/10.1117/12.2665918