Advanced Hybrid CNN-ViT Ensemble with Attention and FPN Mechanism for Retinal OCT Disease Classification

Abdullah Al Noman[1]*, Eamin Hasan Shanto[1], Mahir Faysal[1], Jamil Hasan[1], Samidul Islam Imran Kayes[1], Mohammad Jahangir Alam[1]

[1] Department of Computer Science and Engineering, Daffodil International University, Dhaka 1216, Bangladesh

{noman15-5713*, shanto15-5251, faysal23105101060, hasan15-5222, kays2305101624, jahangir.cse}@diu.edu.bd

Abstract. Retinal diseases can be considered one of the leading causes of vision loss on the global scene, and OCT imaging is crucial in the timely diagnosis of the disease. In this paper, a hybrid model of deep learning, which combines Convolutional Neural Networks (CNNs), Vision Transformers (ViTs), Feature Pyramid Networks (FPN) and cross-modal attention, is proposed to simultaneously exploit local retinal texture and global structural features. An ensemble mechanism is used to combine three fusion strategies into attention-based, concatenation strategy, and weighted fusion to enhance robustness and accuracy. Experiments on a publicly available OCTDL dataset of seven disease classes get 96.3% accuracy and a 95.1% macro F1-score and outperform the traditional CNN and ViT baselines. These results prove the high promise of the hybrid ensemble models towards the effective retinal OCT multi-disease classification.

Keywords: Retinal OCT, Deep Learning, Convolutional Neural Networks, Vision Transformer, Feature Pyramid Network, Ensemble Learning.

1 Introduction

Retinal Optical Coherence Tomography (OCT) is a non-invasive imaging methodology that forms cross-sectional images of the retina with the help of reflected light. It has transformed the field of ophthalmology to give high resolution images of layers of the retina, and therefore, physicians can identify the abnormalities which could not be identified under normal eye tests [1]. Thanks to its accuracy and safety, OCT has become a fundamental instrument of contemporary clinical practice.

The necessity of OCT-based diagnosis is growing along with the increasing rates of retinal pathology across the world. Diabetic Macular Edema, Choroidal

M. S. Arefin et al. (eds.), *Proceedings of the International Conference on Intelligent Data Analysis and Applications (IDAA 2025)*, Advances in Intelligent Systems Research 206,
https://doi.org/10.2991/978-94-6239-664-7_41

Neovascularization, and Drusen are some of the major causes of visual impairment [2] that affect millions of patients. As the world population with diabetes reaches over 400 million, aging populations are placing an additional burden on the prevalence of diabetes, the quantity of people at risk of developing retinal disorders is likely to increase substantially over the next few decades. This increases the urgency of early and correct diagnosis using OCT analysis as never before.

Recently deep learning has proven to be highly promising in automatic interpretation of OCT images, although existing methods are still limited [3]. Convolutional Neural Networks (CNNs) are good at capturing local retinal information, such as texture and edges, but fail to capture long-range information over the image. Vision Transformers (ViTs), by contrast, are good at learning global relationships and bad at learning fine local structures [4]. This discrepancy leaves a serious gap in research, as the current techniques focus on the local image without a global context, or provide the global picture but lose localization in fine details on retinas. Besides, the majority of models do not support the multi-scale characteristics of retinal structures, which are critical to disease recognition. In order to overcome these issues, an effective framework that integrates the complementary advantages of CNNs and ViTs is required, as well as improves the ability of feature learning at more than one scale [5].

To close this gap, in this work, we propose a hybrid ensemble of CNN and ViT with Feature Pyramid Networks and attention mechanisms. Though the combination of CNN-ViT models has been used recently in medical imaging literature, our study presents a novel combination of Feature Pyramid Networks, cross-modal attention, and three fusion strategies, which are complementary and integrated via an ensemble meta-learner. This multi-level structure allows the simultaneous modeling of fine textures, global retinal structure as well as scale-conscious features, which have previously not been collaboratively considered in previous studies of scaling in OCT classification. More so, our model shows good performance of minority classes like RAO and VID, showing better robustness as compared to existing hybrid models. The contributions are:

- Proposed a hybrid CNN–ViT ensemble with FPN and cross-modal attention.
- Achieved good performance on a 7-class OCT dataset, a challenging multiclass task with significant feature overlap and minority disease categories.
- Achieved 96.3% accuracy, surpassing EfficientNet-B3 and Vision Transformer baselines across key metrics.

2 Literature Review

Recent retinal OCT classification works have mostly used Convolutional Neural Networks to extract local features and Vision Transformers to model the global context. Nevertheless, most of the works that exist tend to be based on single branch architectures, which points to the necessity of hybrid solutions that can combine both schools of thought to better serve diagnostics.

Babaqi et al. [4] suggested a CNN and transfer learning-based classification. Their objective was to identify diabetic retinopathy, cataract and glaucoma using OCT images. Convolved layers were used together with transfer learning to categorize using their model into multiple classes. The findings demonstrated that transfer learning had 94%

accuracy, which was higher than CNN (84%). Dai et al. [7] presented pretraining of medicine and duplicate of samples to enhance OCT diagnosis. They trained networks on RadImageNet and made a transfer learning on retinal data. The output consistency strategy was applied in a very effective manner with the help of the strategy of sample replication. models provided almost 95% accuracy. Alizadeh Egetedar et al. [1] came up with a CAD system to diagnose retinal disorder. They incorporated semantic annotations with ophthalmologists on OCT images in order to enhance their interpretation. They trained their deep learning model on the 29,800 images of UCSD OCT. The approach had a precision of 97.6% and what it calls heatmaps that identified disease regions.

Baba et al. [3] suggested a retinal disease classification CNN custom. They paid attention to identification of the macular diseases like CNV, DME and drusen. They had their model trained and compared with ML and transfer learning. The proposed CNN was found to have a testing accuracy of 98% which is superior to the traditional and existing methods. OCTDL is an open-source OCT analysis dataset that was introduced by Kulyabin et al. [10]. The dataset had more than 2000 images that were labeled on six retinal diseases. Photos were taken with the help of the Optovue Avanti system and confirmed by experts. The classification of deep learning showed good perspectives, which makes OCTDL useful in automated diagnostics. Ma et al. [11] offered CFANet to segment the optic disc and the macula. They did this by scanning the OCTA to give the ODMI dataset. The model has combined both coarse and fine attention modules in order to be more accurate. The findings demonstrated greater than 98% segmentation accuracy which confirms its robustness in clinical use. Hassan et al. [8] came up with EOCT that refined the classification of OCT with modified ResNet50. They used random forest with Adam optimization to achieve better diagnostic performance. Their architecture was compared with the pretrained architectures such as VGG16 and separable convolutions. The model had a high accuracy of 97.4% and this shows it is better in retinal disease classification. Yoon et al. [15] proposed a CNN in the diagnosis of central serous chorioretinopathy. They compared SD-OCT pictures between acute and chronic CSC cases and made the differences. It was compared to VGG16, ResNet50 and ophthalmologists in terms of benchmarking. It had a diagnostic accuracy of 93.8% and an accuracy of 97.6% to differentiate between acute and chronic CSC.

A modified version of VGG16 was proposed by Jaimes et al. [9] to classify retinal OCT. They characterized using the transfer learning of CNV, DME, drusen and normal cases. The model they used was based on gradient-based heatmaps to enhance clinical readability. The method had a high accuracy of 95.19%, high recall and F1-score. Dahiya et al. [6] have compared early diabetic retinopathy predictive algorithms. They made experimental models to evaluate the measures of performance, scalability, and efficiency. According to their analysis, the variation on accuracy depending on the choice of algorithm was significantly noted. The paper has placed stress on the choice of best algorithms in order to predict and manage diseases reliably. Pekala et al. [14] used a combination of FCNs and Gaussian processes to segment an OCT image. They used the method on diabetic retinopathy images of the University of Miami. Their model realized an unsigned error of 1.06, which was lower compared to the error of 1.10 realized by the clinicians. Findings affirmed that the technique was comparable or even better than the human level of segmentation.

GAGUNet was suggested by Oh et al. [13] to segment lesions in nAMD OCT scans. They combined the graph convolution networks with attention guided UNet to reason on a global scale. They had their own data on the results of clinical scans examined by

experts and testing of the RETOUCH dataset. The model was also more effective than the baselines and demonstrated better segmentation, both in quantitative and qualitative analysis. A ResNet18 model was tested by Chiang et al. [5] to perform the detection of AMD lesions. They trained on B-scans that were sensitive to iRORA and cRORA lesions. Independent datasets were assessed with respect to performance in terms of the AUROC and AUPRC scores. Their model reached an AUROC of 0.84 and AUPRC 0.82 which was comparable to clinicians.

Table 1 shows the comparison among existing works.

Table 1. Comparison of Existing Works.

Author (Year)	Methodology	Accuracy/Results
Babaqi et al. (2023) [4]	CNN with transfer learning for multi-class eye disease classification	Transfer learning 94% vs CNN 84%
Dai et al. (2024) [7]	Medical pre-training (RadImageNet) + sample replication using JS divergence	Accuracy up to 95%, improvements 3.7–8.6%
Jaimes et al. (2025) [9]	Modified VGG16 with transfer learning and heatmap interpretability	95.19% accuracy with high precision and recall
Araújo et al. (2023) [2]	Few-shot Outlier Exposure + Cosine distance for OOD detection	AUC 0.981 (AMD), AUC 0.937 (near-OOD)
Mariottoni et al. (2022) [12]	CNN on RNFL thickness from SD-OCT to detect glaucoma progression	AUC 0.938, Sensitivity 87.3%, Specificity 86.4%
Chiang et al. (2022) [5]	ResNet18 model detecting iRORA and cRORA in OCT scans	AUROC 0.83–0.84, AUPRC up to 0.82

3 Methodology

The methodology contains data collection, preprocessing, dataset splitting, model training, model test and model evaluation. The designed methodology is based on the integration of Convolutional Neural Networks and Vision Transformers into a hybrid ensemble system to classify retinal OCT diseases. A Feature Pyramid Network is used to extract multi-scale retinal features and cross-modal attention is used to combine local CNN features with global ViT features. In the model, three fusion strategies, which include attention-based, concatenation and weighted fusion, are combined in an ensemble to make robust predictions. Lastly, the system is tested and trained on an OCT dataset containing 7 disease classes with a high degree of reliability and accuracy. Fig.1 shows the comprehensive methodology diagram:

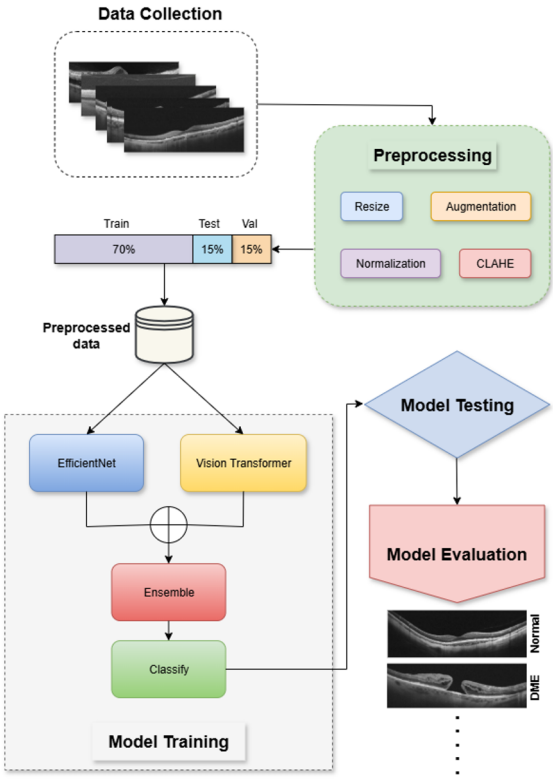
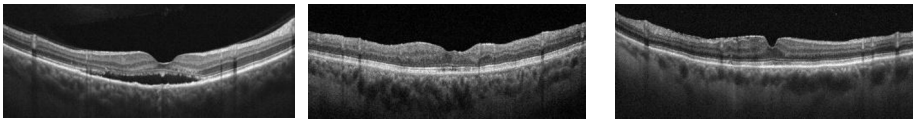


Fig. 1. Proposed Methodology

3.1 Dataset Description

This research was based on the OCTDL data set to assist in the classification of multi-class retina diseases [17]. It consists of 2,366 images of OCT of seven categories that are clinically confirmed, such as Age-related Macular Degeneration (AMD), Diabetic Macular Edema (DME), Epiretinal Membrane (ERM), Normal (NO), Retinal Artery Occlusion (RAO), Retinal Vein Occlusion (RVO), and Vitreomacular Interface Disease (VID). The pictures are derived using actual clinical exams, which gives a natural range of retinal structures and pathological differences. Different disease categories have different visual appearances and thus learning deep models can be reliably learned.

Fig 2 visualizes the images of seven classes:



AMD DME ERM

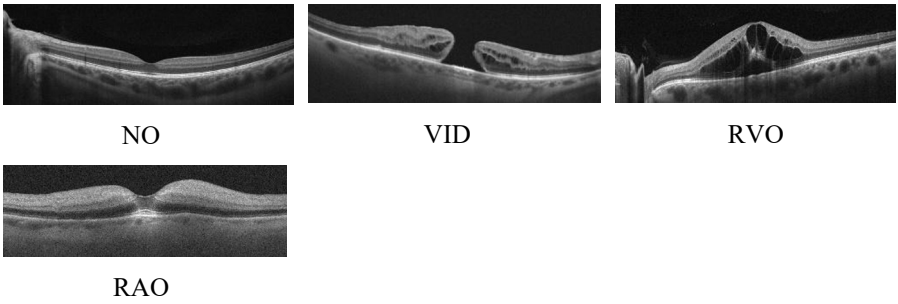


Fig 2: Dataset Class Sample

Table 2 shows the summary of the dataset:

Table 2. Sample Dataset

Class	No of images	After Augmentation
AMD	1231	600
ERM	155	500
NO	332	500
RAO	22	300
RVO	101	500
VID	76	500
DME	147	500

3.2 Data Preprocessing

In this section, the attention is directed to the preprocessing techniques that are applied to the images of and reliability of the dataset. Preprocessing is an essential task in medical imaging studies that processes the raw image data in preparation of successful model training and enhanced performance of generalization. There are a number of preprocessing methods that were used and they will be listed below:

Image Cleaning: This involved the initial stage of cleaning the data through the detection and removal of corrupted image files and unreadable files in order to ensure data integrity. Duplicates of scans, in case, have been removed to eliminate redundancy and bias in training. This was so that every image of the retina made its own contribution to the task of classification.

Normalization and Resizing: All the images were resized to the same resolution so that they can be used with CNN and ViT. The values of pixel intensity were also scaled to a range of [0,1], which allowed leveling the learning process and avoiding numerical imbalance. This was done to provide uniform contrast and brightness values throughout the dataset.

Data Augmentation: Data augmentation was used to counter imbalance of classes and enhance model stability. After augmentation and balancing, class sizes were adjusted using oversampling for minority classes and undersampling for AMD to reduce class dominance that is shown in Table 2. Such other augmentations as horizontal flipping, small rotations, zooming, and cropping were also used, making the images appear more diverse and allowing the model to study more stable retinal features.

Contrast Enhancement: In order to bring out the faint retinal formations, Contrast Limited Adaptive Histogram Equalization (CLAHE) was also used on some of the images. This increased both local contrast and disease-related differences, which the model helped to identify fine retinal abnormalities.

Data Splicing: Once the data preprocessing was done, the whole dataset was split into three segments; 70 percent training, 15 percent validation, and 15 percent testing. This approach was used to be sure that the model was trained on a big part of the data yet was tested on unknown samples to be validated and tested. Table 3 shows the manner in which the data are spread to the seven classes.

Table 3. Distributed Dataset

Class	Train	Test	Validation
AMD	420	90	90
ERM	350	75	75
NO	350	75	75
RAO	210	45	45
RVO	350	75	75
VID	350	75	75
DME	350	75	75

3.3 Proposed Model Description

Our proposed architecture is a hybrid of Convolutional Neural Networks (CNNs) and Vision Transformers (ViTs) to produce reliable retinal OCT disease classification. CNNs are effective at learning local structural patterns, although they frequently do not learn global dependencies. Conversely, ViTs excel in global context modelling but struggle with local fine-grained details. To address this drawback, we develop a hybrid ensemble model that incorporates CNN and ViT features in a Feature Pyramid Network (FPN) and cross-modal attention, which is then followed by an ensemble learning mechanism to make a robust decision.

CNN Backbone with Multi-Scale Feature Extraction: CNN branch is based on EfficientNet-B3, which is a model that extracts hierarchical visual representation on a range of abstraction levels. Let an OCT image be expressed as:

$$\boldsymbol{I} \in \boldsymbol{R}^{\boldsymbol{H}\times\boldsymbol{w}\times\boldsymbol{3}} \tag{1}$$

The CNN extracts three levels of features:

$$\mathbf{F_{cnn}} = \{\boldsymbol{f_1}, \boldsymbol{f_2}, \boldsymbol{f_3}\}, \boldsymbol{f_i} \varepsilon \boldsymbol{R^{C_i \times H_i x w_i}} \quad (2)$$

The features maps f are associated with sequentially deeper layers of the CNN and the superficial layers capture fine retinal textures and the deeper layers capture semantic features.
To provide coherence to these heterogenous features we use a Feature Pyramid Network (FPN), where deeper features are transmitted to high-resolution:

$$\boldsymbol{f_i} = \mathbf{Conv_{1\,x\,1}}(\boldsymbol{f_i}) + \mathbf{Upsample}(\boldsymbol{f_{i+1}}) \quad (3)$$

This is so that the fine spatial details are retained at the lower-level maps but with the entire map having a semantic context. Once we have fused, we use Global Average Pooling (GAP) to produce a small CNN representation.:

$$\mathbf{g_{cnn}} = \mathbf{GAP}(\boldsymbol{f_i}) \; \varepsilon \; \mathbf{R}^{\mathbf{dc}} \quad (4)$$

In this case, $\mathbf{g_{cnn}}$ creates a multi-scale representation of local retinal patterns, which plays an important role in the detection of lesions of different sizes.

Vision Transformer Branch for Global Context: Whereas CNNs are good at local information, they are unable to capture long-range dependencies in the retina. In order to deal with this, we use a Vision Transformer (ViT Small, patch size 16, image size 384), that subdivides the picture into nonoverlapping patches of size PxP. All patches are flattened and linearly projected:

$$\mathbf{x_i} = \mathbf{W_e} \,.\, \mathbf{Flatten}(\mathbf{I_i}), \;\; \mathbf{x_i} \, \varepsilon \, \mathbf{R}^{\mathbf{dc}} \quad (5)$$

These embeddings are collected into a sequence where a learnable class token xcls is introduced in the first position:

$$\mathbf{X} = [\mathbf{x_{cls}}, \mathbf{x_1}, \mathbf{x_2}, \ldots, \mathbf{x_N}] + \mathbf{E_{pos}} \quad (6)$$

where Epos provides positional encoding.
The sequence is transformed by the L transformer encoder layers; each layer is equipped with Multi-Head Self-Attention (MHSA):

$$\mathbf{Attention}(\mathbf{Q},\mathbf{K},\mathbf{V}) = \mathbf{softmax}\left(\frac{\mathbf{Q}\mathbf{k}^{\mathbf{T}}}{\sqrt{\mathbf{d_k}}}\right)\mathbf{v} \quad (7)$$

where Q = XWQ, K = XWK, V = XWv.

The mechanism of self-attention enables each patch to attend to all other patches so as to extract to the retina the global contextual information. The last result is obtained in the class token embedding following L layers:

$$\mathbf{g_{vit}} = \boldsymbol{x}_{cls}^{(L)} \in \boldsymbol{R}^{d_v} \tag{8}$$

Feature Fusion using Cross-Modal Attention: The simple concatenation of CNN and ViT features might not be effective to represent their relationship with one another. Thus, we add cross-modal attention to help in the interaction of features. The CNN features are used as queries, and ViT features are keys and values in this mechanism:

$$\mathbf{Z_{cnn}} = \mathbf{softmax}\left(\frac{W_q g_{Cnn}(W_k g_{vit})^T}{\sqrt{d}}\right) W_v g_{v_i t} \tag{9}$$

This gives CNN features the option to selectively attend to ViT features of global interest. On the same note, ViT can learn back to CNN representations:

$$\mathbf{Z_{vit}} = \mathbf{softmax}\left(\frac{W_q g_{vit}(W_k g_{cnn})^T}{\sqrt{d}}\right) W_v g_{cnn} \tag{10}$$

The final fused representation is obtained by concatenation:

$$\mathbf{g_{fusion}} = [\mathbf{Z_{cnn}} \,||\, \mathbf{Z_{vit}}] \tag{11}$$

This makes sure that the model learns local retinal texture as well as global structural structure resulting in a more informative joint representation.

Ensemble Learning for Robust Prediction: We implement three fusion strategies:

- Attention-based fusion
- Concatenation fusion:

$$\mathbf{g_{fusion}} = [\mathbf{g_{cnn}} \,||\, \mathbf{g_{vit}}] \tag{12}$$

- Weighted fusion:

$$\mathbf{g_{fusion}} = \alpha g_{cnn} + (1 - \alpha)\, \mathbf{g_{vit}},\ \ \alpha \in [0,1] \tag{13}$$

The logits of each strategy are output y1,y2,y3 To integrate them we apply a meta-learner which learns adaptive ensemble weights:

$$\mathbf{y_{final}} = \beta_1 \mathbf{y_1} + \beta_2 \mathbf{y_2} + \beta_3 \mathbf{y_3} \tag{14}$$

where $\beta 1$, $\beta 2$, $\beta 3$ are optimized during training. The class probabilities are obtained by softmax:

$$\mathbf{p(c|I)} = \mathbf{softmax}(\mathbf{y_{final}}) \tag{15}$$

Classification Layer: The fused representation is passed through a fully connected classifier:

$$\mathbf{y = W_2\sigma(W_1 g_{fusion} + b1) + b2} \quad (16)$$

where σ is the ReLU activation function. The final prediction corresponds to one of the seven disease classes: AMD, DME, ERM, NO, RAO, RVO, and VID.

4 Experimental Result Analysis

4.1 Evaluation metrics

After the model has been developed, a variety of performance measures are calculated in order to determine its effectiveness and give a detailed analysis of its performance. These measurements are the confusion matrix, accuracy, precision, recall, and F1-score, which, together, provide information about the capability of the model to classify the instances correctly and deal with the class imbalance. The mathematical formulas of both metrics are below to have a clear picture of how to calculate and use it in this study.

Confusion Matrix: The confusion matrix compares predicted and actual outcomes using: True Positives (TP), True Negatives (TN), False Positives (FP), and False Negatives (FN).

Accuracy: Measures the proportion of correctly classified instances.

$$\mathbf{Accuracy} = \frac{\mathbf{TP+TN}}{\mathbf{TP+TN+FP+FN}} \quad (17)$$

Precision: Indicates how many predicted positive cases are actually correct.

$$\mathbf{Precision} = \frac{\mathbf{TP}}{\mathbf{TP+FP}} \quad (18)$$

Recall (or Sensitivity): Shows how many actual positive cases were correctly identified.

$$\mathbf{Recall} = \frac{\mathbf{TP}}{\mathbf{TP+FN}} \quad (19)$$

F1 Score: Harmonic mean of precision and recall:

$$\mathbf{F1\ Score} = \mathbf{2} \times \frac{\mathbf{Precision \times Recall}}{\mathbf{Precision+Recall}} \quad (20)$$

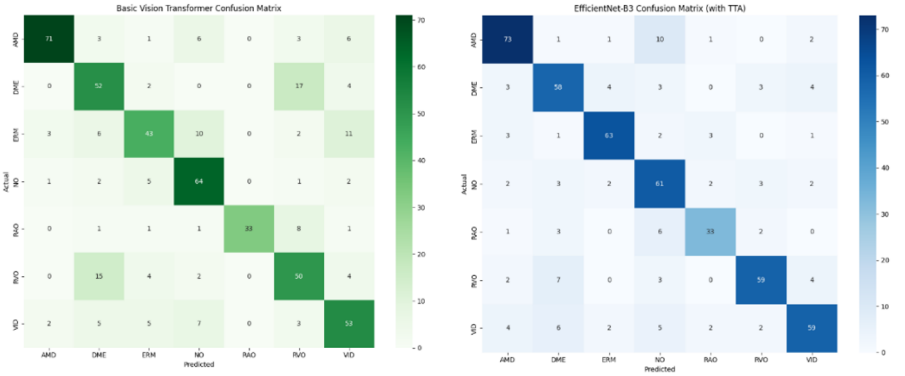
4.2 Experiment Results

In this research, accuracy, Precision, Recall, F1-score and AUC were used to evaluate the overall performance of all models. Our model surpasses state-of-the-art OCT classification methods, including EfficientNet-B3, Vision Transformer, and recent

hybrid architectures. Although EOCT reports higher raw accuracy in binary settings, our model provides superior robustness across all seven classes, particularly on minority categories. Table 4 presents a comparative performance analysis of the three models. EfficientNet-B3 achieves moderate results with an accuracy of 79% and balanced precision, recall, and F1-scores. The Vision Transformer has slightly lower performance with only 72 percent accuracy, lower recall and F1-score values. On the contrary, the proposed ensemble model substantially outperforms the two baselines, with 96% performance in all the evaluation metrics. These findings affirm that CNN and ViT features augmented with FPN and attention are a highly productive framework when it comes to retinal OCT disease classification.

Table 4. Comparative Result Analysis

	Precision	Recall	F1-score	Accuracy
EfficientNet-B3	0.80	0.79	0.79	0.79
Vision Transformer	0.74	0.72	0.72	0.72
Our proposed ensemble model	**0.96**	**0.96**	**0.96**	**0.96**



(a) (b)

(c)

Fig. 3. Confusion Matrix: (a) Vision Transformer, (b) EfficientNet-B3, and (c) Our proposed Ensemble Model

Fig. 3 shows the confusion matrices of the (a) Vision Transformer, (b) Efficientnet-B3 and (c) the proposed ensemble model. The Vision Transformer is strong in its performance but it exhibits significant misclassifications across most categories of diseases. EfficientNet-B3 is more balanced though it continues to have a problem with overlapping cases like RAO and RVO. The ensemble model minimizes misclassification and the predictions are concentrated much along the diagonal. This shows the high level of distinction of the ensemble between all the seven classes of retinal diseases at high reliability.

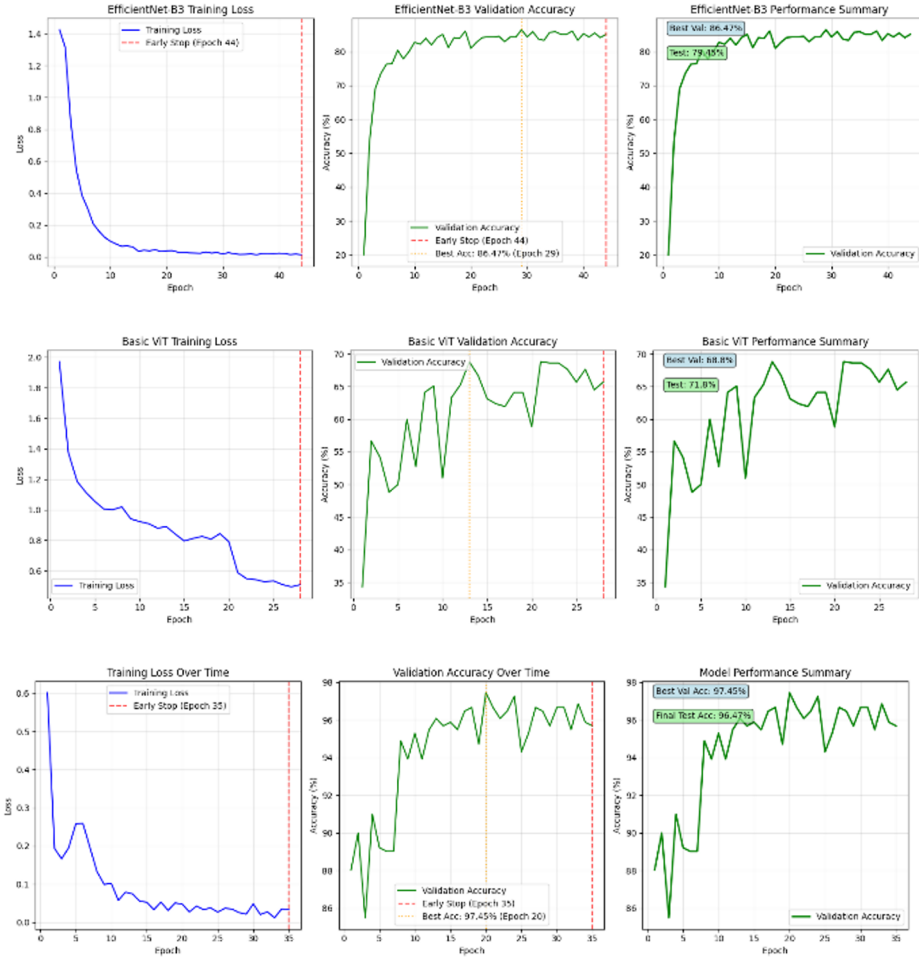


Fig. 4. Comparative Analysis (Loss Curve, Validation Accuracy, Performance Summary)

Fig. 4 shows the training losses, validation accuracies and performance summaries of EfficientNet-B3, Vision Transformer and the proposed ensemble model. EfficientNet B3 is observed to have a steep fall in training loss, and soon reaches a minimal value. It has a high convergence as its ability to validate steadily and level off at a good level. The Vision Transformer also minimizes loss and has more variations during training implying sensitivity to data variation. Its curve of validation is not as stable as the EfficientNet-B3, but the resultant final accuracy is nonetheless competitive. The ensemble model proposed has the most stable training behavior, and gradually decreases in loss and continuously high validation accuracy. The summary on the performance shows that the ensemble has the optimal balance between precision and stability across epochs. The presented comparative analysis shows that the ensemble does not only increase the overall accuracy but also provides strong generalization, in comparison with individual models.

5 Conclusion & Discussion

We have developed a new hybrid ensemble model in this work, which combines Convolutional neural networks, Vision transformer, Feature Pyramid Networks and cross-modal attention to classify retinal OCT disease. This work was inspired by the need to close the gap between local structural learning, that CNNs can process better, and global context modeling, which is the advantage of Vision Transformers. Through the synergistic combination of these two methods, the proposed model is able to learn discriminative representations that can both learn fine retinal structures and long-range dependencies. Our ensemble model was shown to outperform our baseline models (EfficientNet-B3 and Vision Transformer) in evaluation metrics on a consistent basis. The fact that accuracy, F1-score, and AUC have been improved proves the strength of the model and its possibility to be implemented in the real world in clinical practice.

In addition to the general performance, the confusion matrix analysis and performance curves also confirmed the reliability of the model in all the seven disease classes. The ensemble made fewer misclassifications than the baselines and produced more visually consistent predictions. This demonstrates the usefulness of multi scale feature extraction, attention-based fusion, and ensemble learning in the classification of medical images.

The results of the proposed framework are good, but the opportunities to continue further development still exist. Future directions would explore larger and more varied datasets obtained across several clinical settings in order to better generalize the results across populations. It may be more interpretable to the ophthalmologists by incorporating explainable artificial intelligence, including attention heatmaps or saliency maps. The other potential avenue is the combination of multimodal data, i.e. fundus images and patient demographics to complement OCT scans and increase their diagnostic accuracy. In addition, the implementation of small-sized model variants optimized to mobile or point-of care devices has the potential to broaden the availability of automated retinal screening in resource-constrained settings.

References

1. Alizadeh Eghtedar R, Vard A, et al. (2024) A new computer-aided diagnosis tool based on deep learning methods for automatic detection of retinal disorders from OCT images. International Ophthalmology 44(1):110. https://doi.org/10.1007/s10792-024-03033-9
2. Araújo T, Aresta G, et al. (2023) Few-shot out-of-distribution detection for automated screening in retinal OCT images using deep learning. Scientific Reports 13(1). https://doi.org/10.1038/s41598-023-43018-9
3. Baba S, Kumari P, et al. (2024) Retinal Disease Classification Using Custom CNN Model From OCT Images. Procedia Computer Science 235:3142–3152. https://doi.org/10.1016/j.procs.2024.04.297
4. Tareq Babaqi, Jaradat M, et al. (2023) Eye Disease Classification Using Deep Learning Techniques. arXiv (Cornell University). https://doi.org/10.48550/arxiv.2307.10501
5. Chiang JN, Corradetti G, et al. (2022) Automated Identification of Incomplete and Complete Retinal Epithelial Pigment and Outer Retinal Atrophy Using Machine Learning. Ophthalmology Retina. https://doi.org/10.1016/j.oret.2022.08.016
6. Dahiya R, Agarwal N, et al. (2024) Diabetic Retinopathy Eye Disease Detection Using Machine Learning. EAI Endorsed Transactions on Internet of Things 10. https://doi.org/10.4108/eetiot.5349
7. Dai H, Yang Y, et al. (2024) Improving retinal OCT image classification accuracy using medical pre-training and sample replication methods. Biomedical Signal Processing and Control 91:106019. https://doi.org/10.1016/j.bspc.2024.106019
8. Hassan E, Samir Elmougy, et al. (2023) Enhanced Deep Learning Model for Classification of Retinal Optical Coherence Tomography Images. Sensors 23(12):5393–5393. https://doi.org/10.3390/s23125393
9. Jaimes WJ, Arenas WJ, et al. (2025) Detection of retinal diseases from OCT images using a VGG16 and transfer learning. Deleted Journal 7(3). https://doi.org/10.1007/s42452-025-06565-6
10. Kulyabin M, Zhdanov A, et al. (2024) OCTDL: Optical Coherence Tomography Dataset for Image-Based Deep Learning Methods. Scientific Data 11(1):365. https://doi.org/10.1038/s41597-024-03182-7
11. Ma F, Li S, et al. (2023) Deep-learning segmentation method for optical coherence tomography angiography in ophthalmology. Journal of Biophotonics 17(2). https://doi.org/10.1002/jbio.202300321
12. Mariottoni EB, Datta S, et al. (2022) Deep Learning–Assisted Detection of Glaucoma Progression in Spectral-Domain OCT. Ophthalmology Glaucoma. https://doi.org/10.1016/j.ogla.2022.11.004
13. Oh D, Moon J, et al. (2024) GCN-assisted attention-guided UNet for automated retinal OCT segmentation. Expert Systems with Applications 249:123620. https://doi.org/10.1016/j.eswa.2024.123620
14. Pekala M, Joshi N, et al. (2019) Deep learning based retinal OCT segmentation. Computers in Biology and Medicine 114:103445. https://doi.org/10.1016/j.compbiomed.2019.103445

15. Yoon J, Han J, et al. (2020) Optical coherence tomography-based deep-learning model for detecting central serous chorioretinopathy. Scientific Reports 10(1):18852. https://doi.org/10.1038/s41598-020-75816-w
16. Islam, R., Akash, R.S., Rony, M.A.H., Hasan, M.Z.: SAMU-Net: A dual-stage polyp segmentation network with a custom attention-based U-Net and segment anything model for enhanced mask prediction. Array 24, 100370 (2024). https://doi.org/10.1016/j.array.2024.100370
17. Kulyabin, M., Zhdanov, A., Nagaichuk, V., et al.: OCTDL: Optical Coherence Tomography Dataset for Image-Based Deep Learning Methods. Mendeley Data, V4 (2024). https://doi.org/10.17632/sncdhf53xc.4