Indonesian Journal of Electrical Engineering and Computer Science
Vol. 43, No. 2, August 2026, pp. 495~506
ISSN: 2502-4752, DOI: 10.11591/ijeecs.v43.i2.pp495-506

OCTModNet: a deep learning-based framework for optical coherence tomography image multi-class classification

Saja Ataallah Muhammed[1], Abbas M. Ali[2], Dler Salih Hasan[1]
[1]Department of Computer Science and Information Technology, College of Science, Salahaddin University-Erbil, Kurdistan Region, Iraq
[2]Department of Software Engineering, Faculty of Engineering, Salahaddin University-Erbil, Kurdistan Region, Iraq

Article Info

Article history:

Received Dec 30, 2025
Revised Jun 16, 2026
Accepted Jun 27, 2026

Keywords:

Deep learning
Medical image analysis
Multi-class classification
OCT image classification
Retinal disease

ABSTRACT

Sight is one of the most significant senses in human beings. Losing sight can change a human's life dramatically. Early diagnosis and detection of retinal diseases will prevent sight loss. Optical coherence tomography (OCT) imaging is helpful in ocular imaging with its high resolution and non-invasive imaging for diagnosing retinal diseases. This study proposes OCTModNet, a novel multi-class hybrid classification model that is capable of classifying seven different retinal diseases using a combination of publicly available datasets, Kaggle OCT-C8, and OCTDL datasets. For effective extraction of the significant features, the collected data is resized and normalized. Augmentation techniques are used to improve the training and learning process. For model development, we applied transfer learning and fine-tuned several state-of-the-art deep learning models, ResNet50, InceptionV3, DenseNet121, VGG16, and EfficientNetV2S. Two models that performed best, VGG16 and EfficientNetV2S, were selected as the model backbones with integration of a squeeze and excitation block. The results showed that the OCTModNet achieved an outstanding performance with 99% test accuracy, 99% precision, 99% recall, 99% F1-score, and an area under the curve (AUC) as 99% across the unseen data. These results point out the robustness and reliability of the proposed model to classify OCT images and have the potential to enhance clinical decision-making and assist ophthalmologists in the early detection of diseases.

Corresponding Author:

Saja Ataallah Muhammed
Department of Computer Science and Information Technology, College of Science
Salahaddin University-Erbil
Kurdistan Region, Iraq
Email: saja.muhammed@su.edu.krd

1. INTRODUCTION

The retina is a thin, layered structure of cells at the posterior part of the eyeball. This thin layer acts as the bridge between the light received by the eyes and the images we see. The light received is converted into signals and passed to the brain through the optic nerve to be processed. Any damage to these layer cells could affect the vision quality or could cause permanent vision loss. Many techniques were introduced over the years to capture the layered structure of the retina to determine whether it's normal or not. One of these helpful techniques is optical coherence tomography (OCT) imaging, which is a vital non-invasive imaging modality in diagnosing, treating, and following many retinal diseases (e.g., diabetic retinopathy (DR), epiretinal membrane (ERM), drusen, diabetic macular edema (DME), macular hole (MH), and central serous retinopathy (CSR)). Early diagnosis and tracking of ocular diseases are realized by OCT.

Nonetheless, classifying OCT image modalities is still challenging because of diverse image quality, acquisition protocol, and imbalanced datasets. Although machine and deep learning models have shown significant promise in automatically classifying retinal diseases, they have largely failed to enable differentiation between OCT modalities, thereby limiting their potential for automated diagnosis and clinical decision-making [1], [2]. Deep learning-based approaches for OCT image analysis have gained interest recently and have been studied in a few recent publications [3]. With improved classification performance in OCT images, contrastive learning was used to enhance representation learning [4]. To overcome the limitation of the small number of labelled images corresponding to rare retinal diseases [5], few-shot learning methods have been proposed. Moreover, deep convolutional neural networks (CNNs) like ResNet, InceptionV3, and DenseNet have been approached for retinal disease classification [6], [7]. On the other hand, these models have a single feature extraction mechanism with limitations to the higher hierarchical representation of OCT images. Not featuring adaptive learning rate strategies also means poor and less stable convergence. These drawbacks emphasize the urgency of proposing a deep learning framework using an integrated hybrid-model feature extractor, advanced feature fuses, and optimized training strategies.

This research presents a novel hybrid deep learning-based framework called OCTModNet for OCT image modality classification. Fusion of EfficientNetV2S and VGG16 for multi-scale feature extraction and the use of attention mechanism, squeeze and excitation block (SE) to improve classification accuracy and robustness. It uses a feature fusion strategy to combine the extracted features, improving discriminative ability. Besides, the learning rate schedule was implemented, as the cosine decay restart was applied to the optimizer to stabilize convergence better and prevent overfitting. By combining different state-of-the-art deep learning architectures with an adaptive optimization strategy and augmented feature processing methods, we demonstrate improved classification performance compared to state-of-the-art architectures, which is the novelty of this work. This work presents a structured and systematic framework characterized by several key contributions:

a) Hybrid multi-backbone integration: a dual-backbone architecture combining two EfficientNetV2S and VGG16 to examine complementary multi-scale feature extraction and improve discriminative capability in OCT retinal imaging.
b) Attention-enhanced feature refinement: to enhance feature recalibration and robustness to subtle retinal structural variations. SE channel attention mechanism was integrated and assessed.
c) Comprehensive comparative evaluation: a systematic evaluation of widely used backbone models such as ResNet50, InceptionV3, DenseNet121, VGG16, and EfficientNetV2S, to position the proposed framework within existing OCT classification methodologies.
d) Optimized training strategy for robust convergence: the deployment and empirical validation of adaptive optimization using Nadam with cosine decay restarts to improve convergence stability and generalization across heterogeneous OCT datasets.

The clinical significance and the unique contribution of this study can be summarized as the introduction of a novel hybrid framework that fuses EfficientNetV2S and VGG16 backbones integrated with SE attention for the first time for the classification of seven different retinal diseases. By combining complementary feature extraction mechanisms with attention-driven channel recalibration and a structured feature fusion strategy, the proposed OCTModNet enhances discriminative performance for subtle retinal abnormalities. The integration of adaptive optimization through cosine decay restart scheduling and Nadam further improves training stability and generalization under heterogeneous and partially imbalanced multi-dataset conditions. Clinically, the framework supports comprehensive automated screening of multiple retinal diseases within a unified system, offering potential as a robust decision-support tool for ophthalmic diagnostics and real-world OCT deployment.

The structure of the remainder of this paper is as follows. Section 2 is a tabulated literature review. The third section contains the overview of the proposed method, which comprises the dataset preprocessing, deep learning architecture, and training strategies. The experimental results are reported in the 4th section. The 5th Section summarizes and concludes the work with a discussion of the implications of the findings.

2. LITERATURE REVIEW

OCT has emerged as a key imaging modality in assessing retinal disease due to its non-invasive, high-resolution cross-sectional retina imaging. Lately, deep learning has witnessed a rapid development; automated analysis of OCT images is making further strides in efficiency and accuracy in clinical diagnosis. OCT image classification has been approached with different methodologies, such as CNNs, transfer learning, GANs, and ensemble learning.

Baba *et al.* [8] introduced a custom convolutional model for macular disorder classification, achieving 97% training accuracy, 93% validation accuracy, and a test accuracy reached up to 98% on the

Kermany dataset. Talaat *et al.* [9] demonstrated that attention models do not always reach near-perfect performance [9]. They proposed an attention-augmented DensNet for retinal disease diagnosis; the study reported validation accuracy of 0.9167 [9]. Miladinović *et al.* [10] highlighted challenges in OCT classification performance. They showed that different classifiers have performed differently under data scarcity and label noise [10]. In this study, experiments proved that Inception V3 was the best performer, registering performance greater than 90% [10]. In another study on retinopathy detection using VGG16 and semi-supervised learning [11], proposed a network that achieved around 94% and 93% on small OCT datasets, stating that performance can be considerably below 98% in limited data settings [11]. A recent study has performed Xception and InceptionV3 with specific augmentation techniques on the OCT-C8 dataset, and both networks registered a good classification accuracy. The Inception model achieved 94.82%, while the Xception model performed better by registering 95.25 across the 8 classes [12]. Some of the studies that carried out research on different OCT datasets for different disease detection and classification have been reviewed and summarized in Table 1.

Table 1. Summary of literature review findings

Ref.	Year	Methodology	Dataset scope	Identified limitations
Parra-Mora *et al.* [3]	2021	Deep CNN for ERM detection	Local Dataset (ERM only)	Binary classification, single disease detection, no feature fusion or attention modules for broader pathology representation.
Yoo *et al.* [5]	2021	Few-shot learning framework	Kermany2018	Single feature extractor which restricts representational diversity; few-shot learning rather than large-scale multi-class classification; no fusion or attention modules.
Altan [7]	2022	Explainable CNN	Disease-specific OCT dataset	Specific disease category; lacks multi-backbone representation learning; static training pipeline without adaptive learning rate strategies or feature fusion.
Baba *et al.* [8]	2024	Custom CNN Architecture	Kermany2018	Single-backbone architecture that limits multi-scale feature extraction; lacks attention mechanisms to emphasize pathology-relevant retinal layers; fixed learning rate training strategy reduces convergence efficiency and generalization.
Talaat *et al.* [9]	2024	Attention-based DenseNet	Kermany2018	Validation accuracy ~91.7%; Although attention is used, the architecture relies on a single backbone, restricted complementary feature representation, conventional LR scheduling without adaptive optimization.
Miladinović *et al.* [10]	2024	Comparative deep learning evaluation	Heterogeneous OCT datasets	Focuses on evaluation rather than architectural innovation; no feature fusion, attention integration, or adaptive optimization to address domain shift across datasets.
Luo *et al.* [11]	2021	VGG16 with semi-supervised learning	Kermany2018	~94% accuracy; single architecture; lacks attention modules for discriminative feature weighting; no multi-scale or multi-backbone feature fusion strategy.
Noor *et al.* [12]	2026	Transfer learning (Xception, InceptionV3)	Multi-class OCT-C8 dataset	~94–95% accuracy; single-backbone models with no feature fusion; absence of attention mechanisms reduces the ability to emphasize disease-specific regions; employs static optimization strategies without dynamic learning rate adaptation.

Previously reviewed OCT classification studies have demonstrated good performance using single backbone CNNs (e.g., VGG16, Inception, Xception, and DensNet) [3]-[12]. For instance, Parra-Mora *et al.* [3] proposed a deep CNN architecture for ERM detection using OCT images; despite their approach focused on single-disease classification using a feature extractor. Similarly, Yoo *et al.* [5] investigated few-shot learning strategies to improve OCT diagnosis of rare retinal diseases, yet their proposed framework relied on a single backbone architecture without multi-model feature fusion. While Altan [7] introduced DeepOCT, an explainable CNN model for macular edema analysis, their study remains disease-specific and did not incorporate complementary multi-scale feature extraction or adaptive learning rate scheduling.

However, most existing frameworks rely on a single feature extractor, which can limit the capture of complementary retinal patterns across scale and textures, particularly when class boundaries are subtle. In addition, training on static learning rates or convolutional schedules may result in unstable convergence and suboptimal generalization under heterogeneous OCT acquisitions. Conversely, OCTModNet addresses these limitations through dual-backbone fusion (EfficientNetV2S and VGG16) to learn complementary multi-scale features. Moreover, to recalibrate informative retinal responses, an SE-block channel attention was used, and an adaptive learning rate strategy was used (cosine decay restarts with Nadam) to enhance convergence stability and robustness. This combination advances beyond prior single-stream OCT classifiers by explicitly integrating multi-model feature fusion and adaptive optimization within a unified seven-class OCT disease classification framework.

3. MATERIALS AND METHODS

3.1. Dataset description

A mix of two high-quality retinal OCT datasets from the Kaggle OCT-C8 dataset and the OCTDL has been used to develop the model and test it for the diagnosis of retinal diseases [13], [14]. Combining samples from these two datasets introduces inter-dataset variability, which normally has different acquisition protocols, different scanners, and demographic diversity. This mix will improve the model's robustness and generalization capacity. Six different classes from the OCT-C8 dataset have been chosen; each class has DR, CSR, DME, DRUSEN, MH, and NORMAL. The dataset is split into train, validation, and test sets [13]. Images from this dataset are not equal in size; the size of these images differs. The ERM class was chosen from the OCTDL dataset to add diversity; the sizes of images in this dataset are not equal either [14]. Different samples are shown in Figure 1. Data distribution across the seven different classes is shown in Table 2.

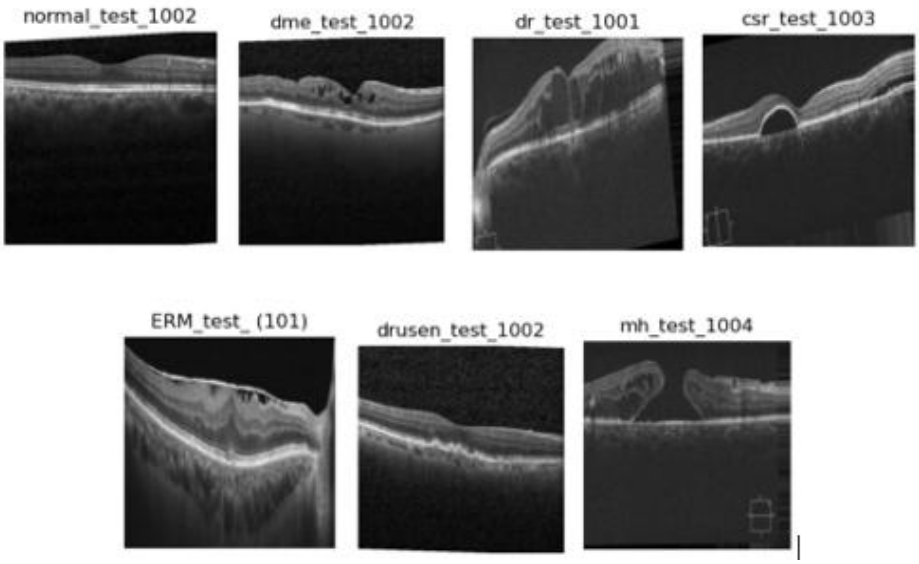


Figure 1. Samples from 7-class dataset

Table 2. Sample distribution across the 7-class dataset

Class	Tarin	Validation	Test	Total per class
CSR	2,300	350	350	3,000
DME	2,300	350	350	3,000
DR	2,300	350	350	3,000
DRUSEN	2,300	350	350	3,000
Macula Hole	2,300	350	350	3,000
Normal	2,300	350	350	3,000
ERM	2,353	490	483	3,326
Total	16, 153	2, 590	2,583	21,326

To assess the generalization of the proposed model and mitigate dataset-specific bias, the model was evaluated on two dependent datasets: Kermay2018 [15] and OCT-C8 [13] datasets. The Kermany2018 consists of approximately 207,130 images, collected from different patients and devices between 2013 and 2017. The distribution of samples across the four classes is (CNV(n=37.2k), DME(n=11,3k), DRUSEN(n=8,616), and NORMAL(n=26,3k)) in the training subset, while the validation set contains 250 images per class, and the test set 242 samples per each of the four classes.

Experiments were extended to the retinal OCT-C8 dataset. It consists of 8 classes, each with 4,600 images: Age-Related Macular Oedema, CNV, CSR, DME, DR, DRUSEN, MH, and NORMAL. The validation and test subsets have 700 images/class. The standard OCT classification setting has been expanded in this dataset to comprise eight retinal disorder categories, wrapping both pathological and normal conditions. The inter-class visual similarity and the elevated number of classes make this dataset more difficult and challenging than the Kermany2018 dataset.

3.2. Data preprocessing

3.2.1. Image resizing

All the images in the mixed dataset were resized to have the exact dimensions and, hence, are uniform and ready to be utilized to feed into the deep learning models. Images are resized to 256×256 pixels, keeping the aspect ratio due to distortion. Resizing images to a standard resolution '256×256' is crucial in the development of the model, allowing an effective and uniform process of images by the convolution neural networks.

3.2.2. Normalization

It is the process of converting pixel values (intensities) into a consistent range. Normalizing pixel intensity values helps the network to learn image patterns more efficiently and accurately. In this study, all intensity values of image pixels were scaled from [0,255] to [0, 1] using (1).

$$l_n = I_0/255 \quad (1)$$

3.2.3. Data augmentation

Due to the fact that huge amounts of data are required by deep learning models to be able to predict and infer correctly, working on medical images poses a challenge because of data scarcity and availability. Researchers had to use Augmentation techniques to boost the amount of data and improve the learning process and model generalization ability to recognize new unseen data. The usefulness of these techniques depends on the task at hand or the problem to be solved. In this work, several augmentation techniques were used, as listed in Table 3.

Table 3. Augmentation techniques used in the study

Augmentation method	Description
Rotation	Rotates the input image randomly by angle [-30, 30] degrees
Horizontal flipping	Flips the input image around the horizontal axis, performing a mirror inversion
Zooming	Randomly zooms the input image in or out.
Shear range	The input image will be distorted along the horizontal axis.
Fill mode	Fills the empty areas (borders) caused by flipping or rotation with the nearest valid value
Brightness and contrast adjustment	Updates the contrast and the brightness of the input image

3.3. Proposed architecture

This study proposes OCTModNet, a hybrid deep learning framework designed for multi-class retinal OCT image classification. The framework brings together two CNN backbones, EfficientNetV2S and VGG16, operating in parallel to extract complementary feature representations from the same input image. An SE attention block is integrated within the EfficientNetV2S branch to enhance channel-wise feature importance. The extracted features from the two branches are then fused through concatenation and passed to a fully connected classification head. The overall architecture of the model is illustrated in Figure 2.

3.4. Backbone model selection

To identify suitable backbone architectures, five different, widely used CNN architectures were first evaluated on the dataset: ResNet50, InceptionV3, DenseNet121, VGG16, and EfficientNetV2S. Preliminary experiments showed that EfficientNetV2S performed better than the other models by achieving the strongest classification performance, as shown in Table 4. Although DenseNet121 was the second best performer, but VGG16 was selected instead for its complementary architectural characteristics. VGG16 employs a simple and sequential convolutional design that effectively captures fine-grained spatial patterns and texture information in OCT images. In contrast, EfficientNetV2S focuses on efficient hierarchical feature extraction through optimized convolutional blocks and compound scaling.

Combining these two architectures introduces representational diversity, enabling the proposed model to learn both detailed local structure and globally optimized feature representations. This complementary combination improves the effectiveness of the feature fusion stage and contributes to the overall performance improvement of the OCTModNet model. Furthermore, combining different design architectures has been shown to improve the performance of hybrid models by reducing feature redundancy and increasing representational diversity in multi-backbone networks [16].

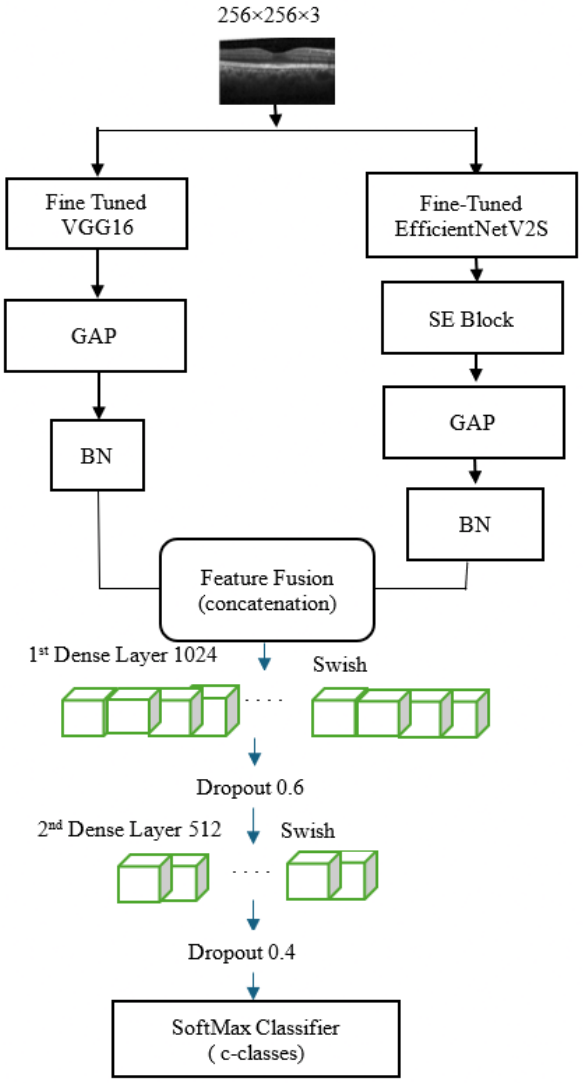


Figure 2. OCTModNet architecture

3.5. EfficientNetV2S

A convolution neural network, an improved, smaller version of the EfficientNet family. The main architecture block of this network is MBConv and Fused-MBConv. The main idea behind using Fused-MBConv is to improve the training speed [17]. The main modification made was removing the fully connected layer (FC) to be used only as a feature extractor, allowing the architecture to extract high-level visual features such as textures, edges, and structural patterns relevant to retinal OCT images. The model was fine-tuned to adapt its learned representations to domain-specific retinal structures [17].

3.6. Squeeze-and-excitation block

The main goal behind using SE-block was to improve the representation quality that a network introduces. This block allows feature recalibration, which enables the learning process to carefully highlight features that are informative and filter out the less informative or less useful ones [18]. The SE-block has two main phases: the squeeze phase and the excitation phase. In the squeeze phase, the model generates channel wise statistics through globally pooling features across the spatial dimensions to capture the most significant information for each channel. In the second phase, the excitation phase, the learned channel wise statistics from the first phase will be utilized to produce attention weights [19]. These weights represent the influence of each channel on the final decision or representation. Through this process, the network will be enabled to give more attention to the most relevant or most informative channels and block out the less informative channels [19].

3.7. VGG16

It is a type of CNN introduced by the Visual Geometry Group in 2014. It became one of the architectures that gained significant popularity and was deployed in many vision tasks due to its effectiveness, simplicity, and transferability. The main goal of the proposed architecture was to demonstrate the critical role of the depth of the network in achieving high performance in vision tasks. The total layer number of this model is 16, consisting of 13 convolutions and three FC layers. The design of the network improves its capacity to extract complex features from the input image [20].

In the current work, a pretrained VGG16 model on ImageNet is used as the second backbone network. The main modification made to this model was removing the top classifier layer of the model, allowing the model to operate as a feature extractor. The extracted feature maps are then processed using a global average pooling (GAP) layer, followed by batch normalization (BN) to produce a compact feature representation before the feature fusion stage. The main goal of using this architecture is to capture fine-grained spatial patterns in OCT images, thus complementing the hierarchical feature representations learned by EfficientNetV2S in the other branch of the framework [20].

3.8. Feature fusion

The resulting tensors extracted from the two different backbones were concatenated. Each of these two feature vectors was processed independently and passed through the different layers of these two different base models. The knowledge obtained from these branches was combined. Each of these two models has learned a different set of patterns from the same input. The fusion of these feature maps by concatenation will allow the next FC layer to learn from these combined feature representations to assist in the process of the final decision-making by the classifier layer.

3.9. Classifier head

The final classification decision is made by the classification head that receives the fused feature maps from the two base models and decides to which class the input Retinal OCT image belongs. The classification head consists of two dense layers with 1024 and 512 units with swish activation. Each dense layer is followed by a dropout layer by 0.6, and 0.4, respectively to prevent model overfitting. The final multi-class classification was made by the SoftMax layer. This layer converts the output from the last model layer into a probability distribution. Where each value in this probability distribution represents the probability of belonging the input Retinal OCT image into one of the seven classes. The higher value represents the true class of the input image.

3.10. Handling class imbalance

To handle the class imbalance in the (7-class and Kermany2018)datasets, class weighting was applied during training of the framework. Every class received a weight inversely proportional to its frequency in the training set. By ensuring that underrepresented classes contribute proportionately to the loss function, this method keeps the model from being skewed in favour of majority classes. The overall workflow of the proposed framework is shown in Algorithm 1.

Algorithm 1. OCTModNet pipeline for OCT image classification

Input: OCT image dataset D, number of classes C, learning rate α, dropout p.
Output:
Predicted class $\hat{y}$.
Steps:

1. **Preprocessing:**
   - Resize images to 256×256, normalize pixel values, and apply augmentation
2. **Feature Extraction:**
   - Compute feature map $F_E = \mathcal{F}_E(I)$ using EfficientNetV2S backbone.
   - Apply SE-block to recalibrate F_E, Then use Global Average Pooling (GAP) and Batch Normalization to obtain v_E
   - Compute feature map $F_V = \mathcal{F}_V(I)$ using VGG16, followed by GAP and Batch Normalization to obtain v_v
3. **Feature Fusion**
   - Concatenate the extracted feature vectors: $F_{concat} = [v_E; v_v]$.
4. **Classification:**
   - Pass F_{concat} through the Dense(1024, swish) + Dropout(0.6), then Dense(512, swish) + Dropout(0.4).
   - Apply SoftMax to produce class probabilities and assign the final label.

5. **Training:**
   - Train the model using categorical cross-entropy loss, class weighting, and update weights using the Nadam optimizer.
   - Apply Cosine Decay Restart for learning rate.
   - Train for 50 epochs with batch size 16.

6. **Evaluation:**

Validate on D_{val}, test on D_{test}, save the trained model M.

4. RESULTS AND DISCUSSION

4.1. Experimental setup

The model was trained on a system with an Intel Core i9-12900K CPU, 64GB RAM, and an NVIDIA RTX 3090 (24GB VRAM), which was used for all of the training and validation processes. Python 3.8 was used for training, while TensorFlow 2.9 with the Keras backend was used for the deep learning framework implementation. The dataset was processed using OpenCV and NumPy, and performance was visualised using Matplotlib. Pixel intensities were normalised within the interval [0,1], and all OCT pictures were converted to 256×256×3 RGB format. The proposed framework was trained for 50 epochs, batch size equal to 32, using the Nadam Optimization technique with an initial learning rate of 0.0001.

4.2. Evaluation metrics

To evaluate the performance of OCTModNet for OCT image classification, several performance metrics were used to assess both overall and class-wise classification behavior. For a multi-class classification problem with N samples and C classes, standard classification metrics were adopted [21], [22] in (2)-(5) express the metrics used for model evaluation:

The overall classification accuracy is defined as:

$$Accuracy = \frac{\sum_{i=1}^{N} 1(\hat{y}_i = y_i)}{N} \tag{2}$$

Precision, recall, and F1-score were calculated per class as follows [21], [23].

$$Precision = \frac{TP}{TP+FP} \tag{3}$$

$$Recall = \frac{TP}{TP+FN} \tag{4}$$

$$F1 - score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \tag{5}$$

For multi-class evaluation, macro-averaging was adopted to ensure equal contribution of each class irrespective of class imbalance [23].

For further assessing discriminative capability, the area under the receiver operating characteristic curve (AUC-ROC) was computed using the One-vs-Rest strategy [24], [25]. For each class c, a binary ROC curve was constructed by treating class c as positive and all other classes as negative. The class-wise area under the curve (AUC) values were then aggregated using macro averaging (13) and (14).

$$AUC = \int_0^1 TPR(FPR)d(FPR)$$

$$AUC_{macro} = \frac{1}{C}\sum_{c=1}^{C} AUC_c$$

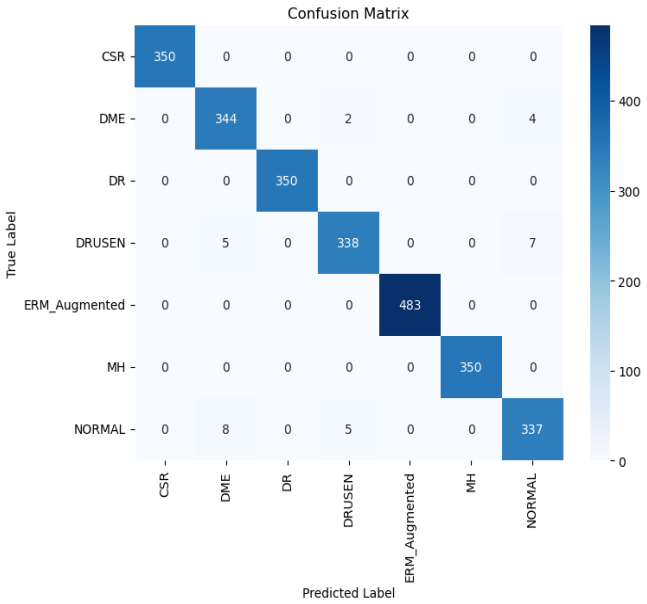
All evaluation metrics were computed on the independent test set that had not been used during training the model or hyperparameter tuning.

4.3. Results

OCTModNet was evaluated on an independent test set consisting of seven different retinal disorders. The proposed model achieved an overall test accuracy of 99% and a macro-averaged AUC of 99%, outperforming the evaluated single-backbone models (RESNet50, InceptionV3, DenseNet121, VGG16, and EfficientNetV2S) under identical training and validation conditions.

Class-wise precision, recall, and F1-scores (see Figure 3) show consistent performance across all categories, indicating balanced discriminative capability despite moderate class imbalance. The confusion matrix, Figure 3, further confirmed limited inter-class misclassification, especially among visually similar classes. Multi-class ROC analysis showed high sperability across all classes, with macro-averaged AUC reflecting robust generalization within the tested dataset. On the other hand, as mentioned earlier, to improve generalization, the proposed model was trained on Kermany2018 and registered an overall accuracy of 99%, and macro-averaged precision, recall, and F1-score of 0.99, proving balanced class-wise discrimination Figure 4.

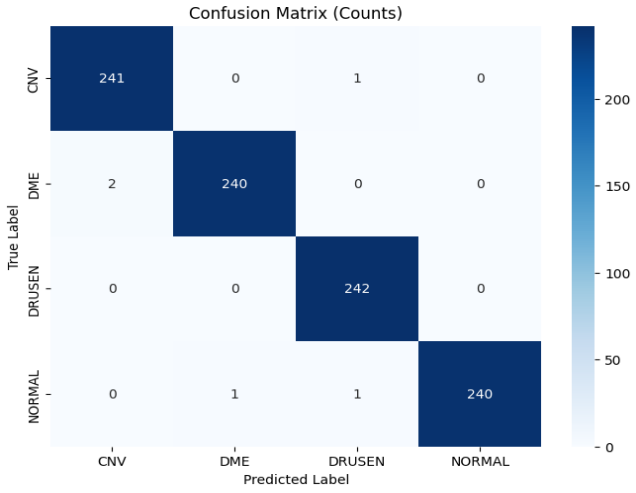
On the OCT-C8 dataset, which contains a more complex 8-class distribution, the model achieved 98% test accuracy on the test set, with a macro-averaged F1-score of 0.98. minor variations observed among visually overlapping classes, while the overall performance remained stable, as can be noticed in Figure 5. These findings suggest that the integration of complementary feature extractors and adaptive optimization improves classification stability relative to single-stream architectures, and that the dual-backbone architecture maintains consistent discriminative capability across datasets with different class structures and distributions.



Classification Report:

	precision	recall	f1-score	support
CSR	1.00	1.00	1.00	350
DME	0.96	0.98	0.97	350
DR	1.00	1.00	1.00	350
DRUSEN	0.98	0.97	0.97	350
ERM_Augmented	1.00	1.00	1.00	483
MH	1.00	1.00	1.00	350
NORMAL	0.97	0.96	0.97	350
accuracy			0.99	2583
macro avg	0.99	0.99	0.99	2583
weighted avg	0.99	0.99	0.99	2583

Figure 3. OCTModNet confusion matrix and classification report on 7-class dataset



	precision	recall	f1-score	support
CNV	0.99	1.00	0.99	242
DME	1.00	0.99	0.99	242
DRUSEN	0.99	1.00	1.00	242
NORMAL	1.00	0.99	1.00	242
accuracy			0.99	968
macro avg	0.99	0.99	0.99	968
weighted avg	0.99	0.99	0.99	968

Figure 4. Confusion matrix and classification report across Kermany2018

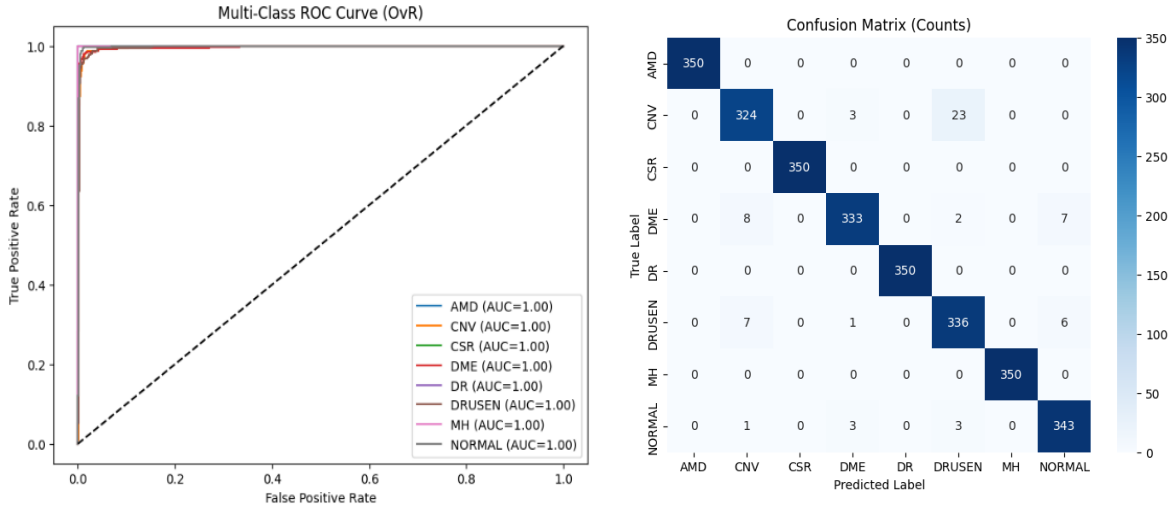


Figure 5. Model's ROC curve and confusion matrix across the OCT-C8 dataset

4.4. Comparison with baseline models

The comparative analysis assesses OCTModNet based on state-of-the-art, single backbone deep learning architectures such as ResNet50, InceptionV3, DenseNet121, VGG16, and EfficientNetV2S (see Table 4). The performances indicate that OCTModNet's test accuracy and macro-averaged AUC outperform the previous methods, depicting the benefits of optimizing multi-model feature extraction, fusion, and training strategy. This underlines its better classification performance compared to traditional single-model architectures. In comparison, the strongest single-backbone model (EfficientNetV2S) achieved 96.05% test accuracy and 97% AUC. Improvements were also observed across macro-averaged precision, recall, and F1-score metrics, indicating balanced class-wise performance rather than gains limited to specific categories.

Table 4. Comparison with baseline models

Model	Test Acc. (%)	Precision (%)	Recall (%)	F1-Score (%)	AUC (%)
ResNet50	92.85	91.60	92.30	91.95	94.10
InceptionV3	94.12	93.80	93.95	93.87	95.25
DenseNet121	95.37	95.00	95.15	95.07	96.45
VGG16	93.25	92.90	93.10	93.00	94.35
EfficientNetV2S	96.05	95.80	96.00	95.90	97.00
OCTModNet	99.0	99.0	99.0	99.0	99.0

4.5. Ablation study

Table 5 reveals that the framework's full configuration has registered a peak performance (99%) in terms of test accuracy and 0.99 AUC. Excluding individual components has degraded the performance and accuracy of the model. The synergy between complementary feature integration and the optimized learning schedules is essential for achieving refined, high-level classification results.

Table 5. Ablation study

Model configuration	EfficientNetV2S	VGG16	Feature fusion	Cosine decay restart	Adaptive feature processing	Test Acc. (%)	AUC (%)
Only VGG16	✗	✓	✓	✓	✓	91.84	94.2
Only EfficientNetV2S	✓	✗	✓	✓	✓	93.62	95.8
Without feature fusion	✓	✓	✗	✓	✓	94.27	96.1
Without cosine decay	✓	✓	✓	✗	✓	95.31	96.9
Without adaptive feature processing	✓	✓	✓	✓	✗	96.42	97.5
Full model OCTModNet	✓	✓	✓	✓	✓	98.8	99

4.6. Discussion

Fusing the EfficientNetV2S for capturing global hierarchical representations with VGG16's stable local texture extraction has enabled robust detection of both contextual structure and subtle local texture extraction. Significant challenges are imposed by the use of different devices at different eye examining centers and different acquisition techniques, making the OCT classification problem non-trivial. These challenges are addressed by the OCTModNet.

Performance evaluation demonstrates consistently high precision, recall, and F1-score throughout most retinal disorders and diseases, along with confusion matrices showing strong class distinction and separability. Despite this high performance, the misclassification cases occur among classes that have overlapping structural features in early stages, such as the DME and DRUSEN. Notably, OCTModNet preserves stable macro-level performance across multiple datasets with varying class configurations.

Regardless of the high performance of the model even across different datasets, the model suffers from limitations like validation on multi-center clinical datasets to confirm real-world applicability. Plus, the increased complexity due to the dual backbone architecture results in 37 million trainable parameters and an average inference time of 48ms per image. Thereafter, future studies should focus on validating the proposed framework on multi-center clinical datasets and exploring lightweight model variants to improve computational efficiency.

5. CONCLUSION

This paper proposed OCTModNet, a hybrid deep learning framework that combines two complimentary feature extraction models, EfficientNetV2S and VGG16, to classify multi-class OCT images. The study's suggested architecture uses an attention-based approach to refine the features fused from the two integrated feature extraction models, as well as adaptive learning rate scheduling to optimise performance. Experimental findings demonstrated the model's superior performance over commonly used CNN baselines across different OCT illness categories. Although the model is very accurate, its computational complexity and reliance on public datasets highlight the need for additional validation using different clinical data. Future work will focus on increasing model generalisability and lowering computing costs.

REFERENCES

[1] Z. Liu, C. Wang, X. Cai, H. Jiang, and J. Wang, "Discrimination of diabetic retinopathy from optical coherence tomography angiography images using machine learning methods," *IEEE Access*, vol. 9, pp. 25063–25073, 2021, doi: 10.1109/ACCESS.2021.3056430.

[2] Y. A. Bayhaqi, A. Hamidi, N. F. Canbaz, A. A. Navarini, P. C. Cattin, and A. Zam, "Deep-learning-based fast optical coherence tomography (OCT) image denoising for smart laser osteotomy," *IEEE Transactions on Medical Imaging*, vol. 41, no. 10, pp. 2615–2628, Oct. 2022, doi: 10.1109/TMI.2022.3168793.

[3] A. Parra-Mora, A. Cazanas-Gordon, R. Proença, and L. A. da Silva Cruz, "Epiretinal membrane detection in optical coherence tomography retinal images using deep learning," *IEEE Access*, vol. 9, pp. 94881–94891, 2021, doi: 10.1109/ACCESS.2021.3095655.

[4] R. Holland *et al.*, "Metadata-enhanced contrastive learning from retinal optical coherence tomography images," *Medical Image Analysis*, vol. 97, Art. no. 103296, 2024, doi: 10.1016/j.media.2024.103296.

[5] T. K. Yoo, J. Y. Choi, and H. K. Kim, "Feasibility study to improve deep learning in OCT diagnosis of rare retinal diseases with few-shot classification," *Medical and Biological Engineering and Computing*, vol. 59, no. 2, pp. 401–415, 2021, doi: 10.1007/s11517-021-02321-1.

[6] Y. Dong *et al.*, "A multi-branch convolutional neural network for screening and staging of diabetic retinopathy based on wide-field optical coherence tomography angiography," *IRBM*, vol. 43, no. 6, pp. 541–551, 2022, doi: 10.1016/j.irbm.2022.04.004.

[7] A. Altan, "DeepOCT: An explainable deep learning architecture to analyze macular edema on OCT images," *Engineering Science and Technology, an International Journal*, vol. 34, Art. no. 101091, 2022, doi: 10.1016/j.jestch.2021.101091.

[8] S. Baba, P. Kumari, and P. Saxena, "Retinal disease classification using custom CNN model from OCT images," *Procedia Computer Science*, vol. 235, pp. 3142–3152, 2024, doi: 10.1016/j.procs.2024.04.297.

[9] F. M. Talaat, A. A. A. Ali, R. ElGendy, and M. A. ElShafie, "Deep attention for enhanced OCT image analysis in clinical retinal diagnosis," *Neural Computing and Applications*, vol. 37, no. 2, pp. 1105–1125, 2025, doi: 10.1007/s00521-024-10450-5.

[10] A. Miladinović *et al.*, "Evaluating deep learning models for classifying OCT images with limited data and noisy labels," *Scientific Reports*, vol. 14, no. 1, Art. no. 30321, 2024, doi: 10.1038/s41598-024-81127-1.

[11] Y. Luo, Q. Xu, R. Jin, M. Wu, and L. Liu, "Automatic detection of retinopathy with optical coherence tomography images via a semi-supervised deep learning method," *Biomedical Optics Express*, vol. 12, no. 5, pp. 2684–2700, 2021, doi: 10.1364/BOE.418364.

[12] M. T. Noor, S. Abrar, J. A. Mahi, M. P. Mia, A. Hridoy, and S. Ghosh, "RetinaVision: XAI-driven augmented regulation for precise retinal disease classification using deep learning framework," *arXiv preprint arXiv:2602.19324*, Feb. 2026.

[13] O. S. Naren, "Retinal OCT Image Classification - C8," *Kaggle Dataset*, Oct. 2024, doi: 10.34740/KAGGLE/DSV/2736749.

[14] M. Kulyabin, "OCTDL: Optical coherence tomography dataset for image-based deep learning methods," *Mendeley Data*, vol. 4, 2024, doi: 10.17632/sncdhf53xc.4.

[15] D. S. Kermany *et al.*, "Identifying medical diagnoses and treatable diseases by image-based deep learning," *Cell*, vol. 172, no. 5, pp. 1122–1131.e9, Feb. 2018, doi: 10.1016/j.cell.2018.02.010.

[16] Z.-H. Zhou, Ensemble Methods: Foundations and Algorithms. Boca Raton, FL, USA: CRC Press, 2012, doi: 10.1201/b12207.

[17] M. Tan and Q. Le, "EfficientNetV2: Smaller models and faster training," *in Proceedings of the 38th International Conference on Machine Learning*, vol. 139, pp. 10096–10106, 2021.
[18] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Salt Lake City, UT, USA, 2018, pp. 7132–7141, doi: 10.1109/CVPR.2018.00745.
[19] E. Yusufoğlu *et al.*, "A comprehensive CNN model for age-related macular degeneration classification using OCT: Integrating inception modules, SE blocks, and ConvMixer," *Diagnostics*, vol. 14, no. 24, Art. no. 2836, 2024, doi: 10.3390/diagnostics14242836.
[20] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, Sep. 2014.
[21] T. Fawcett, "An introduction to ROC analysis," *Pattern Recognition Letters*, vol. 27, no. 8, pp. 861–874, 2006.
[22] C. M. Bishop, Pattern Recognition and Machine Learning. New York, NY, USA: Springer, 2006.
[23] M. Sokolova and G. Lapalme, "A systematic analysis of performance measures for classification tasks," *Information Processing and Management*, vol. 45, no. 4, pp. 427–437, 2009, doi: 10.1016/j.ipm.2009.03.002.
[24] J. Davis and M. Goadrich, "The relationship between precision-recall and ROC curves," in *Proceedings of the 23rd International Conference on Machine Learning*, Pittsburgh, PA, USA, 2006, pp. 233–240, doi: 10.1145/1143844.1143874.
[25] D. J. Hand and R. J. Till, "A simple generalisation of the area under the ROC curve for multiple class classification problems," *Machine Learning*, vol. 45, no. 2, pp. 171–186, 2001, doi: 10.1023/A:1010920819831.

BIOGRAPHIES OF AUTHORS

Saja Ataallah Muhammed is an assistant lecturer of computer science at Salahaddin University. Her research focuses on medical image processing using artificial intelligence. She studied for a master's degree in the same college. Now, she is a Ph.D. student and works as the head of the information technology department and an assistant lecturer in the College of Science. She has published several research papers in the field of Artificial Intelligence, medical image processing using (machine learning and deep learning) techniques. She can be contacted at email: swarajsharma6663@gmail.com.

Abbas M. Ali is an assistant professor. He holds a Ph.D. in computer vision from UKM, Malaysia. Currently, he works as a part-time lecturer at many universities in KRG. In addition, he has published various papers and has participated in reviewing conference papers and articles for international journals. He is also a member of many professional bodies in his specialization. He served as the head of the software and informatics engineering department from 2022 to 2024 in the college of engineering at Salahaddin University. Furthermore, he supervised many M.Sc. and Ph.D. students in his area of specialization. He is working as a staff member in software and informatics engineering at Salahaddin University, Erbil, Kurdistan, Iraq. He can be contacted at email: abbas.mohamad@su.edu.krd.

Dr. Dler Salih Hasan is an assistant professor at the Department of Computer Science and Information Technology in the College of Science at Salahaddin University-Erbil. He earned his Ph.D. in robotics engineering as a split-site student between the University of Salahaddin-Erbil and the University of Florida (USA). His research interests are in robotics and the internet of things (IoT), as evidenced by his published works. He currently holds the position of head of the computer science department. He has taught a range of subjects to BSc. and MSc. students, including OOP, IoT, embedded systems, robotics engineering, and fundamentals of information technology. He can be contacted at email: dler.hasan@su.edu.krd.