Original Article

From Image to Sequence: Exploring Vision Transformers for Optical Coherence Tomography Classification

Amirali Arbab[1], Aref Habibi[1], Hossein Rabbani[2], Mahnoosh Tajmirriahi[2]

[1]Department of Electrical and Computer Engineering, Isfahan University of Technology, Isfahan, Iran, [2]Medical Image and Signal Processing Research Center, School of Advanced Technologies in Medicine, Isfahan University of Medical Sciences, Isfahan, Iran

Amirali Arbab and Aref Habibi contributed equally to this work.

Abstract

Background: Optical coherence tomography (OCT) is a pivotal imaging technique for the early detection and management of critical retinal diseases, notably diabetic macular edema and age-related macular degeneration. These conditions are significant global health concerns, affecting millions and leading to vision loss if not diagnosed promptly. Current methods for OCT image classification encounter specific challenges, such as the inherent complexity of retinal structures and considerable variability across different OCT datasets. **Methods:** This paper introduces a novel hybrid model that integrates the strengths of convolutional neural networks (CNNs) and vision transformer (ViT) to overcome these obstacles. The synergy between CNNs, which excel at extracting detailed localized features, and ViT, adept at recognizing long-range patterns, enables a more effective and comprehensive analysis of OCT images. **Results:** While our model achieves an accuracy of 99.80% on the OCT2017 dataset, its standout feature is its parameter efficiency–requiring only 6.9 million parameters, significantly fewer than larger, more complex models such as Xception and OpticNet-71. **Conclusion:** This efficiency underscores the model's suitability for clinical settings, where computational resources may be limited but high accuracy and rapid diagnosis are imperative.

Code Availability: The code for this study is available at https://github.com/Amir1831/ViT4OCT

Keywords: *Computer vision, convolutional neural network, deep learning, multi-headed self-attention, optical coherence tomography, vision transformers*

Submitted: 21-Aug-2024 **Revised:** 03-Dec-2024 **Accepted:** 04-Dec-2024 **Published:** 09-Jun-2025

Introduction

Imaging plays a critical role in modern medicine, providing invaluable insights for diagnosis, treatment planning, and clinical trial design.[1] In the context of retinal disorders, advanced imaging techniques, such as optical coherence tomography (OCT), enable the detailed visualization of retinal layers, facilitating the early detection and management of conditions that can lead to vision loss.[2-4]

Diabetic macular edema (DME) is characterized by the accumulation of fluid in the retinal layers, leading to swelling and thickening of the retina. This condition occurs as a result of leaky blood vessels in the retina, a common complication of diabetes that can cause vision impairment if untreated.[5,6] Age-related macular degeneration (AMD), on the other hand, is a progressive condition that leads to the degeneration of the macula, often resulting in the thinning of retinal layers. The dry form of AMD is marked by the accumulation of drusen, yellow deposits beneath the retina, whereas the wet form involves abnormal blood vessel growth, which can lead to bleeding and scarring.[7,8] OCT scans enable the detailed visualization of these structural changes, making it a critical tool in diagnosing and managing these retinal disorders.

Convolutional neural networks (CNNs) have shown strong capabilities in analyzing OCT images, particularly in extracting localized features.[9,10] However, CNNs face limitations in capturing global relationships and long-range dependencies, which are crucial in complex medical images. Recently, transformers have emerged as powerful alternatives that excel at modeling these global relationships but at a high computational cost.[11,12] Hybrid models, which combine CNNs and

This is an open access journal, and articles are distributed under the terms of the Creative Commons Attribution-NonCommercial-ShareAlike 4.0 License, which allows others to remix, tweak, and build upon the work non-commercially, as long as appropriate credit is given and the new creations are licensed under the identical terms.

For reprints contact: WKHLRPMedknow_reprints@wolterskluwer.com

How to cite this article: Arbab A, Habibi A, Rabbani H, Tajmirriahi M. From image to sequence: Exploring vision transformers for optical coherence tomography classification. J Med Signals Sens 2025;15:18.

Address for correspondence:
Dr. Mahnoosh Tajmirriahi, Medical Image and Signal Processing Research Center, School of Advanced Technologies in Medicine, Isfahan University of Medical Sciences, Isfahan 81746734641, Iran.
E-mail: mata.riahi@yahoo.com

Access this article online

Website: www.jmssjournal.net

DOI: 10.4103/jmss.jmss_58_24

Quick Response Code:

transformers, have been proposed to leverage the strengths of both architectures. For instance, architectures such as BEFUnet[10] and TransUNet[13] have demonstrated improved performance by fusing CNN's local feature extraction with transformers' long-range modeling capabilities.[14]

This paper proposes a novel hybrid model that addresses the limitations of both individual architectures. We leverage the strengths of CNNs and transformers by combining them into a single framework for improved OCT performance. Our model employs a three-dimensional (3D) convolutional preprocessing step to efficiently extract localized features from medical image sequences. These features are fed into a transformer encoder, enabling the model to capture the crucial long-range dependencies and global context within the images.

We begin by reviewing existing methods and datasets in the field of OCT imaging in related work. Then, the method section details our proposed methodology, including the architecture and training process. In the experiment section, we present the dataset employed for evaluation, along with the metrics used to assess performance. To gain deeper insights into the model's behavior, we conduct ablation studies and interpret the results. Finally, the discussion and conclusion summarize our findings.

Related work

Traditional methods in retinal disease diagnosis

Retinal diseases pose a significant threat to vision and can lead to blindness globally. Analyzing and accurately classifying these diseases through retinal image analysis plays a critical role in early detection and timely intervention. Traditionally, this field relied on techniques centered on image processing and feature extraction.[15] The initial step involves image quality enhancement, followed by feature extraction. This process focuses on identifying and extracting crucial and informative features from the image, such as edges, textures, colors, and shapes. Principal component analysis,[16] Gabor filters, and wavelet transforms were commonly employed for feature extraction. In addition, feature descriptors such as histogram of oriented gradients (HOG),[17] local binary patterns (LBP),[18] and scale-invariant feature transform (SIFT)[19] were utilized. HOG[17] excelled at extracting edge and texture-related features, whereas LBP[18] was adept at texture analysis. SIFT,[19] on the other hand, is used to pinpoint and describe key and distinctive locations within the image. In Table 1,[20] we summarized some relevant traditional methods..

Deep learning revolution

In recent years, the landscape of retinal disease diagnosis has undergone a paradigm shift, witnessing a surge in the utilization of powerful deep learning (DL) architectures.[23,26,27] This trend is driven by the remarkable potential of these techniques to achieve exceptional accuracy in distinguishing various retinal pathologies. DL revolutionizes medical image analysis using neural networks to automatically extract essential features for classification.[28,29] This eliminates the previously prevalent need for manual feature engineering, a laborious and time-consuming process.

One such example is the pioneering work by Awais *et al*.,[30] who developed a novel deep classification system. Their approach leveraged a hybrid architecture, combining the feature extraction capabilities of VGG16 with the classification strengths of K-nearest neighbors and random forest algorithms. This resulted in a robust system capable of effectively differentiating between normal retinas and those afflicted with DME. Similarly, Lee *et al*.[26] adopted a standalone VGG16 architecture for binary classification purposes, achieving promising results in detecting age-related macular edema. CNNs have emerged as leaders in this field, demonstrating remarkable successes in classifying retinal pathologies using OCT images. Compared to traditional multi-block methods, CNNs offer a streamlined and efficient approach. In Table 2,[20] we summarized some DL methods.

The power of transfer learning

Recent studies have explored the potential of pretrained CNNs for OCT image classification, particularly in

Table 1: Some traditional machine learning-based methods in optical coherence tomography image classification

Authors, year	Model	Dataset	Performance measures
Albarrak *et al*., 2013[21]	Volume decomposition, LBP, Bayesian classifier	Private (140 3D OCT)	Accuracy: 91.4% Sensitivity: 92.4% Specificity: 90.5%
Srinivasan *et al*., 2014[22]	Multi-scale histogram, SVM	Duke[22]	Accuracy: 95.56%
Lemaître *et al*., 2016[23]	Five-step OCT classification: Preprocessing, LBP/LBP-Top feature extraction, classification	SERI private[23]	Sensitivity: 81.2% Specificity: 93.7%
Sun *et al*., 2017[24]	Sparse coding, dictionary learning, multiclass linear support vector machine classifier	Duke[22] + private	Accuracy: 97.78%
Venhuizen *et al*., 2017[25]	Bag of words algorithm with frequency vector classifier for automatic AMD severity grading	Eugenda	Sensitivity: 98.2% Specificity: 91.2%

OCT – Optical coherence tomography; LBP – Local binary patterns; SVM – Support vector machine; AMD – Age-related macular degeneration; 3D – Three dimensional

scenarios where access to large datasets might be limited. Karri *et al*.[44] successfully employed the GoogleLeNet model on a publicly available OCT dataset curated by Srinivasan *et al*.[22] This approach underscores the ability of transfer learning to mitigate the need for extensive training data. Similarly, Li *et al*.[45] adopted a transfer learning strategy using the VGG16 model, further demonstrating the versatility and adaptability of pretrained models. Beyond classification tasks, researchers have explored the application of these pretrained models in other areas of retinal disease diagnosis. Gómez-Valverde *et al*.[46] conducted a comparative analysis of various pretrained models including VGG19, GoogLeNet, ResNet50, and DeNet for glaucoma diagnosis using color fundus images. Their work highlights the adaptability and strong performance of these pretrained architectures. Based on this study, Cheng *et al*.[47] proposed a deep hashing algorithm based on ResNet50, demonstrating the potential of these models in image retrieval and classification tasks, further expanding the applications of transfer learning in retinal disease diagnosis. In Table 3, we summarized some transfer learning methods.

Methods

Our model builds upon the original architecture of vision transformer (ViT)[50] with a key modification. We propose a preprocessing step that leverages the strengths of CNNs for effective feature extraction, ultimately improving OCT image classification performance. The block diagram of the proposed method is depicted in Figure 1.

The proposed method consists of following steps

Input image

The input to our model consists of a batch of images denoted as $x \in R^{B\times C\times H\times W}$ where B represents the batch number, C is the number of channels, H is the image height, and W is the image width.

Break down into patches

To handle two-dimensional (2D) images with transformer, that typically operates on 1D embedding as input,[50] N nonoverlapping patches are extracted with size $P \times P$, ensuring they span the entire image $x \in R^{C\times H\times W}$. This converts a single image , to sequence of patches $x' \in R^{C\times N\times P\times P}$ with N number of frames in a sequence, where $N = \frac{HW}{P^2}$.

Table 2: Some deep learning methods in optical coherence tomography image classification

Authors, year	Model	Dataset	Performance measures
Lee *et al*., 2017[26]	Modified VGG16 CNN	Private dataset (48,312 normal and 52,690 AMD OCT scans)	Accuracy: 87.63% (OCT level), accuracy: 88.98% (volume level), accuracy: 93.45% (patient level)
Serener and Serte, 2019[31]	AlexNet and ResNet18 for dry and wet AMD	OCT2017[32]	ResNet18-Dry AMD Accuracy: 99.8%, sensitivity: 98.0%, specificity: 100.0% ResNet18-Wet AMD Accuracy: 98.8%, sensitivity: 95.6%, specificity: 99.9%
Fang *et al*., 2019[33]	IFCNN	2nd version of the OCT2017[32] + MURA[34]	Accuracy: 87.3%
Huang *et al*., 2019[34]	LGCNN	2nd version of the OCT2017[32] + HUCM[34]	Accuracy: 88.4%
Rasti *et al*., 2018[35]	MCME	NEH[35]	Precision: 99.36% Recall: 99.36% F1-score: 99.34%
Das *et al*., 2019[36]	MDFF	2nd version of the OCT2017[32]	Accuracy: 99.6% Sensitivity: 99.6% Specificity: 99.87%
Thomas *et al*., 2021[37]	Multi-scale CNN structure[37]	OCT2017[32]	Accuracy: 99.73%
Fang *et al*., 2019[38]	LACNN	OCT2017[32]	Accuracy: 90.1%
Das *et al*., 2020[39]	BACNN	DUIA[40] + NEH[35]	Accuracy: 90.1% (NEH) Accuracy: 97.1% (DUIA)
Hassan *et al*., 2021[41]	RAG-FW	Duke1[40], Duke2[42], Duke3[22], BIOMISA[43], OCT2017[32]	Accuracy: 98.6% Sensitivity: 98.27% Specificity: 99.6%

OCT – Optical coherence tomography; AMD – Age-related macular degeneration; CNN – Convolutional neural network; IFCNN – Iterative fusion CNN; BACNN – B-scan attentive CNN; LACNN – Lesion-aware CNN; MDFF – Multi-scale deep feature fusion; MCME – Multi-scale convolutional mixture of expert; LGCNN – Layer guided CNN; RAG-FW – Deep retinal analysis and grading framework; NEH – Noor Eye Hospital

Table 3: Some transfer learning methods in optical coherence tomography image classification[8]

Authors, year	Model	Dataset	Performance measures
Kermany *et al*., 2018[32]	Transfer learning with inceptionV3	OCT2017[32]	Accuracy: 96.6% Sensitivity: 97.8% Specificity: 97.4%
Li *et al*., 2019[45]	Transfer learning with VGG16	OCT2017[32]	Accuracy: 98.6% Sensitivity: 97.8% Specificity: 99.4%
Hwang *et al*., 2019[48]	Transfer learning (VGG16, inceptionV3, ResNet50)	Private dataset + OCT2017[32]	Accuracy: 91.20%–96.93% Sensitivity: 95.87%–97.65%
Kaymak and Serener, 2018[49]	AlexNet for retinal OCT pathologies	OCT2017[32]	Accuracy: 97.1% Sensitivity: 99.6% Specificity: 98.4%

OCT – Optical coherence tomography

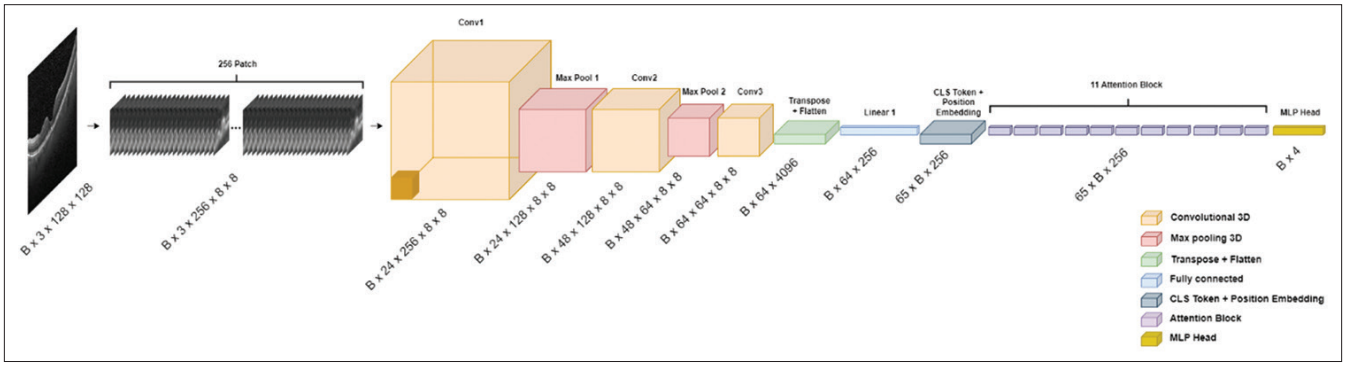

Figure 1: The block diagram of the proposed method. In the first step, our model converts a single image to a sequence of smaller images. Then, pass them through a series of Conv3D and Maxpool3D to extract features and then use the attention mechanism[51] to extract information across the entire image and finally use a linear classification to map our data to corresponding labels

In OCT imaging, each column of the image represents an A-scan which captures the signal intensity as a function of depth, displayed through pixel brightness. By arranging the patches in column-wise order (sequencing patches from top to bottom within each column), we preserve the inherent time sequence of the A-scans. This ordering maintains the continuity of the depth information and leverages the sequential nature of the A-scans, leading to improved performance in our hybrid model. Therefore, we modify the patch sequence so that patches within the same column are in sequential order, effectively capturing the temporal dynamics of the OCT data as shown in Figure 2.

Feature extraction

After converting the single OCT image into a sequence of patches, we employ three stacked Conv3D layers with ReLu activation[52] to extract spatio-temporal features of the sequence. While Conv3D layers are commonly used for processing 3D data, to the best of our knowledge, this is the first work to integrate them with a ViTs for 2D image classification tasks like OCT analysis. This approach allows us to not only capture spatial relationships within the OCT image but also extract temporal information across the patch sequence. An experiment was conducted where the patches were randomly shuffled, but this did not yield satisfactory accuracy compared to the fixed sequential arrangements. We also tested both horizontal and vertical sequential arrangements. Among these, the vertical arrangement achieved the highest accuracy, as it preserves the temporal and spatial coherence of related A-scans. This result highlights the importance of patch ordering in maintaining meaningful relationships between patches to improve model performance. While random ordering proved suboptimal, the sequential vertical arrangement demonstrated significant advantages, achieving unprecedented accuracy on this dataset. Further exploration of other ordering methods may still be an interesting direction for future studies.

To leverage spatial relationships in 2D OCT images, we divided each image into sequential patches, treating them as a pseudo-temporal input. This approach conceptually aligns with the application of 3D convolutions, allowing us to capture cross-patch dependencies more effectively. Our implementation utilized Conv3D layers to process these sequences, which offered a unique advantage over traditional Conv2D architectures in our experiments.

To further improve efficiency and reduce the computational burden, we incorporate MaxPool3D[53] between the Conv3D layers, with a kernel size of (2,1,1) and a stride of (2,1,1). This configuration is specifically chosen to reduce the sequence

length along the temporal dimension, without altering the spatial dimensions of the patches. By applying the pooling operation only in the temporal dimension, we ensure that the spatial resolution of each patch is preserved, allowing the model to retain critical retinal information while focusing on reducing the number of tokens passed to the transformer layers.

Hence, after applying these layers, we face a sequence $x' \in R^{C' \times N' \times P \times P}$ where C' represents the number of output feature channels extracted by the final convolutional layer and $N' = \frac{N}{4}$ reflects the reduced sequence length due to the MaxPool operation.

Flattening

Subsequently, each patch in the sequence is flattened into a 1D vector of size $C' \times P \times P$. This flattening operation transforms the sequence $\tilde{x}$ into a 2D tensor of size $N' \times (C'P^2)$. Therefore, for a single patch in the sequence, we have $x_p \in R^{C'P^2}$, where $p = 1,...,N'$ is patch index.

Linear embedding

In this stage, each patch is linearly mapped into an embedding vector:

$$z_p^{(0)} = Ex_p + e_p^{(pos)} \tag{1}$$

Where $z_p^{(0)} \in R^D$ is related to input vector x_p by adding $e_p^{(pos)} \in R^D$ to the transformation of original input x_p, with matrix $E \in R^{D \times C'P^2}$. Where $e_p^{(pos)}$ is a learnable position embedding to retain positional information.

To use transformers for classification problems, as the original BERT[54] do, a learnable classification token is added in the first position of the embedding sequence. BERT, which stands for Bidirectional Encoder Representations from Transformers, is an innovative model that employs a transformer architecture to process language. It captures complex semantic information through its multi-headed attention mechanism and bidirectional training approach. The final embedding vector can be considered the sequence of embedded text words that are processed by transformers in NLP.

Transformer layers

The core of the proposed model is a transformer encoder that processes the sequence of patches. Each transformer block follows this sequence:

$$y^l = MHSA\left(ln\left(z^l\right)\right) + z^l \tag{2}$$

$$z^{l+1} = MLP\left(ln\left(y^l\right)\right) + y^l \tag{3}$$

- Layer Normalization. The input patch sequence (x) is normalized using layer normalization[55] to reduce the training time by applying normalization on the inputs to the neurons in a layer
- Multi-headed Self-attention (MHSA). Is a mechanism that allows the model to focus on different parts of the input simultaneously, improving its ability to capture global relationships. This concept, first introduced in the transformer architecture,[51] is central to our model's ability to classify complex OCT images. Equation 4 represents the attention mechanism that is used in the MHSA blocks.

$$Attention\left(Q,K,V\right) = Softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{4}$$

In our model, the self-attention mechanism is employed to dynamically focus on different segments of the input OCT images. By generating queries Q, keys K, and values V through distinct linear projections, each attention head can capture unique information from the input embeddings.[51] This capability allows the model to adaptively enhance or suppress features depending on their relevance to the specific task of classifying OCT images.

The self-attention mechanism enhances the model's interpretability and efficiency in handling complex spatial relationships inherent in medical imaging data, as demonstrated in works such as reference,[50] where transformers were successfully applied to image data. By focusing on the most informative parts of an image, the model can efficiently process large inputs while maintaining high accuracy. This aspect is particularly crucial for medical diagnosis, where precision and detail are paramount.[13]

Moreover, the flexibility of this approach enables our model to better generalize across different OCT devices and conditions by learning to prioritize image features that are most diagnostic, regardless of variations in image quality or presentation. This results in robust performance, as evidenced by improved classification accuracy in our experimental results, making it highly effective for clinical applications where diverse and unpredictable image data are common.

- Multi-Layer Perceptron. Sequence is further processed by a two-layer MLP with GELU activation[56] and dropout[57] for regularization.

Linear classifier

Finally, a linear classifier is used to classify z_{cls}^L token, which is used to predict the final image class:

$$y = MLP\left(ln\left(z_{cls}^L\right)\right) \tag{5}$$

However, the quadratic complexity of the attention mechanism[58] makes processing all image tokens

Table 4: Our model hyperparameters overview

Parameter	Value
Image size	128×128
Patch size	8×8
Batch size	64
Number of patches	256
Number of heads	4
Number of encoder	11
Embed dim	256
Transformer feed-forward hidden layer size	512

computationally expensive. As we mentioned, the patch sequence is passed through a series of Conv3d[52] and Maxpool3d[53] operations to extract spatiotemporal features of the sequence and then pass to the transformers layers. This significantly reduces the number of tokens processed by the transformer layers, leading to faster training times and a more efficient model.[59]

Table 4 shows the hyperparameters used in our model, which were carefully chosen to balance accuracy and efficiency during training.

Experiments

Dataset and preliminary processing

We evaluated the proposed model on four publicly available datasets. We initially trained the model on the OCT2017, which was the main dataset, and then fine-tuned it on the remaining three datasets. Each dataset varied significantly in image size, dataset size, and OCT device types, providing a comprehensive testbed for assessing our model's adaptability and robustness across different clinical imaging conditions. These variations are critical for demonstrating the model's flexibility and generalization capabilities in real-world scenarios. The results, as detailed in our manuscript, confirm that the model consistently maintains high-performance levels despite the complexities introduced by different imaging technologies and dataset limitations. This adaptability is crucial for reliable clinical diagnostic tools, ensuring their efficacy across varied clinical settings.

OCT2017:[32] This widely used dataset contains 84,484OCT images, with 83,484 images allocated for training and 1000 for testing. The dataset includes four retinal states: normal, choroidal neovascularization (CNV), DME, and drusen. To ensure consistency across images, we resized them to a uniform size of 128 × 128 pixels. In addition, we preprocessed each image by extracting 256 smaller patches of size 8 × 8 pixels. The dataset is publicly available https://www.kaggle.com/datasets/paultimothymooney/kermany2018.

Noor Eye Hospital (NEH):[35] This dataset, collected from NEH in Tehran, consists of OCT scans categorized into three classes: normal, AMD, and DME. It includes 50 normal scans, 48 AMD scans, and 50 DME scans. While the axial resolution remains consistent at 3.5 μm across scans, both lateral and azimuthal resolutions vary between patients. This variation results in images with either 512 or 768 A-scans, and the number of B-scans per volume can range from 19 to 61 depending on the individual. The dataset is publicly available https://misp.mui.ac.ir/en/dataset-oct-classification-50-normal-48-amd-50-dme-0.

OCTID Dataset:[60] The OCTID dataset encompasses over 500 spectral-domain OCT volumetric scans, available in high-resolution jpeg format and grouped into several disease categories. For the purposes of our study, we extracted images classified under three categories: normal (NO), diabetic retinopathy, and AMD. Each scan from the dataset includes a fovea-centered image selected by a skilled clinical optometrist, with the images resized to 500 × 750 pixels for uniformity. The scans were captured using a Cirrus HD-OCT machine at Sankara Nethralaya Eye Hospital, Chennai, India, providing a detailed basis for diagnosing and categorizing various stages of retinal pathologies. The dataset's diverse range of disease stages offers a robust framework for evaluating the sensitivity and repeatability of diagnostic techniques. The dataset is publicly available https://borealisdata.ca/dataverse/OCTID.

OCTDL Dataset:[61] The OCTDL dataset consists of 2064 high-resolution OCT B-scans, meticulously categorized to represent various retinal diseases and conditions. Each image, centered on the fovea, provides a detailed visualization of the retinal layers, posterior vitreous body, and choroidal vessels. Originally including multiple disease categories, our study specifically utilized images labeled as AMD and DME, focusing on these two classes due to their significant impact on vision quality and prevalence in clinical research. The open-access nature of this dataset makes it a valuable resource for the development of diagnostic algorithms and the advancement of automatic processing techniques aimed at early disease detection. The dataset is publicly available https://data.mendeley.com/datasets/sncdhf53xc/4.

Performance metrics

To evaluate the proposed model, we use three standard metrics: accuracy, sensitivity, and specificity, terms TP, FP, FN, and TN denote true positive, false positive, false negative, and true negative and is the number of samples in test set and is the number of classes. Results of these

Table 5: Test results on optical coherence tomography 2017 dataset

Architectures	Accuracy	Sensitivity	Specificity	Parameters (m)
InceptionV3[32]	96.60	97.80	97.40	23.6
ResNet50[63]	99.30	99.30	99.76	25.64
MobileNet-v2[64,65]	99.40	99.40	99.80	3.4
Xception[65]	99.70	99.70	99.90	22.8
OpticNet-71[62]	99.80	99.80	99.93	12.5
Our model (horizontal)	99.50	99.50	99.83	6.9
Our model (vertical)	99.80	99.80	99.93	6.9

metrics are reported in Tables 5 and 6 and compared to the recent studies.

$$Accuracy = \frac{1}{N}\sum TP \tag{6}$$

$$Sensitivity = \frac{1}{K}\sum \frac{TP}{TP + FN} \tag{7}$$

$$Specificity = \frac{1}{K}\sum \frac{TN}{TN + FP} \tag{8}$$

We also include the confusion matrix Figure 3 in addition to other performance metrics to provide a class-wise breakdown of predictions. According to Table 5[62] and Figure 3, the proposed method has outstanding performance in the classification of OCT images. In addition, it can be seen from Table 5 that the proposed method has fewer parameters than the most comparing methods. The results in Table 5 are taken directly from the cited papers to provide a fair comparison with our model's performance on the same datasets.

We arranged the patches in two different ways. In the first method, horizontal patches were arranged sequentially from left to right and then top to bottom. In the second method, vertical patches were arranged sequentially from top to bottom and then left to right. The second approach achieved the highest accuracy on this dataset. This result can be explained by the fact that each column of OCT data represents an A-scan, which is a temporal signal with its amplitude encoded as pixel brightness. By arranging the patches column-wise, the temporal sequence of the A-scans is preserved, leading to better performance. Our results for both methods can be seen in Tables 5 and 6.

Training process

We trained our model using the stochastic gradient descent optimizer within the PyTorch framework, leveraging an Nvidia RTX 3050 Ti GPU. The initial learning rate was set at 0.01 and adjusted according to a decay schedule over the course of the training to facilitate optimal convergence. The training spanned 150 epochs, with each epoch averaging 244 s on the OCT 2017[32] dataset. Our model demonstrated exceptional performance on this dataset, achieving a competitive accuracy of 99.80% and maintaining remarkable parameter efficiency with only 6.9 million parameters. This efficiency is significant when compared to other state-of-the-art architectures, as shown in Table 5.

Fine-tuning process

To assess the model's adaptability and performance across various datasets, we fine-tuned the pretrained model on additional datasets, including NEH,[35] OCTID,[60] and OCTDL.[61] Pretrained weights from the OCT2017 training phase were utilized, and an additional layer was incorporated at the end of our model. During fine-tuning, we selectively adjusted only the last five layers of the encoder layers, whereas the earlier layers were frozen to retain the generalized features previously learned.

During the fine-tuning phase, we employed a 5-fold cross-validation method, limiting the training to a

Table 6: Additional experiment results

Performance metrics are based on five-fold cross-validation, with each fold trained for 15 epochs

Datasets	Classes	Architectures	Accuracy	Sensitivity	Specificity
NEH[35]	AMD, DME, normal	ResNet50[63]	82±2.1	82±2.4	89.8±1.6
		InceptionV3[32]	81.7±1	82.2±0.7	90.8±0.7
		Our model (horizontal)	81.12±2.1	81.17±1.5	90.21±2.1
		Our model (vertical)	82.08±2.1	81.34±2.44	90±0.87
OCTID[60]	AMD, DR, normal	ResNet50[63]	92.4±2.9	86.22±5.9	95.87±1.54
		InceptionV3[32]	87.8±2.9	77.80±3.0	93.1±1.9
		Our model (horizontal)	91.8±2.5	87.1±5.8	95.5±1
		Our model (vertical)	92.19±3.1	87.27±2.8	95±3.2
OCTDL[61]	AMD, DME	ResNet50[63]	95.06±0.6	63.85±12.4	98.78±1
		InceptionV3[32]	93.38±2.3	44.31±26.6	99.12±1.3
		Our model (horizontal)	93.6±1.2	76.32±5.9	96.38±2.2
		Our model (vertical)	95.33±2.2	87.54±8.6	96.61±2.5

AMD – Age-related macular degeneration; NEH – Noor eye hospital; DME – Diabetic macular edema; DR – Diabetic retinopathy; OCTDL – Optical coherence tomography dataset for image-based deep learning; OCTID – Optical coherence tomography image database

Table 7 : Ablation study on optical coherence tomography 2017 dataset

Architectures	Pretrained	3D feature extractor	Accuracy	Time (s)	Parameters (m)
Proposed method	✘	✔	99.8	244	6.9
Pure ViT	✘	✘	48	600	5.9
ViT-b-16	✔	✘	96.9	1026	85.8

3D – Three dimensional; ✔ – Included/Applied; ✘ – Excluded/Not applied

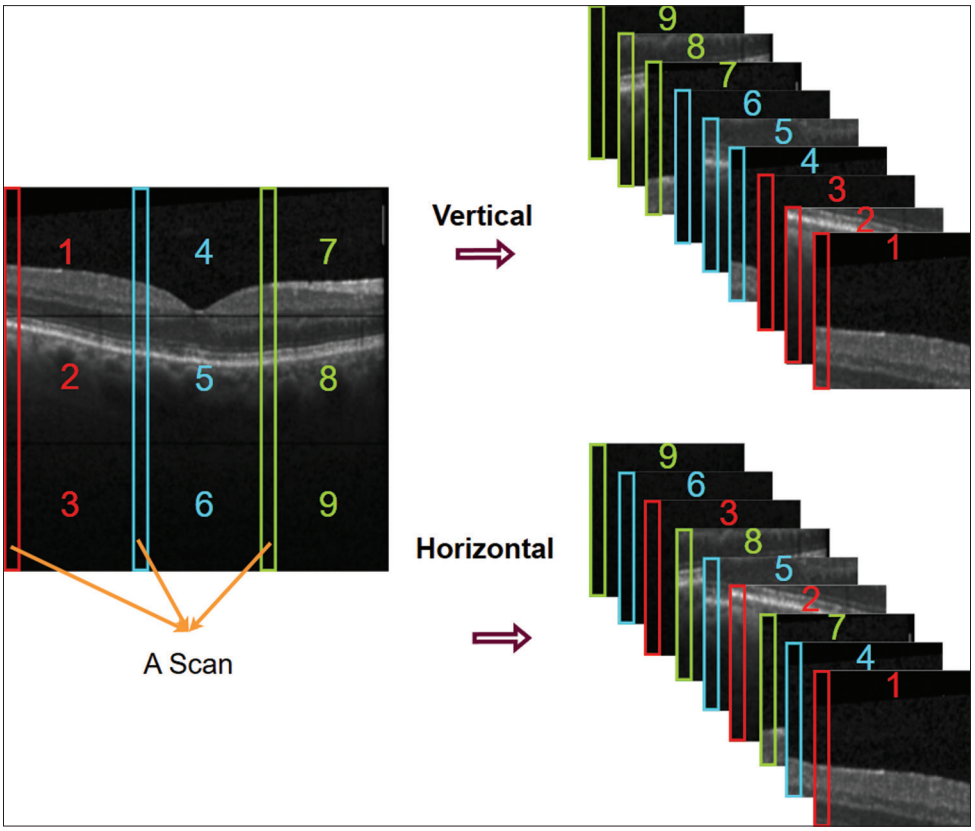

Figure 2: Illustration of patch arrangement methods: Vertical arrangement preserves the temporal sequence of A-scans (left to right and top to bottom), whereas horizontal arrangement processes patches sequentially (top to bottom and left to right)

maximum of 15 epochs per fold to prevent overfitting and ensure robustness across varied datasets.[66]

Results

The model's robustness was further validated across multiple datasets, demonstrating its versatility and effectiveness in different clinical settings. Specifically, the model achieved accuracies of 82.08% on the NEH[35] dataset, 92.19% on the OCTID[60] dataset, and 95.33% on the OCTDL[61] dataset. Notably, it excelled in the detection of Diabetic Macular Edema (DME) B-scans, where it showed a high accuracy of 91.89% on the NEH[35] dataset. This capability is crucial for early diagnosis and intervention, which can prevent potential vision loss, underlining the model's practical significance in real-world medical applications. Table 6 expands on more results, providing a detailed comparison and further insights into the model's performance across these diverse datasets. These results affirm the model's utility in clinical environments, providing a reliable tool for the early detection and treatment of significant retinal conditions.

Ablation experiments

Our proposed model for OCT image classification, a hybrid ViT architecture, achieved a remarkable maximum accuracy of 99.80 on the OCT 2017 dataset. This model also boasts impressive efficiency, requiring only 244 s per training epoch and maintaining a relatively low parameter count of 6.9 million. To understand the contribution of each component to this performance, we conducted a comprehensive ablation study, investigating the impact of various architectural elements.

The core innovation of our model lies in the preprocessing step that utilizes stacked Conv3D and MaxPool3D layers. In our study, we employ Conv3D and MaxPool3D layers for processing OCT images, tailored to the architecture of ViTs by converting images into patch sequences. This transformation leverages spatial-temporal features effectively and ensures the model focuses on significant retinal data while minimizing background noise. Our methodology significantly enhances model performance, making it particularly effective for the unique properties of OCT imaging. This approach ensures a precise alignment with the needs of OCT image analysis.

To assess the effectiveness of this feature extraction approach, we removed these layers from the model. This resulted in a pure ViT architecture without any 3D feature extraction. In a pure ViT architecture, images are split into patches that are flattened into tokens, which are then processed by transformer layers using self-attention to capture global relationships across the image. Unlike CNNs, which use convolutional layers to extract local

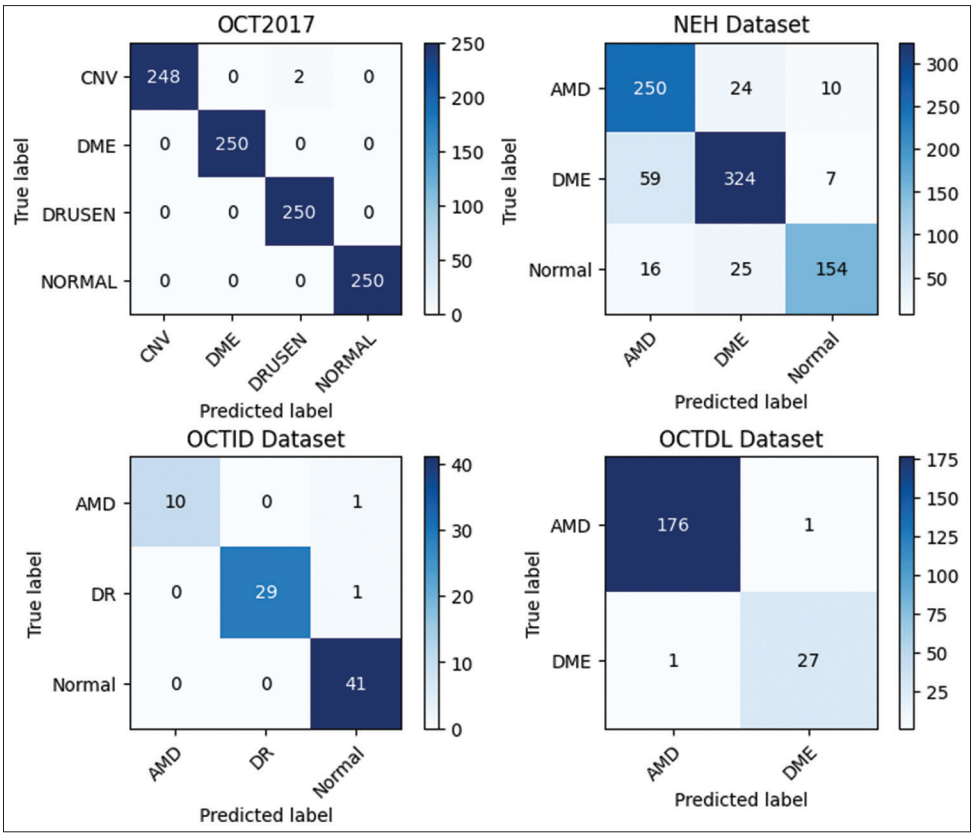

Figure 3: Confusion matrices illustrating the performance of our model in classifying retinal diseases from four datasets: OCT2017, NEH, OCTID, and OCTDL. Each matrix corresponds to one dataset, with rows indicating the actual disease classes and columns showing the predicted classes. These results are derived from the last fold of a five-fold cross-validation process, highlighting the model's diagnostic accuracy in various testing scenarios

features, pure ViT relies solely on the attention mechanism to learn both local and global patterns.

While the model remained functional, its performance significantly dropped compared to the hybrid model. As shown in Table 7, the maximum accuracy ViT model was only 48, highlighting the crucial role of the Conv3D and MaxPool3D layers in capturing relevant spatial and potentially temporal features from the OCT image sequences.

To delve deeper into the influence of pre-training, we compared the proposed method with a pre-trained ViT (vit-b-16), this model leveraged pretrained weights obtained from a large image dataset, promoting transfer learning to the medical image classification task. The pretrained model achieved a considerably higher maximum accuracy of 96.9. This substantial improvement highlights the effectiveness of pretraining in transferring learned features from a large image dataset to the OCT image classification task. However, training the pretrained ViT took 1026 s per epoch, significantly longer than the proposed method (244s) and has a parameter count of 85 million, considerably larger compared to the previous method (6.9 million).

Table 8 expands on this by showing the impact of the preprocessing steps on computational complexity under different configurations. The column "Preprocess Step" indicates whether Conv3D and MaxPool3D layers were used. When these layers are applied, our model achieves the lowest floating-point operations (FLOPs) at 137.07 GFLOPs with 6.98 million parameters, whereas without the preprocessing step (using only Conv2D for patch

Table 8: Impact of the preprocessing steps

Hyper-parameters	Value	Preprocess step	Parameters (m)	FLOPs (GFLOPs)
Batch size	64	✘	6.06	984.5
Image size	256×256			
Patch size	16×16			
Number encoder	11			
Batch size	64	✔	10.13	403.8
Image size	256×256			
Patch size	16×6			
Number encoder	11			
Batch size	64	✘	5.92	960.4
Image size	128×128			
Patch size	8×8			
Number encoder	11			
batch size	64	✔	6.98	137.07
Image size	128×128			
Patch size	8×8			
Number encoder	11			

FLOP – Floating-point operation; GFLOP – Giga floating-point operations; ✔ – Included/Applied; ✘ – Excluded/Not applied

extraction), the FLOPs are substantially higher, reaching up to 984.5 GFLOPs with a similar parameter count. This highlights the effectiveness of our novel preprocessing approach in reducing the computational load while maintaining model accuracy.

Interpretation

DL models have revolutionized various fields, from image recognition to natural language processing. However, their intricate nature often resembles a black box, making it challenging to comprehend their decision-making processes. This has led to the development of interpretation techniques, aimed at shedding light on the inner workings of these powerful models. One such technique utilizes heatmaps to visualize the image regions that significantly influence the model's predictions. These heatmaps highlight the areas that the model deems crucial for classification, providing valuable insights into its decision-making process. For supervised learning tasks, various interpretation methods have emerged, including rule-based explanations and decision trees. These techniques extract interpretable rules or decision paths from the model, allowing for a more transparent understanding of its reasoning.

We employed occlusion sensitivity,[67] a technique to understand which regions of an input image contribute most to the model's prediction for a specific class. It works by systematically occluding (masking) different parts of the image and observing the change in the model's output probability for the target class. Regions that cause a significant decrease in the target class probability when occluded are considered to be important for the model's prediction. We used a sliding window approach with a window size of (3, 16, 16) and strides of (3, 2, 2) to occlude image regions and analyze their impact on the prediction for the target class. The visualizations in Figure 4 present the original input image, along with the heat maps representing the attribution scores for the target class. Positive attribution scores (green) indicate regions contributing to the prediction, while negative scores (red) indicate regions potentially confusing the model. The masked image shows the input with occluded regions based on the occlusion sensitivity analysis. Our analysis revealed that the model heavily relies on the presence of the retinal layers within the OCT image for accurate classification. This is evident in the positive attribution heat map, where these regions exhibit high scores. Conversely, background noise and artifacts seem to have a negative impact, as indicated by the negative scores in certain areas of the heatmap. These findings align with our understanding of the task, where identifying and focusing on the distinct characteristics of the retinal layers (e.g., thickness, reflectivity) is crucial for accurate classification of retinal diseases such as normal, CNV, DME, and drusen.

Table 9: Comparison of computational complexity

Architectures	FLOPs (GFLOPs)	Parameters (m)
ResNet50	525.5	25.557
MobileNet	39.3	3.505
VGG-16	1982	138.357
ViT-b-16	2158	85.8
Our model	137	6.982

FLOP – Floating-point operation; GFLOP – Giga floating-point operations

Computational complexity and efficiency comparison

In this section, we present a detailed computational complexity analysis comparing our proposed model, which utilizes Conv3D and MaxPool3D layers, with other state-of-the-art models, including ViT-B/16, ResNet50, MobileNet, and VGG-16. This analysis includes the number of FLOPs and the number of parameters. The results are summarized in Table 9. All FLOPs, training time, and memory usage measurements were performed using an NVIDIA Tesla T4 GPU.

Our model achieves 137.071 GFLOPs, which is significantly lower than ViT-b-16, which requires 2.158 TFLOPs–approximately 15 times more computational power. This reduction in computational complexity makes our model more efficient and suitable for environments where computational resources are limited, such as real-time clinical settings.

With 6.982 million parameters, our model is relatively compact compared to ViT-b-16, which has 85.802 million parameters–more than 12 times the size. This reduction in parameters leads to faster training times, lower memory consumption, and easier deployment in resource-constrained environments.

While ViT-b-16 is designed to process large and complex datasets, its computational complexity and number of parameters make it resource-intensive. In contrast, our model balances the strengths of ViTs with the computational efficiency of Conv3D and MaxPool3D layers, making it more practical for OCT image classification.

Our model reduces the input sequence size by applying Conv3D and MaxPool3D layers, which compress redundant background information in OCT images. This allows the transformer blocks to process a more compact and relevant set of tokens, improving both computational efficiency and performance.

Discussion and Conclusion

In this study, a hybrid ViT architecture was proposed which can achieve a remarkable maximum accuracy of 99.80 on the OCT2017 dataset. This model also boasts impressive efficiency, requiring only 244 s per training epoch and maintaining a relatively low parameter count of 6.9 million. To understand the contribution of each component to this performance, we

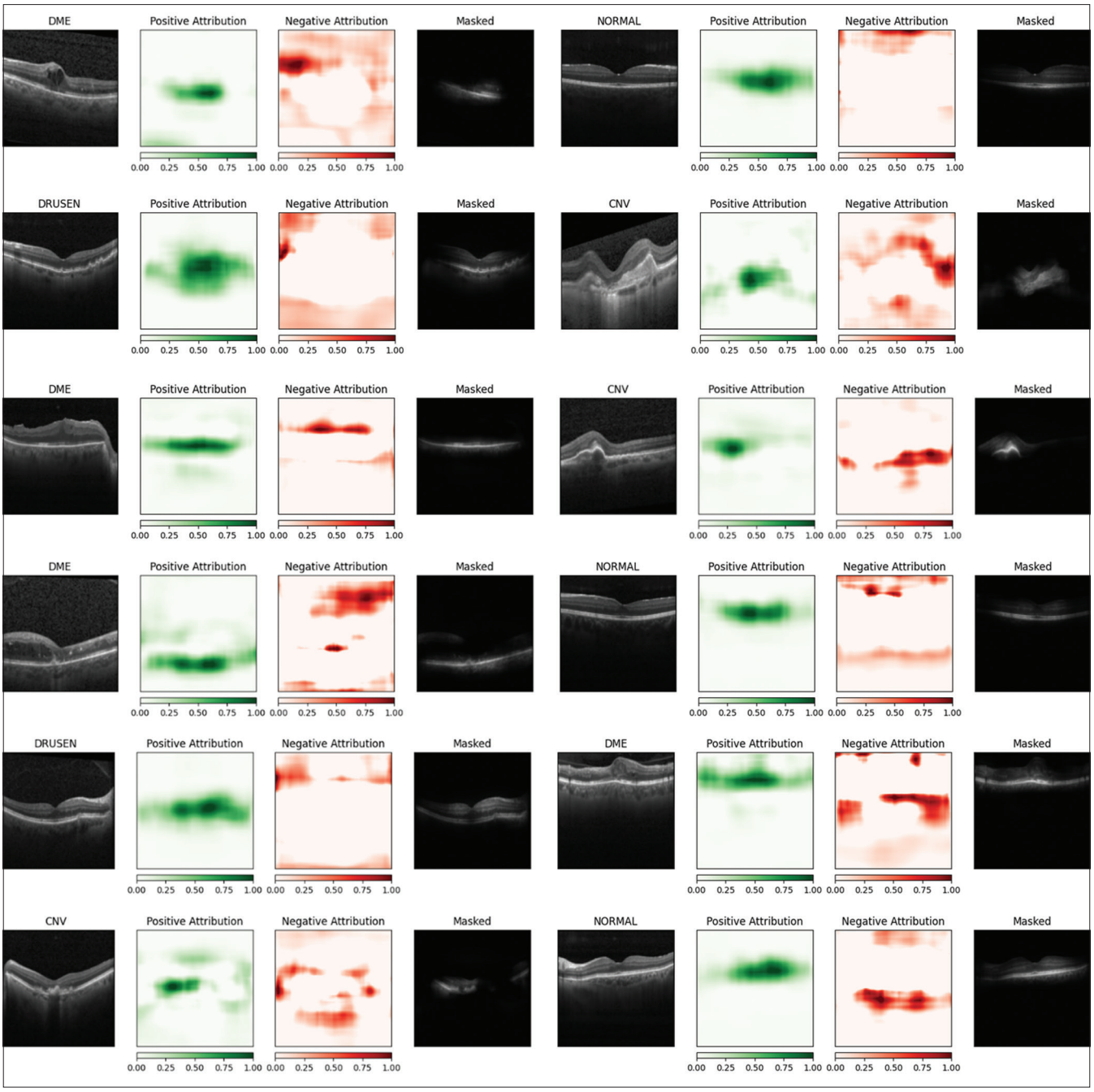

Figure 4: This figure visualizes how the model classifies retinal diseases in OCT images using occlusion sensitivity.[67] The original image (left) shows an OCT scan. The heatmap (center) highlights regions (green) the model relies on for classification (e.g., retinal layers). Areas with negative scores (red) might be confusing the model. The masked image (right) demonstrates how occluding specific regions affects the model's prediction

conducted a comprehensive ablation study. The ablation study revealed several key insights. First, the proposed incorporation of the 3D feature extractor with Conv3D and MaxPool3D layers significantly enhanced performance by capturing relevant spatial and potentially temporal features from the OCT image sequences. Second, pretraining with a large image dataset further improved accuracy but with a trade-off in training time and parameter count.

Importantly, the hybrid ViT architecture is designed to be generalizable across various OCT imaging datasets. Its device-independent property ensures consistent performance regardless of the hardware used, enabling widespread adoption in clinical settings. The insights gained from our ablation study not only validate the effectiveness of our approach but also provide a foundation for future research aimed at optimizing hybrid architectures for enhanced diagnostic accuracy and efficiency in diverse OCT imaging tasks. This positions our method as an advancement in the field, with the potential to improve patient outcomes through timely and accurate detection of conditions such as DME.

Further, we also trained the model on a laptop equipped with a Nvidia RTX3050 Ti, achieving approximately 5 min per epoch. This demonstrates that our model, despite utilizing Conv3D, remains computationally manageable. Comparative tests with Conv2D confirmed that our model does not exhibit a significant increase in parameter count, underlining its efficiency.

Financial support and sponsorship

This work was supported in part by the Vice Chancellery for Research and Technology of Isfahan University of Medical Sciences under Grant number 2402318.

Conflicts of interest

There are no conflicts of interest.

References

1. Leitgeb R, Placzek F, Rank E, Krainz L, Haindl R, Li Q, *et al.* Enhanced medical diagnosis for doctors: A perspective of optical coherence tomography. J Biomed Opt 2021;26:100601.
2. Miri M, Amini Z, Rabbani H, Kafieh R. A comprehensive study of retinal vessel classification methods in fundus images. J Med Signals Sens 2017;7:59-70.
3. Huang D, Swanson EA, Lin CP, Schuman JS, Stinson WG, Chang W, *et al.* Optical coherence tomography. Science 1991;254:1178-81.
4. Keane PA, Sadda SR. Retinal imaging in the twenty-first century: State of the art and future directions. Ophthalmology 2014;121:2489-500.
5. Jamshidi M, Rabbani H, Amini Z, Kafieh R, Ommani A, Lakshminarayanan V. Automatic detection of the optic disc of the retina: A fast method. J Med Signals Sens 2016;6:57-63.
6. Klein R, Klein BE, Moss SE. The Wisconsin epidemiological study of diabetic retinopathy: A review. Diabetes Metab Rev 1989;5:559-70.
7. Ferris FL 3rd, Wilkinson CP, Bird A, Chakravarthy U, Chew E, Csaky K, *et al.* Clinical classification of age-related macular degeneration. Ophthalmology 2013;120:844-51.
8. Mitchell P, Liew G, Gopinath B, Wong TY. Age-related macular degeneration. Lancet 2018;392:1147-59.
9. Rasti R, Mehridehnavi A, Rabbani H, Hajizadeh F. Convolutional mixture of experts model: A comparative study on automatic macular diagnosis in retinal optical coherence tomography imaging. J Med Signals Sens 2019;9:1-14.
10. Manzari ON, Kaleybar JM, Saadat H, Maleki S. BEFUnet: A hybrid CNN-transformer architecture for precise medical image segmentation. arXiv preprint 2024. [doi: 10.48550/arXiv:2402.08793].
11. Li J, Chen J, Tang Y, Wang C, Landman BA, Zhou SK. Transforming medical imaging with transformers? A comparative review of key properties, current progresses, and future perspectives. Med Image Anal 2023;85:102762.
12. Kim JW, Khan AU, Banerjee I. Systematic review of hybrid vision transformer architectures for radiological image analysis. medRxiv 2024. p. 2024-6.
13. Chen J, Mei J, Li X, Lu Y, Yu Q, Wei Q, *et al*. TransUNet: Rethinking the U-Net architecture design for medical image segmentation through the lens of transformers. Med Image Anal 2024;97:103280.
14. Fan Y, Song J, Yuan L, Jia Y. HCT-Unet: Multi-target medical image segmentation via a hybrid CNN-transformer unet incorporating multi-axis gated multi-layer perceptron. Vis Comput 2024. p. 1-16.
15. Nguyen H, Roychoudhry A, Shannon A. Classification of Diabetic Retinopathy Lesions from Stereoscopic Fundus Images. In Proceedings of the 19th Annual International Conference of the IEEE Engineering in Medicine and Biology Society. Magnificent Milestones and Emerging Opportunities in Medical Engineering (Cat. No. 97CH36136), IEEE; 1997. p. 426-8.
16. Maćkiewicz A, Ratajczak W. Principal components analysis (PCA). Comput Geosci 1993;19:303-42.
17. Dalal N, Triggs B. Histograms of Oriented Gradients for Human Detection. In: 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05). IEEE; 2005. p. 886-93.
18. Ojala T, Pietikäinen M, Harwood D. A comparative study of texture measures with classification based on featured distributions. Pattern Recognit 1996;29:51-9.
19. Lowe DG. Object Recognition from Local Scale-Invariant Features. In: Proceedings of the Seventh IEEE International Conference on Computer Vision. IEEE; 1999. p. 1150-7.
20. Sotoudeh-Paima S, Jodeiri A, Hajizadeh F, Soltanian-Zadeh H. Multi-scale convolutional neural network for automated AMD classification using retinal OCT images. Comput Biol Med 2022;144:105368.
21. Albarrak A, Coenen F, Zheng Y. Age-Related Macular Degeneration Identification in Volumetric Optical Coherence Tomography Using Decomposition and Local Feature Extraction. In: Proceedings of 2013 International Conference on Medical Image, Understanding and Analysis; 2013. p. 59-64.
22. Srinivasan PP, Kim LA, Mettu PS, Cousins SW, Comer GM, Izatt JA, *et al.* Fully automated detection of diabetic macular edema and dry age-related macular degeneration from optical coherence tomography images. Biomed Opt Express 2014;5:3568-77.
23. Lemaître G, Rastgoo M, Massich J, Cheung CY, Wong TY, Lamoureux E, *et al.* Classification of SD-OCT volumes using local binary patterns: Experimental validation for DME detection. J Ophthalmol 2016;2016:3298606.
24. Sun Y, Li S, Sun Z. Fully automated macular pathology detection in retina optical coherence tomography images using sparse coding and dictionary learning. J Biomed Opt 2017;22:16012.
25. Venhuizen FG, van Ginneken B, van Asten F, van Grinsven MJ, Fauser S, Hoyng CB, *et al.* Automated staging of age-related macular degeneration using optical coherence tomography. Invest Ophthalmol Vis Sci 2017;58:2318-28.
26. Lee CS, Baughman DM, Lee AY. Deep learning is effective for the classification of OCT images of normal versus age-related macular degeneration. Ophthalmol Retina 2017;1:322-7.
27. Treder M, Lauermann JL, Eter N. Automated detection of exudative age-related macular degeneration in spectral domain optical coherence tomography using deep learning. Graefes Arch Clin Exp Ophthalmol 2018;256:259-65.
28. LeCun Y, Bengio Y, Hinton G. Deep learning. Nature 2015;521:436-44.
29. Ting DS, Pasquale LR, Peng L, Campbell JP, Lee AY, Raman R, *et al.* Artificial intelligence and deep learning in ophthalmology. Br J Ophthalmol 2019;103:167-75.
30. Awais M, Müller H, Tang TB, Meriaudeau F. Classification of Sd-oct Images Using a Deep Learning Approach. In: 2017 IEEE International Conference on Signal and Image Processing Applications (ICSIPA), IEEE; 2017. p. 489-92.
31. Serener A, Serte S. Dry and wet age-related macular degeneration classification using oct images and deep learning. In: 2019 Scientific Meeting on Electrical-Electronics and Biomedical Engineering and Computer Science (EBBT). IEEE; 2019. p. 1-4.

32. Kermany DS, Goldbaum M, Cai W, Valentim CC, Liang H, Baxter SL, *et al.* Identifying medical diagnoses and treatable diseases by image-based deep learning. Cell 2018;172:1122-31.e9.
33. Fang L, Jin Y, Huang L, Guo S, Zhao G, Chen X. Iterative fusion convolutional neural networks for classification of optical coherence tomography images. J Vis Commun Image Represent 2019;59:327-33.
34. Huang L, He X, Fang L, Rabbani H, Chen X. Automatic classification of retinal optical coherence tomography images with layer guided convolutional neural network. IEEE Signal Process Lett 2019;26:1026-30.
35. Rasti R, Rabbani H, Mehridehnavi A, Hajizadeh F. Macular OCT classification using a multi-scale convolutional neural network ensemble. IEEE Trans Med Imaging 2018;37:1024-34.
36. Das V, Dandapat S, Bora PK. Multi-scale deep feature fusion for automated classification of macular pathologies from OCT images. Biomed Signal Process Control 2019;54:101605.
37. Thomas A, Harikrishnan PM, Krishna AK, Palanisamy P, Gopi VP. A novel multiscale convolutional neural network based age-related macular degeneration detection using OCT images. Biomed Signal Process Control 2021;67:102538.
38. Fang L, Wang C, Li S, Rabbani H, Chen X, Liu Z. Attention to lesion: Lesion-aware convolutional neural network for retinal optical coherence tomography image classification. IEEE Trans Med Imaging 2019;38:1959-70.
39. Das V, Prabhakararao E, Dandapat S, Bora PK. B-scan attentive CNN for the classification of retinal optical coherence tomography volumes. IEEE Signal Process Lett 2020;27:1025-9.
40. Farsiu S, Chiu SJ, O'Connell RV, Folgar FA, Yuan E, Izatt JA, *et al.* Quantitative classification of eyes with and without intermediate age-related macular degeneration using optical coherence tomography. Ophthalmology 2014;121:162-72.
41. Hassan T, Akram MU, Werghi N, Nazir MN. RAG-FW: A hybrid convolutional framework for the automated extraction of retinal lesions and lesion-influenced grading of human retinal pathology. IEEE J Biomed Health Inform 2021;25:108-20.
42. Chiu SJ, Allingham MJ, Mettu PS, Cousins SW, Izatt JA, Farsiu S. Kernel regression based segmentation of optical coherence tomography images with diabetic macular edema. Biomed Opt Express 2015;6:1172-94.
43. Hassan T, Akram MU, Masood MF, Yasin U. BIOMISA Retinal Image Database for Macular and Ocular Syndromes. In: Image Analysis and Recognition: 15th International Conference, ICIAR 2018, Póvoa de Varzim, Portugal: Proceedings 15, Springer; 2018. p. 695-705.
44. Karri SP, Chakraborty D, Chatterjee J. Transfer learning based classification of optical coherence tomography images with diabetic macular edema and dry age-related macular degeneration. Biomed Opt Express 2017;8:579-92.
45. Li F, Chen H, Liu Z, Zhang X, Wu Z. Fully automated detection of retinal disorders by image-based deep learning. Graefes Arch Clin Exp Ophthalmol 2019;257:495-505.
46. Gómez-Valverde JJ, Antón A, Fatti G, Liefers B, Herranz A, Santos A, *et al.* Automatic glaucoma classification using color fundus images based on convolutional neural networks and transfer learning. Biomed Opt Express 2019;10:892-913.
47. Cheng S, Wang L, Du A. Histopathological image retrieval based on asymmetric residual hash and DNA coding. IEEE Access 2019;7:101388-400.
48. Hwang DK, Hsu CC, Chang KJ, Chao D, Sun CH, Jheng YC, *et al.* Artificial intelligence-based decision-making for age-related macular degeneration. Theranostics 2019;9:232-45.
49. Kaymak S, Serener A. Automated Age-Related Macular Degeneration and Diabetic Macular Edema Detection on Oct Images Using Deep Learning. In: 2018 IEEE 14th International Conference on Intelligent Computer Communication and Processing (ICCP). IEEE; 2018. p. 265-9.
50. Dosovitskiy A, *et al.* An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint 2020. [doi: 10.48550.arXiv:2010.11929].
51. Vaswani A, *et al.* Attention is all you need. Adv Neural Inf Process Syst 2017;30.
52. Tran D, Bourdev L, Fergus R, Torresani L, Paluri M. Learning Spatiotemporal Features with 3d Convolutional Networks. In: Proceedings of the IEEE International Conference on Computer Vision; 2015. p. 4489-97.
53. Maturana D, Scherer S. Voxnet: A 3d Convolutional Neural Network for Real-Time Object Recognition. In: 2015 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), IEEE; 2015. p. 922-8.
54. Kenton JD, Toutanova LK. Bert: Pre-training of deep bidirectional transformers for language understanding. InProceedings of naacL-HLT 2019;1:2.
55. Ba JL, Kiros JR, Hinton GE. Layer normalization. arXiv preprint 2016. [doi: 10.48550/arXiv:1810.04805].
56. Hendrycks D, Gimpel K. Gaussian error linear units (gelus). arXiv preprint 2016. [doi: 10.48550/arXiv:1606.08415].
57. Srivastava N, Hinton G, Krizhevsky A, Sutskever I, Salakhutdinov R. Dropout: A simple way to prevent neural networks from overfitting. J Mach Learn Res 2014;15:1929-58.
58. Arnab A, Dehghani M, Heigold G, Sun C, Lučić M, Schmid C. Vivit: A Video Vision Transformer. In: Proceedings of the IEEE/CVF International Conference on Computer Vision; 2021. p. 6836-46.
59. Papanastasiou G, Dikaios N, Huang J, Wang C, Yang G. Is attention all you need in medical image analysis? A review. IEEE J Biomed Health Inform 2024;28:1398-411.
60. Gholami P, Roy P, Parthasarathy MK, Lakshminarayanan V. OCTID: Optical coherence tomography image database. Comput Electr Eng 2020;81:106532.
61. Kulyabin M, Zhdanov A, Nikiforova A, Stepichev A, Kuznetsova A, Ronkin M, *et al.* OCTDL: Optical coherence tomography dataset for image-based deep learning methods. Sci Data 2024;11:365.
62. Kamran SA, Saha S, Sabbir AS, Tavakkoli A. Optic-Net: A Novel Convolutional Neural Network for Diagnosis of Retinal Diseases from Optical Tomography Images. In: 2019 18th IEEE International Conference on Machine Learning and Applications (ICMLA), IEEE; 2019. p. 964-71.
63. He K, Zhang X, Ren S, Sun J. Deep Residual Learning for Image Recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition; 2016. p. 770-8.
64. Sandler M, Howard A, Zhu M, Zhmoginov A, Chen LC. Mobilenetv2: Inverted Residuals and Linear Bottlenecks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition; 2018. p. 4510-20.
65. Chollet F. Xception: Deep Learning with Depthwise Separable Convolutions. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition; 2017. p. 1251-8.
66. Ancona M, Ceolini E, Öztireli C, Gross M. Towards better understanding of gradient-based attribution methods for Deep Neural Networks. In6th International Conference on Learning Representations: Arxiv-Computer Science; (ICLR) 2018 (No. 1711.06104, pp. 0-0).
67. Huang G, Liu Z, Van Der Maaten L, Weinberger KQ. Densely Connected Convolutional Networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition; 2017. p. 4700-8.