Neural Computing and Applications (2024) 36:9171–9188
https://doi.org/10.1007/s00521-024-09564-7

ORIGINAL ARTICLE

Automated retinal disease classification using hybrid transformer model (SViT) using optical coherence tomography images

G. R. Hemalakshmi[1] · M. Murugappan[2,3,4] · Mohamed Yacin Sikkandar[5] · S. Sabarunisha Begum[6] · N. B. Prakash[7]

Received: 3 November 2023 / Accepted: 24 January 2024 / Published online: 23 February 2024

Abstract
Optical coherence tomography (OCT) is a widely used imaging technique in ophthalmology for diagnosis and treatment. Recent advances in deep neural networks (DNNs) and vision transformers (ViTs) have paved the way for automated eye/retinal disease classifications and segmentations using OCT or spectral domain OCT (SD-OCT) images. Diabetic macular edema (DME), choroidal neovascularization (CNV), and Drusen are particularly challenging to accurately classify using OCT images because of their subtle differences and intricate features. Currently, the algorithms reported in the literature using DNNs or ViTs are computationally complex, consider fewer diseases, and are less accurate. This study proposes a hybrid SqueezeNet-vision transformer (SViT) model that combines the strengths of SqueezeNet and vision transformer (ViT), capturing local and global features of OCT images to achieve more accurate classification with less computational complexity. The proposed model uses the OCT2017 dataset for training, testing, and validation, and it performs both binary classification (normal vs disorders) as well as multiclass classification (DME, CNV, Drusen, and normal). As compared to state-of-the-art CNN-based and standalone Transformer models, the proposed SViT model achieves an overall classification accuracy of 99.90% for multiclass classification (CNV: 100%, DME: 99.9%, Drusen: 100%, and normal: 100%). With a good generalization ability, the model can be used to improve patient care and clinical decision-making across a broader range of applications.

Keywords OCT · Eye disorders · Retinal diseases · Classification · SqueezeNet · Vision transformer · Hybrid model

1 Introduction

In ophthalmology, early detection and diagnosis of retinal diseases is a key component of patient care. Retinal disease detection is one of the most challenging areas of biomedical optics research. A rapid, high-resolution imaging system is being developed that is capable of detecting disease-specific markers in the retina but hasn't met clinical needs yet. OCT is a non-invasive, contactless imaging modality that provides cross-sectional images of a variety of retinal abnormalities [1]. OCT images are frequently used in ophthalmology to assess Diabetic Macular Edema (DME), diabetic retinopathy (DR), and Choroidal Neovascularization (CNV). OCT has several advantages over other

✉ M. Murugappan
m.murugappan@kcst.edu.kw

[1] School of Computing Science and Engineering, VIT Bhopal University, Bhopal, Madhya Pradesh, India

[2] Intelligent Signal Processing (ISP) Research Lab, Department of Electronics and Communication Engineering, Kuwait College of Science and Technology, Block 4, Doha, Kuwait

[3] Department of Electronics and Communication Engineering, Vels Institute of Sciences, Technology, and Advanced Studies, Chennai, Tamilnadu, India

[4] Center for Unmanned Autonomous Systems (CoEUAS), Universiti Malaysia Perlis, Arau, Perlis, Malaysia

[5] Department of Medical Equipment Technology, College of Applied Medical Sciences, Majmaah University, 11952 Al Majmaah, Saudi Arabia

[6] Department of Biotechnology, P.S.R. Engineering College, Sivakasi 626140, India

[7] Department of Electrical and Electronics Engineering, National Engineering College, Kovilpatti 628503, India

imaging procedures, including real-time imaging, three-dimensional imaging without dilation of the patient, and the use of a harmless imaging probe. Additionally, OCT images can be used in the early diagnosis of retinal diseases by determining the underlying causes of the condition [2].

Classification of OCT images is a critical step in retinal image analysis for medical applications. This includes fully automated or semi-automated OCT image analysis or OCT image-based diagnosis [3]. The accurate classification of OCT images is essential for medical professionals to diagnose retinal diseases quickly and accurately. In contrast, manual interpretation of OCT images can take a long time and be prone to errors. Aside from clinical diagnosis by eminent ophthalmologists, artificial intelligence (AI) methods play an important role in retinal image diagnosis, and they produce results that are as similar as possible to those of an ophthalmologist. The development of automated deep-learning models for OCT image classification has become one of the hottest research topics in ophthalmology in the last few years [4–6]. In addition to being more efficient than conventional machine learning models, deep learning models extract more detailed information from input pathological images for better decision-making.

In the past few years, deep learning has become increasingly popular in retinal image classification based on Convolutional Neural Networks (CNNs) using OCT images [5, 6], transfer learning approaches [6, 7], capsule networks [4], and vision transformers [8]. Deep learning models are effective at the classification of OCT images because they can extract patterns and features from labeled images. As a result, the labeled images are fed into the model, which then learns a relationship between inputs and labels. As a result of the training, the model can be used to classify unknown OCT images based on learned patterns and features. In medical image classification applications, CNNs are the most commonly used type of DNNs [5, 9–11]. CNNs are used to automatically extract relevant features from images. As a result, CNN is especially effective when dealing with images that have a high dimensionality, like those found in medical settings. Over the past few years, several CNN-based image classification models have been proposed [12–14]. However, these models have some limitations, including a requirement for a large number of training images, an inability to capture long-range dependencies between features, and limited ability to incorporate prior knowledge of retinal structure.

Recent developments in OCT image classification have led to the development of vision transformer (ViT) [15] models to address the limitations of existing CNN or DNN-based models. The ViT model converts input images into token sequences that can be encoded by a transformer. ViT can capture long-range dependencies among image features, which contributes to its ability to classify and segment images effectively [20, 21]. The model has demonstrated remarkable performance in retinal image classification tasks [8, 16–19]. Due to its ability to capture long-range dependencies between features, ViT is most suitable for OCT image classification because it is more effective for tasks that require a global understanding of the image. The tokenization process of OCT images can also be enhanced by incorporating the unique structural characteristics. Hence, it is more efficient and cost-effective than other methods to train on a relatively small number of images [8, 16, 17].

Even though ViT performs well in medical image classifications or retinal image classifications, its performance depends only on the number of images and features used. The hybrid Transformer model combines the benefits of CNNs and Transformers to revolutionize image processing across a wide range of fields [16–19, 22]. A hybrid transformer combines the ability of CNNs to capture local spatial features with the ability of transformers to model long-range dependencies. Through these models, it is possible to process medical images efficiently and to capture long-range dependencies between features, which results in better accuracy in tasks like segmentation and classification [23].

In a recent study, Khan et al. [24] examine the interactions between ViTs and CNNs, emphasizing the emergence of hybrid models combining convolution operations and self-attention mechanisms. Besides discussing key components such as attention mechanisms, positional embeddings, and convolution operations, it provides a detailed taxonomy of hybrid architectures that can capture both local and global image features. A total of nine variants are presented in this context for integrating CNNs with ViTs. Nanni et al. proposed a CNN-based image classification system combining CNNs with Transformer models in [25]. In this paper, the ensemble approach is presented as a novel optimization algorithm, and it outperforms traditional optimization methods. The first-of-its-kind research demonstrates a significant improvement in detecting small/medium images by combining CNNs and transformers. Despite this, these models are often difficult to train because they have high computational requirements. Because there is a trade-off between computational efficiency and model accuracy, developing accurate hybrid models that are efficient and accurate at the same time is difficult. A lightweight, efficient model must balance accuracy and computational complexity in automated medical image-driven diagnosis.

The previous works had several major limitations: (i) a less accurate classification of retinal images, (ii) a higher computational complexity, (iii) ViT-based models require more training images for better classification, and (iv) fewer classes are considered for classification (retinal

diseases). Hence, we are motivated to develop a lightweight hybrid DNN-ViT model with less computational complexity and high classification accuracy for retinal image classification. In this study, we present a hybrid OCT image classification method that combines SqueezeNet and ViT for retinal image classification. In comparison to traditional CNNs, SqueezeNet is lightweight and achieves high accuracy with fewer parameters. This is accomplished through SqueezeNet's "fire modules" that reduce the number of parameters required for conventional CNNs. As a result of the research, the following contributions have been made.

1. This study proposes a lightweight hybrid model for OCT image classification called SViT, that extracts low-level and high-level features as well as global dependencies using SqueezeNet and Vision Transformer (ViT).
2. Compared to other hybrid or standalone retinal image classification models, the proposed hybrid model uses fewer parameters and requires a shorter inference time.
3. A hybrid model was developed for binary classification (normal and retinal diseases) and multi-class classification (normal, DME, CNV, and Drusen).
4. An explainable analysis was performed on the hybrid model to interpret its learned features and shed light on its significant features.

SViT is tested and trained on the publicly available OCT-2017 [26] dataset, which demonstrates superior performance metrics and computational efficiency compared to existing state-of-the-art models. In the remainder of this paper, we organize the information as follows. This paper is organized as follows: Sect. 2 presents a review of relevant literature, emphasizing the most recent studies on similar topics. Section 3 describes the dataset and methodology used in this study. We present the mathematical description of the classification problem and the model's architecture in Sect. 4. In Sect. 5, empirical findings, an explanation of the analysis, a graphic representation, and an ablation study are presented. In Sect. 6, we conclude this paper with future research directions.

2 Related works

Over the past several decades, deep learning techniques for classifying OCT images have evolved and diversified. According to Omid et al., they categorized twelve different types of pathologies based on the MedMNIST-2D database using the medical vision transformer (MedViT). The proposed CNN/vision transformer hybrid model has achieved a mean classification rate of 0.961 for four classes (DME, CNV, normal, and Drusen) using a limited number of parameters (45.8 M). Ma et al. propose a hybrid ConvNet-Transformer network (HCTNet) has been proposed for classifying retinal images into four classes from two datasets (Spectral Domain OCT image (SD-OCT) [6], OCT2017) [16]. With the OCT2017 dataset, they achieved a maximum classification rate of 91.56%, and with the SD-OCT dataset acquired in [6], they achieved a maximum classification rate of 86.18%. A hybrid CNN-Vit (Hybrid CNN-ViT) method was used for segmenting retina layers from OCT images [17]. In this hybrid network, called transformer segmentation network (TransSegNet), two datasets have been used to train and test it. They found that their proposed network performed better than fully convolutional networks (FCNs), segmentation networks (SegNets), and U-Nets in terms of segmentation performance. In accordance with TransSegNet, datasets 1 and 2 exhibit 88.28/82.40 precision, recall, and dice similarity coefficient (DSC) of 92.29/83.57, 89.76/81.48, respectively.

The ViT proposed by Jiang et al. al. [18] is used to classify three classes (AMD, DME, and normal) in OCT images, as reported. A comparison was conducted between the proposed ViT and conventional CNNs such as VGG16, ResNet50, DenseNet121, and EfficientNet. A maximum mean classification rate of 99.69% was achieved by the proposed ViT in classifying three classes with an inference time of 17 ms. A new method of classifying retinal diseases was developed using texture features extracted from OCT images by using InceptionV3 and ResNet-50 DNNs, as well as shape features extracted from vision transformers (ViT) [19]. In multiclass classification using images from the OCT2017 dataset, the maximum F1-score was 0.92 with an accuracy of 0.9237 achieved by the researchers. A deep multi-layered CNN was trained by Kuwayama et al. to categorize OCT images into healthy, dry age-related macular degeneration (AMD), wet AMD, and diabetic macular edema (DME). In a study, Islam et al. [28] and Li et al. [29] used deep transfer learning models to automatically diagnose diabetic retinopathy in OCT images, demonstrating the power of pre-trained networks in automatic diagnosis. A Multi-scale Deep Feature Fusion (MDFF) network developed by Das et al. [30] contributed significantly to this field by combining discriminative features with complementary information for more accurate classification.

In this study, Huang et al. [31] proposed a Layer-Guided CNN (LGCNN) to improve the classification of OCT images of normal retinas and common macular pathologies like CNV and DME. The ensemble learning models for retinal thickness assessment and classification developed by Cazaas-Gordón et al. [32] and Anoop et al. [33] built upon the success of CNN-based approaches. In a new study, Tsuji et al. [4] propose a capsule network based on OCT images to classify eye diseases. As a result, they

achieved a mean accuracy of 99.60% in classifications. The authors of [34] proposed using OCT images to construct a network for the classification of retinal images based on a fusion of networks. Three DNNs (Inception V2, Inception-ResNet, and Xception) have been combined to identify retinal images as normal, CNV, DME, and Drusen. As a result of their proposed fusion network, they were able to achieve a maximum mean classification rate of 98.7% with an AUC of 99.1%. This network also achieved a maximum recall and specificity of 0.987 and 0.996, respectively. Many factors are limiting the use of CNN-based OCT image classification, such as their high computational complexity and long training times, which make it unsuitable for real-time clinical applications. Moreover, they require extensive annotated datasets, making it difficult to obtain expert-labeled data [35]. It is difficult to interpret CNNs because of their black-box nature, resulting in a lack of acceptance among medical professionals. The inconsistency of OCT devices, inconsistencies in acquisition protocols, and variations in image quality also compromise their performance, emphasizing the necessity of robust models. It is imperative that CNNs be able to overcome these limitations in clinical settings if they are to be used more widely.

In terms of OCT image classification, the ViT algorithm [36] might be able to address the limitations of CNNs. As images are treated as sequences of tokens, ViTs allow an understanding of the global context, which improves model interpretability and addresses the black-box issue. Due to their ability to learn with fewer samples, ViTs reduce the need for large, annotated datasets by incorporating self-attention mechanisms. Moreover, ViTs can be trained on large-scale external datasets and fine-tuned for specific tasks using transfer learning. As a result of their scalability and computational efficiency, ViTs are suitable for real-time clinical applications despite variations in image quality and acquisition protocols.

ViTs have been widely employed to detect eye disorders and other image classification problems. The classification of glaucomatous eye conditions based on fundus images by ViT-based ensembles is effective by Wassel et al. [37]. In this paper, they evaluated several vision transformer models and suggested an ensemble approach based on six publicly available datasets. The best standalone model of ViTs has a sensitivity of 92.57%, a specificity of 96.94%, and an AUC of 97.9%, which indicates that they can be used to diagnose glaucoma. Using state-of-the-art technology, Fan et al. [38] examined the applicability and interpretability of ViT for glaucoma detection. These comprehensive experimental results demonstrate strong generalization across diverse ethnic groups represented in the external test data, according to the authors. In addition, the ViT used in this study could locate the neuroretinal rim and detect glaucoma based on its features.

OCT images present numerous challenges in the detection and classification of eye disorders, including image quality issues, high dimensionality, heterogeneous appearances, imbalanced data, inter- and intra-observer variability, as well as interpretability and generalizability. By leveraging self-attention mechanisms that capture global context and relationships, ViTs can meet these challenges effectively with high-dimensional OCT images. Using the Transformer architecture, ViTs can handle heterogeneity in appearance and provide better interpretability through attention maps, making their decision-making process more understandable. Furthermore, ViTs can benefit from pre-training on large datasets, reducing the scarcity of large, annotated OCT datasets and improving the generalization of models across multiple device types and patient populations. Consequently, ViTs is capable of providing a more robust and reliable solution for detecting and classifying eye disorders based on OCT images. Wen et al. [39] developed a novel Lesion Localization Convolution Transformer (LLCT) for ophthalmic disease classification and lesion detection using OCT images. As compared to conventional deep learning architectures, LLCT combines convolution and self-attention mechanisms to improve classification accuracy, sensitivity, and specificity. As a result of the experiments, overall accuracy improved by 7.6%, sensitivity improved by 10.9%, and specificity improved by 9.2%. Furthermore, LLCT successfully localized lesions in retinal OCT images without labeling them.

According to He et al. [40], an interpretable Swin-Poly Transformer network was developed to automate retinal OCT image classification. A network forms connections between adjacent non-overlapping windows in the preceding layer by adjusting the window partition, allowing features to be modeled across several scales. Furthermore, by altering the significance of polynomial bases, the Swin-Poly Transformer improves classification by fine-tuning cross-entropy. Additionally, a confidence score map is provided as part of this approach to guide medical professionals in understanding how the model makes its decisions. Based on the OCT2017 dataset, this model shows an overall classification accuracy of 99.80% for CNV, DME, Drusen, and Normal cases. The authors also evaluated this dataset using ViT, Swin Transformer [41], and LLCT, all of which showed marginally lower performance.

Swin Transformer's complexity results in higher computational costs and memory requirements, which could inhibit its use in real-world scenarios, especially when dealing with large datasets or resource-constrained devices. Furthermore, Swin Transformer adjusts window partitions,

so the choice of window size and partitioning strategy may influence its effectiveness. Considering these limitations, hybrid transformer models, such as SqueezeNet-ViT, appear to be a compelling option. Through the combination of SqueezeNet and Vision Transformer strengths, a hybrid approach is intended to achieve high classification performance while maintaining a lightweight network structure.

3 Materials and methods

The proposed hybrid model (SviT) for retinal image classification using OCT images is shown in Fig. 1. In this section, the dataset and underlying methods used in this investigation are described, allowing replication of the experiments and validation of the results. In addition to providing information about the nature, origin, quality, and preparation of the dataset, the comprehensive section provides the context needed for accurate interpretation and application of the results.

3.1 Dataset and preprocessing of OCT images

In this research, the proposed SviT is trained and tested on the OCT2017 [42] dataset. According to Table 1, this dataset contains images of four different classes: CNV, DME, Drusen, and Normal. The images are all JPEG formats with different dimensions, which are resized to 227 × 227 as required by SqueezeNet.

3.2 SqueezeNet model

The SqueezeNet architecture is a lightweight deep CNN design developed to reduce parameters while maintaining competitive performance on image classification tasks [11]. SqueezeNet's architecture is shown in Fig. 2 and summarized as follows:

1. **Initial convolutional layer**: A single 3 × 3 convolutional layer with 96 filters and has a stride factor of 2.
2. **Max-pooling layer**: A 3 × 3 max-pooling layer with a stride factor of 2.
3. **Fire modules (FM)**: A sequence of 8 Fire modules (Fire2 to Fire9), each consisting of a squeeze layer and an expand layer.
   - Squeeze layer: A 1 × 1 convolutional layer with a smaller number of filters. This layer is responsible for reducing the number of channels in the input feature map.
   - Expand layer: A combination of 1 × 1 and 3 × 3 convolutional layers that increase the number of channels, effectively expanding the compressed feature map obtained from the squeeze layer.
4. **Final convolutional layer**: A 1 × 1 convolutional layer with a number of filters equal to the number of classes in the classification task (e.g., 1000 filters for ImageNet).
5. **Global average pooling (GAP) layer**: A global average pooling layer that reduces the spatial dimensions of the feature map to 1 × 1.
6. **Softmax activation**: A softmax activation function applied to the output of the global average pooling layer to produce class probabilities.

The following steps are taken to incorporate SqueezeNet for feature extraction:

1. **Removal of the final classification layers:** The architecture of SqueezeNet has been modified to remove the final convolutional layer, global average pooling layers, and softmax layers. In this way, high-level feature maps can be generated before the feature extractor is able to classify them.
2. **Processing of input images:** OCT images are fed into the modified SqueezeNet architecture, which produces feature maps at the highest level. Feature maps constructed with SqueezeNet retain their efficiency and compactness while capturing important input image information.
3. **Integration with ViT:** A feature map is imported into ViT after being extracted from SqueezeNet. In this step, feature maps are flattened and divided into nonoverlapping patches, which ViT linearly embeds and processes.

3.3 Vision transformer

A ViT architecture for image classification includes embedding, position encoding, and Transformer encoder layers [8]. It consists of several sub-layers, including Multi Head Self Attention (MHSA), Feed Forward Network (FFN), and Layer Normalization (LN). Below is a description of the key components of the ViT, based on the schematic shown in Fig. 3.

1. **Input image processing:** First, the input image is resized to a fixed resolution and divided into equal-sized non-overlapping patches. By using a linear projection layer, patches are flattened into 1D vectors and linearly embedded into continuous representations. Hence, a sequence of embeddings of fixed-size patches is generated.
2. **Positional encoding:** A positional encoding is added to patch embeddings to incorporate spatial information. It uses these encodings to distinguish patches based on their positions in the input image, and they can either be learned or fixed.

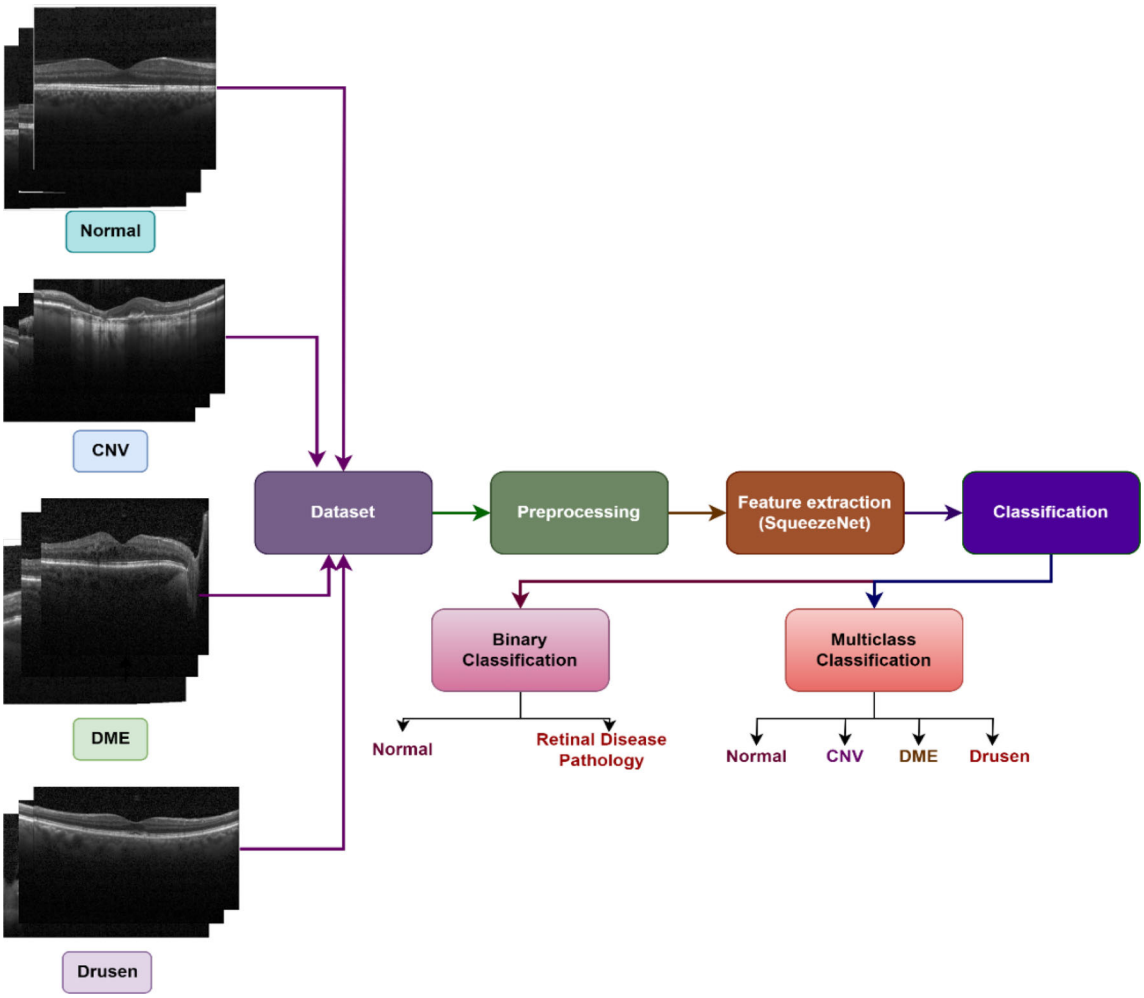


Fig. 1 Overall proposed methodology for retinal image classification

Table 1 OCT2017 Data Distribution [27]

Classes	No. of training images	No. of testing images
CNV	37205	250
DME	11348	250
Drusen	8616	250
Normal	26315	250
Total	83484	1000

3. **Transformer encoder layers:** This sequence of patch embeddings is fed into a Transformer architecture, which consists of identical encoder layers. There are multiple sub-layers in an encoder layer, including a multiheaded self-attention (MHSA) layer, a feedforward layer, a normalization layer, and a dropout layer. During the multi-head attention layer, relationships between patches are captured, while in the feedforward layer, embeddings are transformed nonlinearly. Additionally, the LN stabilizes the training process and prevents overfitting. Finally, the output is produced by the encoder's last normalization layer.
4. **Classification head:** During the image classification task, the embedding associated with the first position (usually called a "classification token") is used to generate the output probabilities. It is usually achieved by applying a linear layer followed by a softmax activation function.

ViT has demonstrated strong performance in image recognition, rivaling or even surpassing traditional CNNs. Furthermore, the model's performance improves as the size and scale of the dataset increase, proving that it is capable of efficiently utilizing large datasets and computational resources.

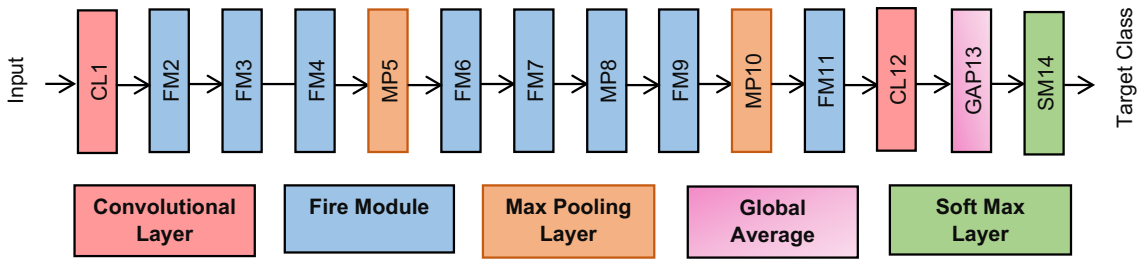


Fig. 2 SqueezeNet architecture

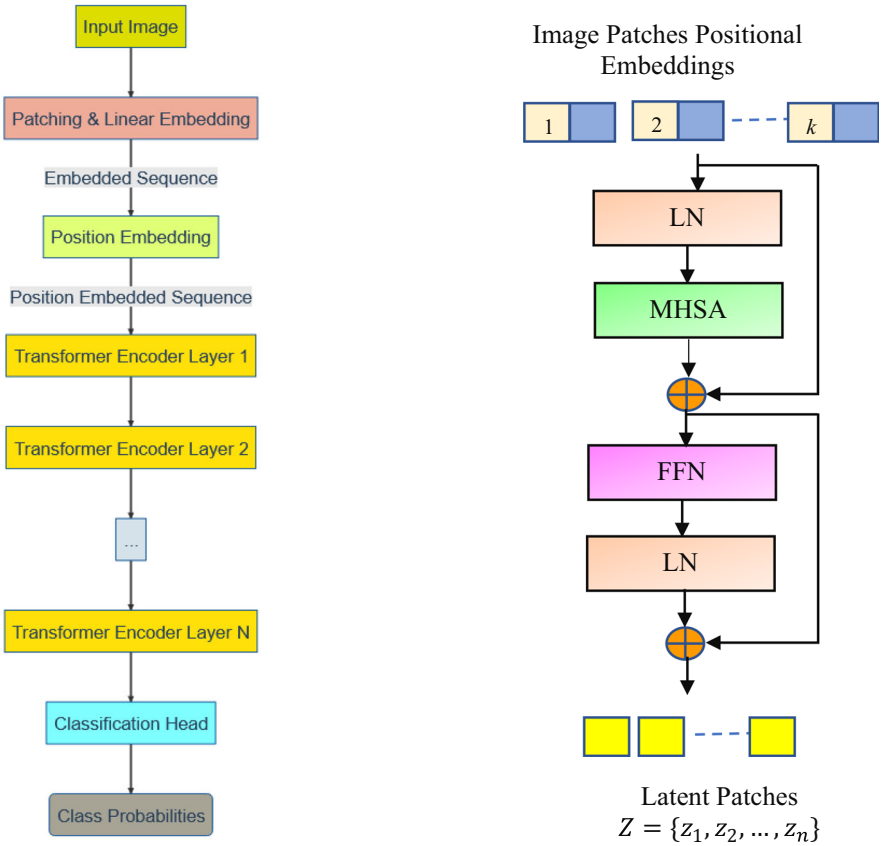


Fig. 3 **a** ViT architecture for image classification, **b** transformer encoder

a. ViT Architecture for Image Classification **b.** Transformer Encoder

4 Proposed OCT image classification model

The mathematical formulation of the proposed OCT classification problem and the architecture of the SViT classifier are described in detail in this section. The proposed SViT classifier is illustrated in Fig. 4.

4.1 Problem definition

The OCT classification problem is modeled as a multiclass classification problem. Given a training dataset $\mathrm{Tr} = (x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$, where x_i is an OCT image and $y_i \in \{\mathrm{CNV}, \mathrm{DME}, \mathrm{Drusen}, \mathrm{Normal}\}$ is its corresponding label, the goal is to learn a function $f : X \rightarrow Y$ that maps each input image x_i to its correct label y_i. The function f is trained to minimize a loss function defined as the average of the per-class cross-entropy loss $\mathcal{L}_{\mathrm{CLS}}$ and the regularization loss $\mathcal{L}_{\mathrm{REG}}$. The loss function is given as in Eq. (1), where f_θ is the learned mapping function, λ is the regularization coefficient and n is the number of samples.

$$L(\theta, \mathrm{Tr}) = \frac{1}{n}\sum_{i=1}^{n} \mathcal{L}_{\mathrm{CLS}}(x_i, y_i, f_\theta) + \lambda \mathcal{L}_{\mathrm{REG}}(f_\theta) \quad (1)$$

The loss function $\mathcal{L}_{\mathrm{CLS}}$ is calculated using cross entropy, which is defined as in Eq. (2). The loss function $\mathcal{L}_{\mathrm{REG}}$ is to

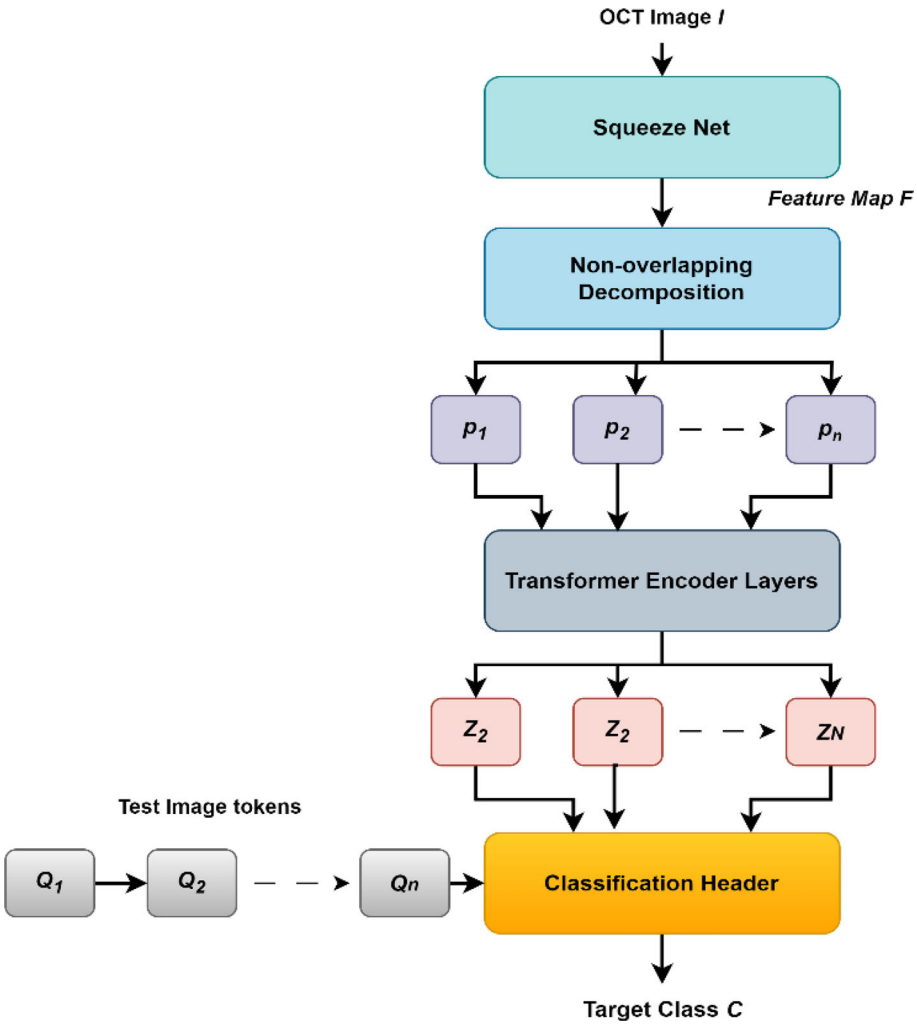


Fig. 4 SqueezeNet-ViT classifier model

encourage the network to learn compact and non-overlapping regions, defined as in Eq. (3).

$$\mathcal{L}_{\mathrm{CLS}}(x_i, y_i, f_\theta) = -\frac{1}{n}\sum_{i=1}^{n} y_i \log(f_\theta(x_i)) \tag{2}$$

$$\mathcal{L}_{\mathrm{REG}}(f_\theta) = \frac{1}{2n}\sum_{i=1}^{n} \|f_\theta(x_i) - x_i\|_2^2 \tag{3}$$

4.1.1 Local feature extraction with SqueezeNet

A local feature extraction process using SqueezeNet is described in this section. Let I be an input image of size $H \times W \times C$, where H is the height, W is the width, and C is the number of channels. Image transformations through various layers are given in the following steps.

i. **Convolutional layers:** The image I is passed through several convolutional layers with various filter sizes and strides. Layers that extract local features from an input image apply convolution operations between the input image and the learnable filters. The operation can be described as in Eq. (4), where $*$ is the convolution operation, F is the filter size of $m \times n$, i and j are the spatial indices of the output feature map.

$$(I * F)(i,j) = \sum_m \sum_n I(i-m,\ j-n) * F(m,n) \tag{4}$$

ii. **Fire module:** Several building blocks make up SqueezeNet, but the Fire module is the most important. The structure is composed of two layers, one squeezed and one expanded.

   a. **Squeeze layer:** Squeeze layers apply 1×1 convolutions to reduce the number of channels in the feature maps, thus compressing the feature representation. It generates the output calculated using Eq. (5), where W_{sqz} and b_{sqz} are the learned weights and biases of the convolutional filters.

$$O_{\text{sqz}} = \text{Conv1} \times 1\left(I,\ W_{\text{sqz}},\ b_{\text{sqz}}\right) \tag{5}$$

b. **Expand layer**: Expand layers use a combination of 1 × 1 and 3 × 3 convolutions to expand the number of channels in feature maps, resulting in richer feature representations. The output of this layer is represented mathematically by Eq. (6), where W_{exp1}, b_{exp1}, W_{exp3} and b_{exp3} are the learned weights and biases of the 1 × 1 and 3 × 3 convolutional filters, respectively.

$$O_{\text{exp}} = \text{Concat}\left(\text{Conv1} \times 1\left(O_{\text{sqz}}, W_{\text{exp1}}, b_{\text{exp1}}\right), \text{Conv3} \times 3\left(O_{\text{sqz}}, W_{\text{exp3}}, b_{\text{exp3}}\right)\right) \tag{6}$$

iii. **Pooling layers**: Following transformations by the fire modules and convolutional layers, the output of the final convolutional layer is input to the GAP layer to reduce the spatial dimensions of the feature maps. The output of the GAP is shown in Eq. (7).

$$F_{\text{map}} = \text{Pool}(\text{Conv1} \times 1\left(O_{\text{exp}}\right) \tag{7}$$

4.1.2 Global feature extraction with vision transformer (ViT)

In this section, we describe the process of extracting global features from SqueezeNet using Vision Transformer (ViT). Let F_{map} be the input feature map of size $H' \times W' \times C'$, obtained from the SqueezeNet as described in Eq. (7). The target label is assigned to the input image$I$, by processing the F_{map} as below.

i. **Patch extraction and tokenization**: The input feature map is divided into non-overlapping patches of size $P \times P$, resulting in a total of $(N = H'W')/P^2$ patches. Each patch is then linearly embedded (flattened) into a 1D vector of length D, where D is the dimension of the transformer's input as in Eq. (8).

$$x_p = \text{Flatten}\left(\text{Patch}\left(F_{\text{map}}\right)\right) \tag{8}$$

ii. **Positional encoding**: Positional embeddings are concatenated with the patch embeddings to retain spatial information as in Eq. (9). The combined patch and positional embeddings form the input sequence for the transformer.

$$x'_p = x_p \odot \text{PE}_p \quad \forall p = 1, 2, \cdots, N \tag{9}$$

iii. **Transformer encoder:** The input sequence $x'_1, x'_2, \cdots, x'_N$ is fed into the transformer encoder, which consists of multiple layers of multi-head self-attention and feed-forward networks. The output of the transformer encoder is a sequence of transformed feature vectors as in Eq. (10).

$$z_p = \text{TE}\left(x'_p\right) \forall p = 1, 2, \cdots, N \tag{10}$$

iv. **Pooling and classification**: A pooling operation, such as mean pooling or attention pooling, is applied to the sequence of transformed feature vectors to obtain a single global feature vector z_{global} as in Eq. (11). This global feature vector is then passed to the classification head to perform multiclass classification as in Eq. (12). The output y_{pred} represents the predicted class probabilities for the input OCT image.

$$z_{\text{global}} = \text{Pool}(z_1, z_2, \ldots, z_N) \tag{11}$$

$$y_{\text{pred}} = \text{MLP}\left(z_{\text{global}}\right) \tag{12}$$

5 Experimental results and discussions

In this section, we discuss the experimental setup and empirical evaluations of the SViT model using training and testing datasets. An evaluation is conducted using objective performance metrics, visualizations, and comparisons with other approaches. In addition, an explainable analysis is conducted to better understand the model's behavior, and an ablation study is conducted to demonstrate the effectiveness of the method.

5.1 Experimental setup

SViT is implemented in MATLAB R2023a on an Intel Core i10 with 64 GB of RAM and an NVIDIA GeForce RTX 3090 GPU. The implementation of NVIDIA CUDA and its cuDNN library results in significant improvements in training time and overall performance. Based on the OCT2017 dataset, a meticulously designed SViT model is trained, tested, and validated using hyperparameters listed in Table 2. An initial set of parameter values is selected after evaluating the model with the data subset to select the most appropriate hyperparameters. To assess the model's generalization capabilities, these initial values are used to train and validate the model. By exploring various combinations of values with grid search, optimal hyperparameters are identified systematically. The model's performance is closely monitored throughout the search process, and the best-performing hyperparameters are selected. In this way, the model configuration is tailored to the dataset and task at hand, so that it is both effective and

robust. In training, the training dataset is divided by 70:30 between training and validation groups.

A preliminary assessment of the effectiveness of the SViT's learning process is made by examining the training process for different hyperparameters. In Fig. 5, the training progress for one epoch is shown, showing an initial validation accuracy of 90.27% for the model.

5.2 Performance evaluation

In Eqs. (13)–(17), we evaluate the model's accuracy, specificity, sensitivity, precision, and F1 metrics. The True Positive (TP), the True Negative (TN), the False Positive (FP), and the False Negative (FN) values are obtained from evaluating the model on the test dataset. According to Eq. (13), accuracy is the proportion of instances that were correctly classified out of all instances in the dataset. True negative rate (TNR), or specificity, is a measure of whether the model identifies negatives correctly. Sensitivity is also known as True Positive Rate or recall, which is a measure of how well the model identified actual positives (Eq. 15). Positive predictive value (PPV) measures the number of samples that are positive among those predicted to be positive as in Eq. (16). To balance precision and recall, the F1 score, the harmonic mean of precision and recall, is used. Performance is measured on a scale of 0–1, with higher values indicating better performance (Eq. 17). In addition, AUC is another metric used to evaluate classification model performance. The area under the curve (AUC) value ranges from 0 to 1, where 0.5 represents a random classifier, and 1 indicates an excellent classifier.

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \tag{13}$$

$$\text{Specificity} = \frac{\text{TN}}{\text{TN} + \text{FP}} \tag{14}$$

$$\text{Sensitivity} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{15}$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{16}$$

$$\text{F1} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{17}$$

5.3 Classification performance

In this study, the performance of SViT is evaluated with the subsets of training and testing data under binary classification and multiclass classification. Table 3 provides objective metrics for binary, multiclass, and class-wise detections. According to these findings, the SViT model is exceptionally good at both binary and multiclass classification on the OCT dataset, as evidenced by its high scores across all evaluation metrics. Figures 6 and 7 illustrate the confusion matrices corresponding to binary and multiclass classifications, respectively. The analysis of confusion matrices provides deeper insights into the model's classification accuracy since they reveal correctly predicted classes and the distribution of misclassifications. Among the Disorder and Normal categories, there are one and two misclassifications based on binary classification. Multiclass classification shows only one misclassification from CNV to DME, while the other classes are correctly classified. In this way, the classifier was able to learn disorders with a good degree of accuracy without raising false alarms.

Additionally, Receiver Operating Characteristics (ROC) curves are valuable tools for assessing the trade-offs between true positives and false positives across different decision thresholds. According to Figs. 8 and 9, the ROC curves show that the model reaches its best classification performance at around 0.95.

We compare the proposed SViT model with the state-of-the-art models on the OCT2017 dataset to demonstrate its effectiveness and competitiveness. A comparison of the

Table 2 SViT optimal hyperparameters used for retinal image classification

Parameter	Values
Maximum number of epochs	100
Learning rate	0.001
Batch size	32
Optimizer	SGDM
Momentum	0.9
Loss function	Cross-entropy loss
Learning rate schedule	Step decay
Weight initialization	Random
Regularization	L1
No. of transformer layers	3
Stopping criterion	Validation loss does not decrease further

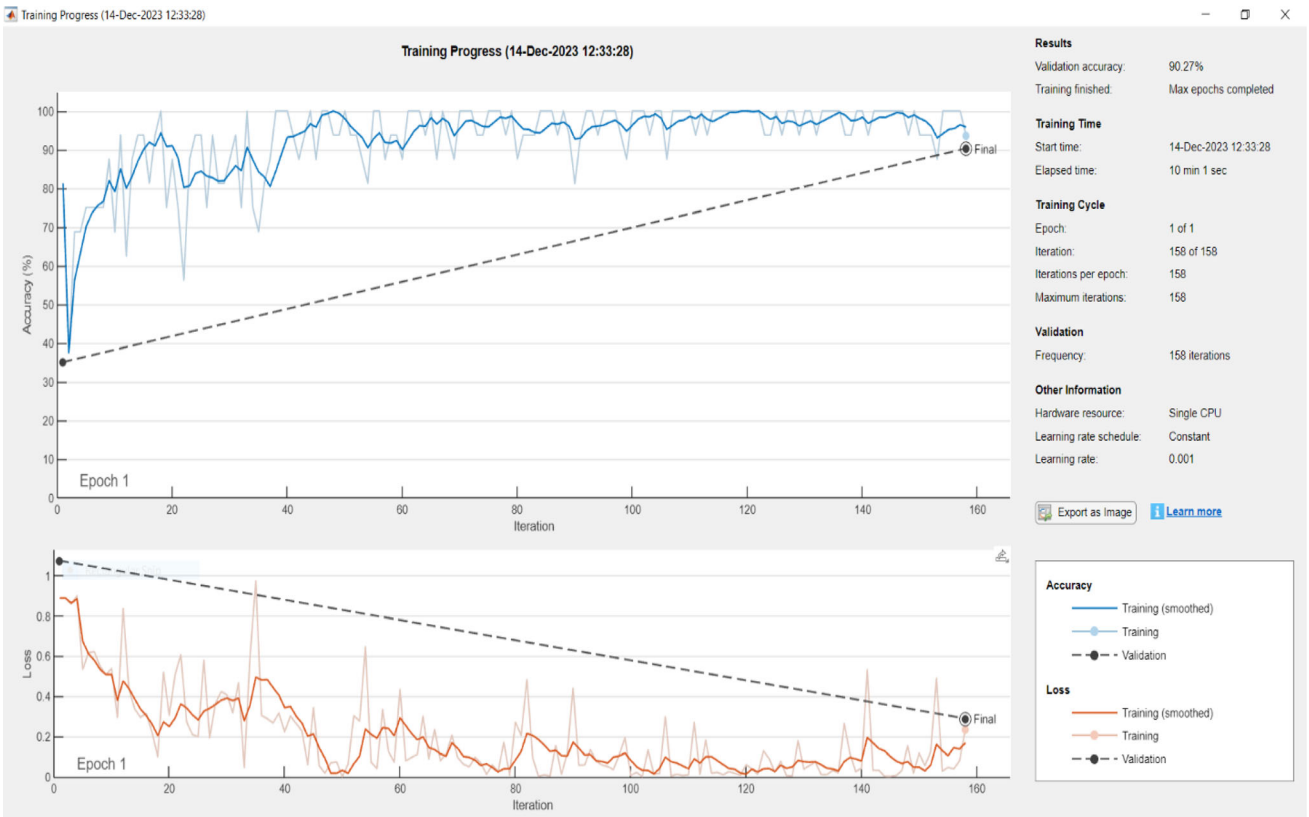


Fig. 5 SViT training progress

Table 3 Performance of retinal image classification using SViT

Classification	Accuracy	Sensitivity	Specificity	Precision	F1	AUC
Binary	0.9970	0.9987	0.9920	0.9973	0.9980	0.9896
Multiclass (Overall)	0.9990	0.9990	0.9997	0.9990	0.9990	0.9961
CNV	0.9960	0.9960	1.0	1.0	0.9980	0.9991
DME	1.0	1.0	1.0	1.0	1.0	1.0
DRUSSEN	1.0	1.0	1.0	1.0	1.0	1.0
Normal	1.0	1.0	1.0	1.0	1.0	1.0

multiclass classification performances of representative models reported in the literature is presented in Table 4.

Most of the earlier studies reported in the literature have used the OCT2017 dataset. The majority of researchers have considered binary classifications (normal vs retinal disorders) and few have considered multiclass classifications (DME, CNV, Drusen, and normal). Compared to conventional DNNs, hybrid models and hybrid DNNs with ViT produce higher accuracy in binary as well as multi-class classification. A multi-class classification using the HCT-Net produced the lowest accuracy (0.9156) of the different works reported in Table 4. It is important to note, however, that the inference time required for classification is much less than in all of the earlier studies, including the present one. In comparison to the other models, the SViT has some differences in performance metrics. The Swin Transformer model lags behind the proposed SViT model in mean accuracy, sensitivity, and precision by 0.0190, 0.0201, and 0.0090, respectively. In the LLCT model, there are differences of 0.0120 in mean accuracy, 0.0169 in mean sensitivity, and 0.0147 in mean precision. Even though Swin-Poly Transformers provide impressive performance, on average their accuracy, sensitivity, and precision are still 0.0008, 0.0009, and 0.0022 behind those of the SViT. In retinal OCT image classification tasks, this numerical difference indicates SViT's superior performance.

As a result of the synergistic combination of SqueezeNet and ViT, the SViT model performs exceptionally well. In addition to offering an efficient and compact architecture, SqueezeNet effectively reduces the size of the model without compromising accuracy. In this way, the model is

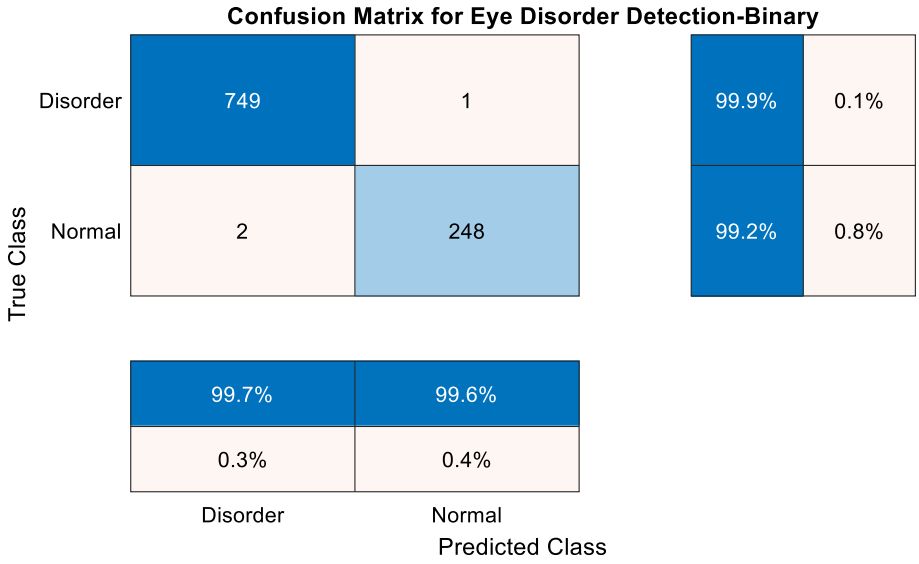


Fig. 6 Confusion matrix for binary classification

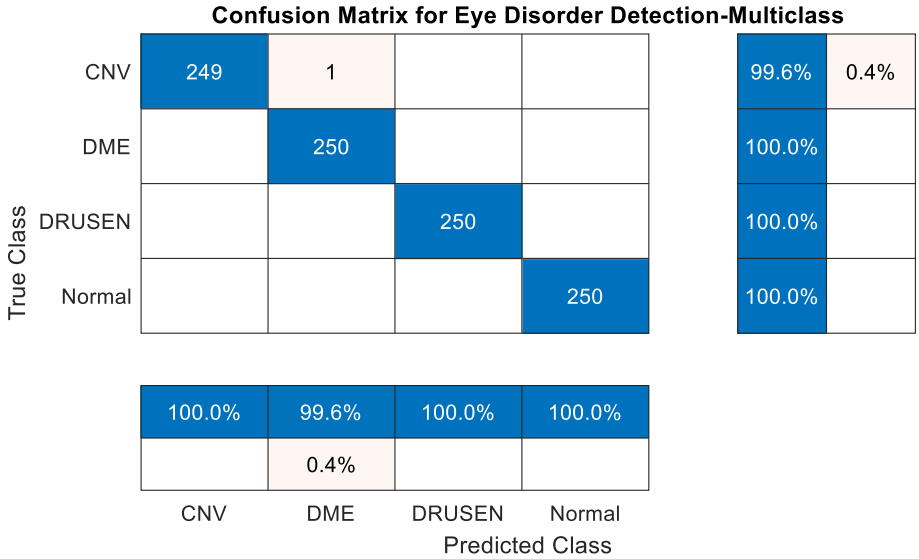


Fig. 7 Confusion matrix for multiclass classification

capable of capturing important local features and processing complex OCT images. Furthermore, the ViT component enhances the model's ability to recognize global contextual information as well as intricate patterns in retinal OCT images. As a result of integrating the strengths of SqueezeNet and ViT, the SViT model achieves exceptional performance in retinal OCT image classification tasks, outperforming both models individually. Furthermore, SViT's inference time is the shortest among the other models after analyzing the inference time. The SViT's performance is attributed to the Transformer's ability to make quick and accurate predictions, based on the SqueezeNet feature maps.

5.4 Explainable analysis

The Grad-CAM method uses Explainable Artificial Intelligence to provide insight into the decision-making process of deep learning models such as SViT, which combines SqueezeNet and ViT. In addition to creating easily understandable class activation maps based on input images, clinicians are able to confirm that model predictions match medical knowledge by reviewing the class activation maps [43]. Grad-CAM validates predictions with meaningful information by visualizing the final convolutional layer in SqueezeNet. Grad-CAM's flexibility and application to different architectures make it ideally suited to XAI analysis of the SViT model, which ultimately contributes to better patient outcomes.

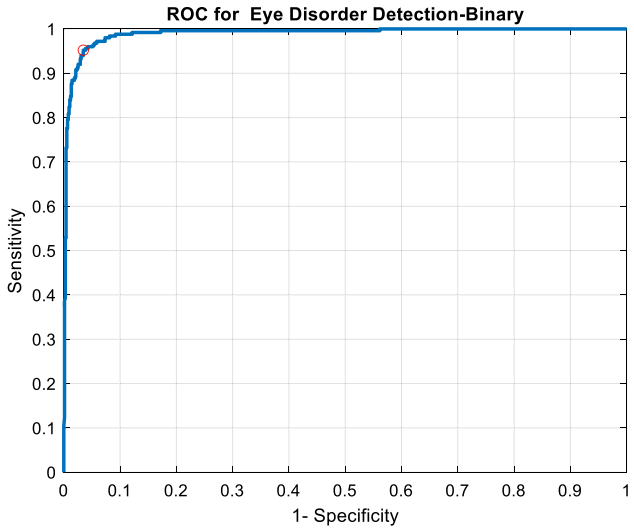


Fig. 8 ROC for binary classification

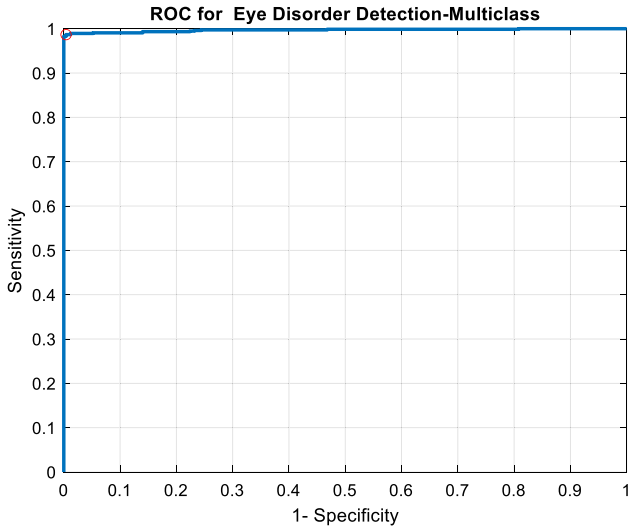


Fig. 9 ROC for multiclass classification

This procedure involves selecting the final convolutional layer in the SqueezeNet portion of the SViT model, performing a forward pass on the OCT image, and computing the gradients of the predicted class score using the feature maps of the target layer. Grad-CAM weights are calculated and used to create weighted activation maps that are then resized to match the input OCT image dimensions. A final step involves overlaying the resized class activation map on the original OCT image to highlight the most significant areas. This assists clinicians in verifying the model's predictions. Figure 10 shows the heat maps and classification scores for Normal, CNV, DME, and Drusen classes based on XAI. It is evident that the model is capable of localizing the infected areas accurately. Based on the results of this analysis, it can be concluded that the model is highly reliable.

5.5 Ablation study

Ablation studies were conducted to compare the performance of the SViT model with various components and design choices. The analysis provides insight into the most important factors contributing to the model's success in OCT image classification. During the ablation study, the following variations were taken into account. This study's results are summarized in Table 5 along with those from the SViT.

1. SqueezeNet only: SqueezeNet architecture was used without the ViT component.
2. ViT only: A standalone ViT model was used without SqueezeNet.

There are substantial differences in performance between the SViT model and its components, according to the ablation study. Based on performance metrics, the SqueezeNet-only model performs worse than the ViT-only model, indicating the limitations of SqueezeNet alone for OCT image classification. Due to its powerful global context capabilities, the ViT-only model outperforms the conventional model. Based on the results of the ablation study, there was a significant difference in the performance metrics among the three model variants. Compared to SViT alone, SqueezeNet increased mean accuracy by 3.9%, sensitivity by 4.8%, precision by 3.5%, and F1-score by 4.15%. Comparison between the ViT-only variant and the SViT-only variant. Even more pronounced increases were observed: 2.4% in mean accuracy, 3.2% in mean sensitivity, 2.1% in mean precision, and 2.65% in mean F1-score. By combining SqueezeNet and ViT, the hybrid SViT model significantly improves the classification performance of OCT images.

5.6 Discussions

Based on retinal OCT images, a hybrid SqueezeNet-Vision Transformer (SViT) model is proposed to classify eye disorders such as CNV, DME, Drusen, and normal cases. In many cases, these disorders exhibit subtle differences and intricate characteristics that can be challenging to distinguish, making their correct classification essential for clinical decision-making. Our objective metrics demonstrate significant improvements in performance compared to other state-of-the-art models. A hybrid architecture that combines the strengths of both SqueezeNet and ViT is key to the SViT model's performance. As a result of the improved model, both local and global features can be captured from the OCT images, which leads to more accurate classifications. In addition, the SViT model exhibits greater robustness and generalization, showing

Table 4 Comparison with State-of-the-art retinal image classification using OCT2017 dataset with ViT and Hybrid ViT networks

References & Year	Model	Database	No of classes	Mean accuracy	Mean sensitivity	Mean precision	Mean F1	Model characteristics
Omid et al. [8] (2023)	MedViT	OCT2017	4	0.9630	N/A	N/A	N/A	Parameters: 45.6 M
Ma et al. [16] (2022)	HCTNet	OCT2017	4	0.9156	0.8857	0.8811	N/A	Inference time: 3.74 ms
Dutta et al. [19] (2023)	Conv-ViT	OCT2017	4	0.9237	N/A	N/A	0.92	N/A
Ai et al. [34] (2020)	ViT	OCT2017	4	0.9906	0.9917	0.9921	0.9907	Parameters:86 M Inference time: 26.9 ms
He et al. [40] (2023)	SwinT	OCT2017	4	0.9801	0.9799	0.9910	0.9824	Parameters: 88 M Inference time: 35.3 ms
	LLCT	OCT2017	4	0.9870	0.9821	0.9853	0.9824	Parameters: $\sim$ 89 M Inference time: 41 ms
	Swin-PolyT	OCT2017	4	0.9982	0.9981	0.9978	0.9987	Parameters:88 M Inference time: 10.9 ms
Ours	**SViT**	**OCT2017**	**4**	**0.9990**	**0.9990**	**1.0000**	**0.9995**	**Parameters: 87.1 M** **Inference time: 6.9 ms**

The bold values refer to the highest performance achieved in the proposed method

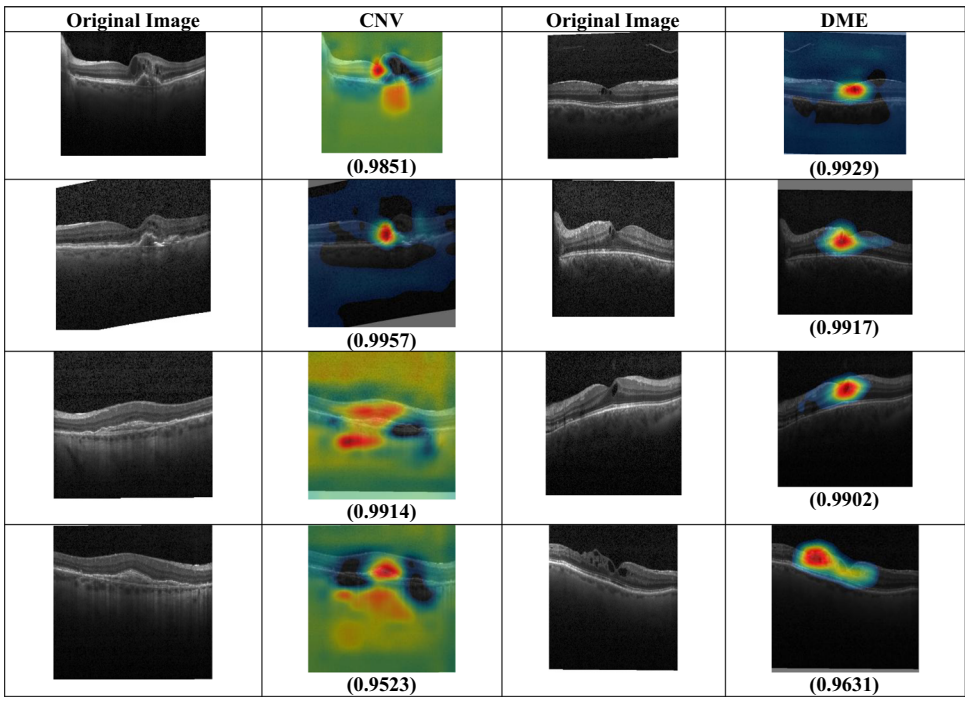
N/A: not available/not reported

that it can be applied to a variety of challenging cases, including CNV and DME, which are likely to have minute differences.

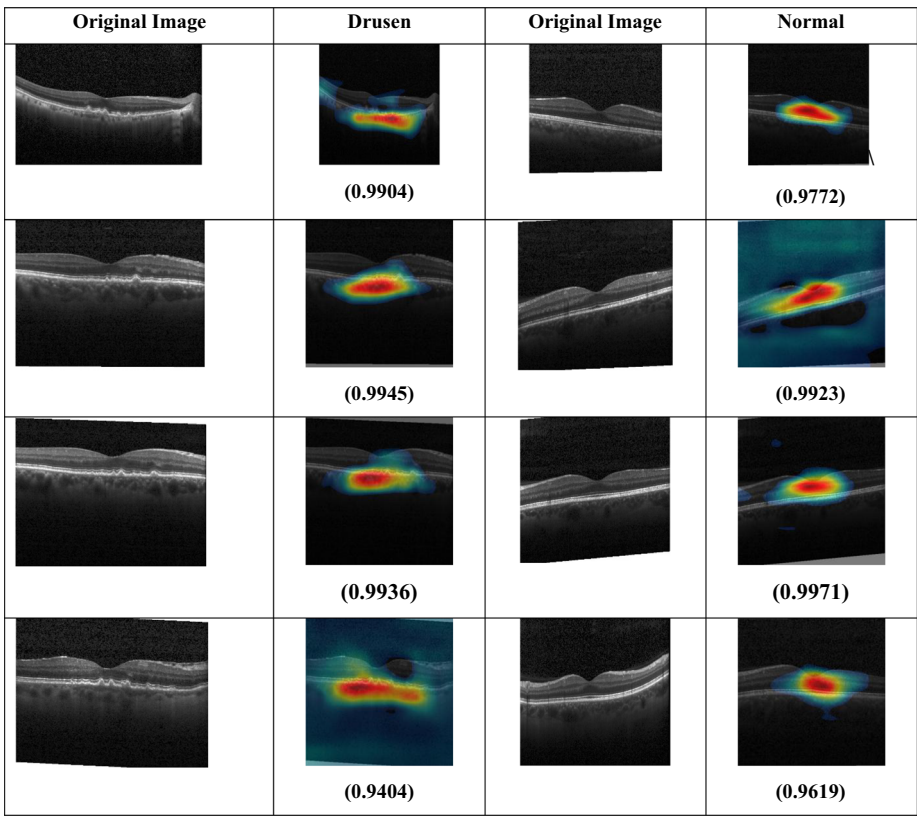
A comparison of the SViT model with other representative works shows that the SViT model represents a significant advance in the detection of retinal disease using OCT images:

1. **Focused Optimization for OCT Images:** The SViT model was specifically designed to optimize OCT images, unlike the general ViTs combined with CNNs described in [17]. By taking a tailored approach, it is able to effectively address the challenges unique to retinal diseases, such as Diabetic Macular Edema (DME), Choroidal Neovascularization (CNV), and Drusen. OCT image analysis model does not simply combine existing technologies but tailors the integration to suit the specific needs of the application.
2. **Seamless Integration for Enhanced Feature Extraction:** The SViT model extends beyond the ensemble of CNNs and transformers explored in [18]. In our proposed work, SqueezeNet's compact and efficient architecture has been seamlessly integrated with ViT's feature extraction capabilities. The integration does more than juxtapose two architectures; it enhances the model's capability to extract local and global features from OCT images. It is particularly important in retinal disease detection, where the distinction between various conditions is often based on minute, but critical, image details.
3. **Superior Computational Efficiency and Accuracy:** SViT leverages SqueezeNet's lightweight architecture, known for computational efficiency. As a result of this choice, the SViT model is both fast and resource-efficient, a significant advantage in clinical settings where quick and accurate diagnosis is crucial. Moreover, the model's exceptional accuracy of 99.90% in multiclass classification, as compared to the ConViT [19], highlights its superiority. An increased level of accuracy, especially in a field like medical imaging, can improve patient outcomes and increase the reliability of clinical decision-making.

The model's focus on localized neuroretinal rim features, as seen in the explainable analysis, aligns with clinical practice in glaucoma management and facilitates differentiation between pathological and normal cases. An ablation study examines the impact of individual components in the SViT model along with design choices. The results indicate that SqueezeNet and ViT components are integral to achieving optimal performance. There is a significant performance difference between the individual



(a) Set of original and heatmap images of CNV and DME



(b) Set of original and heatmap images of Drusen and normal

Fig. 10 XAI analysis of SViT model

components and the hybrid model, demonstrating the effectiveness of the SViT architecture.

Nevertheless, there are some limitations to this research. In addition to showing enhanced performance, the SViT model has not been evaluated with variants of SqueezeNet

Table 5 Comparison with State-of-the-art

Model variant	Mean accuracy	Mean sensitivity	Mean specificity	Mean precision	Mean F1-score
SqueezeNet only	0.9600	0.9510	0.9606	0.9650	0.9580
ViT only	0.9750	0.9670	0.9729	0.9790	0.9730
SViT (Proposed)	**0.9990**	**0.9990**	**1.0000**	**1.0000**	**0.9995**

The bold values refer to the highest performance achieved in the proposed method

and ViT to determine their optimal combination. The optimal number of Transformer Encoder layers for improving SViT performance can be determined by analyzing different Transformer Encoder layers. Nevertheless, these drawbacks can be overcome by performing additional experiments to identify the optimal combination of SqueezeNet and ViT variants and the optimal number of Transformer Encoder layers to build an optimal model without sacrificing performance. Furthermore, the proposed eye disorder detection model would be improved by applying DeepLab3 + [28] based local feature extraction and Few-Shot learning [29]-based classification.

There is room for optimizations and adaptations for a broader range of applications and modalities, including Optical Coherence Tomography Angiography (OCTA), fundus photography, and Adaptive Optics Scanning Laser Ophthalmoscopy (AOSLO). In general, the SViT model is a promising tool for augmenting patient care and clinical decision-making to diagnose and manage a variety of eye disorders. It combines improved performance metrics, robustness, generalizability, and explainability.

6 Conclusion

The purpose of this study is to propose SViT, a hybrid SqueezeNet-ViT (SviT) model for the accurate classification of retinal OCT images, targeting conditions such as CNV, DME, Drusen, and normal cases. In comparison to other state-of-the-art models, this model combines the strengths of both SqueezeNet and ViT, leading to significant improvements in performance metrics, robustness, generalizability, and explainability. The classification accuracy of SViT is 99.70% for binary classifications and 99.90% for multiclass classifications. Additionally, the proposed network requires fewer hyperparameters and requires less inference time than existing models. In this study, the SViT model was shown to be an effective tool to enhance patient care and clinical decision-making for the diagnosis and management of various types of eye disorders. Moreover, SViT can serve as a baseline for future research, providing a solid foundation to explore potential adaptations to different medical imaging modalities and ophthalmology subspecialties, such as anterior segment imaging, ocular oncology, and orbital disorders. This model can play a significant role in improving patient outcomes and care by extending its application to ophthalmic imaging and diagnosis in a broader sense.

Data availability The OCT2017 dataset is available as open-source data and can be accessed from https://www.kaggle.com/paultimothymooney/kermany2018.

Declarations

Conflict of interest The authors have no relevant financial or non-financial interests to disclose.

References

1. Sakata LM, DeLeon-Ortega J, Sakata V, Girkin CA (2009) Optical coherence tomography of the retina and optic nerve–a review. Clin Exp Ophthalmol 37(1):90–99
2. Hui VWK, Szeto SKH, Tang F et al (2022) Optical coherence tomography classification systems for diabetic macular edema and their associations with visual outcome and treatment responses–an updated review. Asia Pac J Ophthalmol (Phila) 11(3):247–257. https://doi.org/10.1097/APO.0000000000000468
3. Krishna KVSSR, Chaitanya K, Subhashini PPS, Yamparala R, Kanumalli SS (2021) Classification of glaucoma optical coherence tomography (OCT) images based on blood vessel identification using CNN and firefly optimization. Traitement du Signal 38(1):239–245
4. Tsuji T, Hirose Y, Fujimori K et al (2020) Classification of optical coherence tomography images using a capsule network. BMC Ophthalmol 20:114. https://doi.org/10.1186/s12886-020-01382-4
5. Stanojevic M, Draškovic D, Nikolic B (2022) Retinal disease classification based on optical coherence tomography images using convolutional neural networks. J Electron Imag 32(3):032004. https://doi.org/10.1117/1.JEI.32.3.032004
6. Srinivasan PP, Kim LA, Mettu PS, Cousins SW, Comer GM, Izatt JA, Farsiu S (2014) Fully automated detection of diabetic macular edema and dry age-related macular degeneration from optical coherence tomography images. Biomed Opt Express 5:3568–3577
7. Chen X, Xue Y, Wu X, Zhong Y, Rao H, Luo H, Weng Z (2023) Deep learning-based system for disease screening and pathologic region detection from optical coherence tomography images. Transl Vis Sci Technol 12(1):29. https://doi.org/10.1167/tvst.12.1.29

8. Omid NM, Hamid H, Hossein K, Shahriar BS, Ahmad A (2023) MedViT: a robust vision transformer for generalized medical image classification. Comput Biol Med 157:106791. https://doi.org/10.1016/j.compbiomed.2023.106791
9. Varadarajan AV, Bavishi P, Ruamviboonsuk P et al (2020) Predicting optical coherence tomography-derived diabetic macular edema grades from fundus photographs using deep learning. Nat Commun 11:130
10. Murugappan M, Bourisly AK, Prakash NB et al (2023) Automated semantic lung segmentation in chest CT images using deep neural network. Neural Comput Appl 35:15343–15364. https://doi.org/10.1007/s00521-023-08407-1
11. Murugappan M, Prakash NB, Jeya R, Mohanarathinam A, Hemalakshmi GR, Mahmud M (2022) A novel few-shot classification framework for diabetic retinopathy detection and grading. Measurement 200:111485. https://doi.org/10.1016/j.measurement.2022.111485
12. Perdomo O, Rios H, Rodríguez FJ, Otálora S, Meriaudeau F, Müller H, González FA (2019) Classification of diabetes-related retinal diseases using a deep learning approach in optical coherence tomography. Comput Methods Programs Biomed 178:181–189
13. Ryu G, Lee K, Park D, Park SH, Sagong M (2021) A deep learning model for identifying diabetic retinopathy using optical coherence tomography angiography. Sci Rep 11(1):23024
14. Das V, Prabhakararao E, Dandapat S, Bora PK (2020) B-Scan attentive CNN for the classification of retinal optical coherence tomography volumes. IEEE Signal Process Lett 27:1025–1029
15. Wu B, Xu C, Dai X, Wan A, Zhang P, Yan Z, Tomizuka M, Gonzalez J, Keutzer K, Vajda P (2020) Visual transformers: token-based image representation and processing for computer vision. arXiv preprint arXiv:2006.03677
16. Ma Z, Xie Q, Xie P, Fan F, Gao X, Zhu J (2022) HCTNet: a hybrid ConvNet-transformer network for retinal optical coherence tomography image classification. Biosensors 12:542. https://doi.org/10.3390/bios12070542
17. Zhang Y, Li Z, Nan N, Wang X (2023) TranSegNet: hybrid CNN-vision transformers encoder for retina segmentation of optical coherence tomography. Life 13:976. https://doi.org/10.3390/life13040976
18. Jiang Z, Wang L, Wu Q, Shao Y, Shen M, Jiang W, Dai C (2022) Computer-aided diagnosis of retinopathy based on vision transformer. J Innov Opt Health Sci 15(02):2250009. https://doi.org/10.1142/S1793545822500092
19. Dutta P, Sathi KA, Hossain MA, Dewan MAA (2023) Conv-ViT: a convolution and vision transformer-based hybrid feature extraction method for retinal disease detection. J Imag 2023(9):140. https://doi.org/10.3390/jimaging9070140
20. Strudel R, Garcia R, Laptev I, Schmid C (2021) Segmenter: transformer for semantic segmentation. In: Proceedings of the IEEE/CVF international conference on computer vision, pp 7262–7272
21. Dai Y, Gao Y, Liu F (2021) Transmed: transformers advance multi-modal medical image classification. Diagnostics 11(8):1384
22. Gao Y, Zhou M, Metaxas DN (2021) UTNet: a hybrid transformer architecture for medical image segmentation. In: Medical image computing and computer assisted intervention–MICCAI 2021: 24th international conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part III 24, Springer International Publishing, pp 61–71
23. He K, Gan C, Li Z, Rekik I, Yin Z, Ji W, Gao Y, Wang Q, Zhang J, Shen D (2022) Transformers in medical image analysis: a review. Intell Med 3(1):59–78. https://doi.org/10.1016/j.imed.2022.07.002
24. Khan A, Rauf Z, Sohail A, Rehman A, Asif H, Asif A, Farooq U (2023) A survey of the vision transformers and its CNN-transformer based variants. arXiv preprint arXiv:2305.09880
25. Nanni L, Loreggia A, Barcellona L, Ghidoni S (2023) Building ensemble of deep networks: convolutional networks and transformers. IEEE Access 11:124962–124974. https://doi.org/10.1109/ACCESS.2023.3330442
26. Kermany D, Zhang K, Goldbaum M (2018) Labeled optical coherence tomography (oct) and chest x-ray images for classification. Mendeley Data 2(2):651
27. Kuwayama S, Ayatsuka Y, Yanagisono D, Uta T, Usui H, Kato A, Takase N, Ogura Y, Yasukawa T (2019) Automated detection of macular diseases by optical coherence tomography and artificial intelligence machine learning of optical coherence tomography images. J Ophthalmol 2019:6319581
28. Islam MM, Yang HC, Poly TN, Jian WS, Li YCJ (2020) Deep learning algorithms for detection of diabetic retinopathy in retinal fundus photographs: a systematic review and meta-analysis. Comput Methods Programs Biomed 191:105320
29. Li T, Gao Y, Wang K, Guo S, Liu H, Kang H (2019) Diagnostic assessment of deep learning algorithms for diabetic retinopathy screening. Inf Sci 501:511–522
30. Das V, Dandapat S, Bora PK (2019) Multi-scale deep feature fusion for automated classification of macular pathologies from OCT images. Biomed Signal Process Control 54:101605
31. Huang L, He X, Fang L, Rabbani H, Chen X (2019) Automatic classification of retinal optical coherence tomography images with layer guided convolutional neural network. IEEE Signal Process Lett 26(7):1026–1030
32. Cazañas-Gordón A, Parra-Mora E, Cruz LADS (2021) Ensemble learning approach to retinal thickness assessment in optical coherence tomography. IEEE Access 9:67349–67363
33. Anoop BN, Pavan R, Girish GN, Kothari AR, Rajan J (2020) Stack generalized deep ensemble learning for retinal layer segmentation in optical coherence tomography images. Biocybern Biomed Eng 40(4):1343–1358
34. Ai Z, Huang X, Feng J, Wang H, Tao Y, Zeng F, Lu Y (2022) FN-OCT: disease detection algorithm for retinal optical coherence tomography based on a fusion network. Front Neuroinform 16:876927. https://doi.org/10.3389/fninf.2022.876927
35. Khan A, Sohail A, Zahoora U, Qureshi AS (2020) A survey of the recent architectures of deep convolutional neural networks. Artif Intell Rev 53:5455–5516
36. Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai X, Unterthiner T, Dehghani M, Minderer M, Heigold G, Gelly S, Uszkoreit J (2020) An image is worth 16x16 words: transformers for image recognition at scale. arXiv preprintarXiv:2010.11929
37. Wassel M, Hamdi AM, Adly N, Torki M (2022) Vision transformers based classification for glaucomatous eye condition. In: 2022 26th international conference on pattern recognition (ICPR), IEEE, pp 5082–5088
38. Fan R, Alipour K, Bowd C, Christopher M, Brye N, Proudfoot JA, Goldbaum MH, Belghith A, Girkin CA, Fazio MA, Liebmann JM, Weinreb RN, Pazzani M, Kriegman D, Zangwill LM (2023) Detecting glaucoma from fundus photographs using deep learning without convolutions: transformer for improved generalization. Ophthalmol Sci 3(1):100233
39. Wen H, Zhao J, Xiang S, Lin L, Liu C, Wang T, An L, Liang L, Huang B (2022) Towards more efficient ophthalmic disease classification and lesion location via convolution transformer. Comput Methods Progr Biomed 220:106832
40. He J, Wang J, Han Z, Ma J, Wang C, Qi M (2023) An interpretable transformer network for the retinal disease classification using optical coherence tomography. Sci Rep 13(1):3637
41. Liu Z, Lin Y, Cao Y, Hu H, Wei Y, Zhang Z, Lin S, Guo B (2021) Swin transformer: hierarchical vision transformer using

shifted windows. In: Proceedings of the IEEE/CVF international conference on computer vision, pp 10012–10022
42. Retinal OCT Images (optical coherence tomography) | Kaggle. https://www.kaggle.com/paultimothymooney/kermany2018. Retrieved on 2 June 2023
43. Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D, Batra D (2017) Grad-cam: visual explanations from deep networks via gradient-based localization. In: Proceedings of the IEEE international conference on computer vision, pp 618–626