mathematics

MDPI

Article

OCTNet: A Modified Multi-Scale Attention Feature Fusion Network with InceptionV3 for Retinal OCT Image Classification

Irshad Khalil [1], Asif Mehmood [1], Hyunchul Kim [2,*] and Jungsuk Kim [1,*]

[1] Department of Biomedical Engineering, College of IT Convergence, Gachon University, 1342 Seongnamdaero, Sujeong-gu, Seongnam-si 13120, Republic of Korea; irshadkhalil@gachon.ac.kr (I.K.); asif@gachon.ac.kr (A.M.)
[2] School of Information, University of California, 102 South Hall 4600, Berkeley, CA 94720, USA
* Correspondence: hyunchul@berkeley.edu (H.K.); jungsuk@gachon.ac.kr (J.K.)

Abstract: Classification and identification of eye diseases using Optical Coherence Tomography (OCT) has been a challenging task and a trending research area in recent years. Accurate classification and detection of different diseases are crucial for effective care management and improving vision outcomes. Current detection methods fall into two main categories: traditional methods and deep learning-based approaches. Traditional approaches rely on machine learning for feature extraction, while deep learning methods utilize data-driven classification model training. In recent years, Deep Learning (DL) and Machine Learning (ML) algorithms have become essential tools, particularly in medical image classification, and are widely used to classify and identify various diseases. However, due to the high spatial similarities in OCT images, accurate classification remains a challenging task. In this paper, we introduce a novel model called "OCTNet" that integrates a deep learning model combining InceptionV3 with a modified multi-scale attention-based spatial attention block to enhance model performance. OCTNet employs an InceptionV3 backbone with a fusion of dual attention modules to construct the proposed architecture. The InceptionV3 model generates rich features from images, capturing both local and global aspects, which are then enhanced by utilizing the modified multi-scale spatial attention block, resulting in a significantly improved feature map. To evaluate the model's performance, we utilized two state-of-the-art (SOTA) datasets that include images of normal cases, Choroidal Neovascularization (CNV), Drusen, and Diabetic Macular Edema (DME). Through experimentation and simulation, the proposed OCTNet improves the classification accuracy of the InceptionV3 model by 1.3%, yielding higher accuracy than other SOTA models. We also performed an ablation study to demonstrate the effectiveness of the proposed method. The model achieved an overall average accuracy of 99.50% and 99.65% with two different OCT datasets.

Keywords: OCTNet; retinal OCT; Choroidal Neovascularization; diabetic macular edema; Drusen; retinal disease classification; feature fusion; multi-scale attention mechanism

MSC: 68T01; 68T10; 68T07

check for updates

Citation: Khalil, I.; Mehmood, A.; Kim, H.; Kim, J. OCTNet: A Modified Multi-Scale Attention Feature Fusion Network with InceptionV3 for Retinal OCT Image Classification. *Mathematics* **2024**, *12*, 3003. https://doi.org/10.3390/math12193003

Academic Editor: Rami Qahwaji

Received: 20 August 2024
Revised: 17 September 2024
Accepted: 24 September 2024
Published: 26 September 2024

Correction Statement: This article has been republished with a minor change. The change does not affect the scientific content of the article and further details are available within the backmatter of the website version of this article.

Copyright: © 2024 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

The retina is a transparent and soft membrane located adjacent to the inner surface of the choroid. It is responsible for sensing incoming light stimuli and plays an important role in vision [1]. The retina serves as the main foundation for the visual system, where the brain processes electrical transmissions as coherent visual representations [2]. In recent years, the number of patients with different retinal diseases has increased [3]. According to the World Health Organization, the number of patients suffering from vision impairments and related problems in 2019 was around 2.2 billion [4]. Around 80% of these diseases can cause permanent blindness if they are not detected at their initial stages [5], which is a major issue for the general health and quality of life of the general public. This is an increased

number of patients with different retinal diseases, which can cause permanent damage to vision and, in severe cases, result in permanent blindness. Among all diseases, Age-related Macular Degeneration (AMD) (2.7 million) and Diabetic Macular Edema (DME) (3 million) a dominant retina diseases with a high number of patients with an estimated 360 million by 2030 [6]. AMD is further divided into dry and wet depending on its pathological changes. In dry AMD, yellowish-white, rounded vitreous warts of different shapes and sizes can be visualized in the posterior pole of the eye (Drusen) while in Wet AMD induces the outward growth of new blood vessels in the Choroidal Capillaries (CNV) [7]. All these symptoms within the retina and pathological features play an important role in diagnosing different retina abnormalities.

To effectively analyze and study conditions of the eyes with various abnormalities, ophthalmologists use two different types of imaging techniques: Optical Coherence Tomography (OCT) and color fundus imaging. Color fundus imaging captures 2D images of the retina. A fundus camera is used to capture the reflection of light from the retina, showing its deformation, but it is unable to provide depth information. Thus, fundus imaging only offers a 2D overview without showing the in-depth structure. OCT on the other hand, is a non-invasive imaging modality that generates cross-sectional images of human retinal abnormalities within the retina [8]. It utilizes light waves to capture different cross-sections in the human retina and captures high-resolution, which is used for analyzing and visualizing eye tissue [9]. OCT Imaging provides a mechanism for detecting different anomalies in the eyes even at their early stages. The benefits of OCT imaging over other modalities are evident in its real-time imaging capability and ability to generate three-dimensional (3D) visuals without patient dilation. OCT imaging helps in the early identification and detection of different retinal diseases by showing their underlying factors [10]. The remarkable capabilities of OCT imaging have revolutionized the field of ophthalmology, enabling more accurate diagnoses are possible for individuals with retinal conditions [11].

The most critical phase of retinal image analysis in medical applications involves the classification of OCT images into different types. Accurately classifying OCT images is one of the key responsibilities of medical practitioners; however, manual interpretation of these OCT images is time-intensive and prone to various errors. Recent advancements in developing automated DL model for OCT retinal image classification has become a trending research area, and different models have been developed for the classification of OCT images [12,13]. Advances in the DL methods have resulted in improved performance comparable to human observers in multiple domains, e.g., retinal image diagnosis [4,14]. Recent advancements in DL and AI have led to better achievements across the field of detection and identification in the different medical domains [15], and are used for classification of retinal and medical images [16–19]. The excellent performance of the Convolutional Neural Network (CNN)-based model inspired the researchers to investigate and explore its performance in medical domains [20–22].

As DL models evolve, the enhanced learning capabilities of neural network architectures are increasingly replacing conventional OCT image analysis techniques. These models learned several features, some of these are vital and others are irrelevant as these models are mainly based on convolution and pooling layers for feature extraction [23]. Recent research utilized the CNN framework for OCT image analysis including retinal layer segmentation [14] and classification. On the other hand, most CNN-based models not only learn relevant features but, unfortunately, also learn from irrelevant features such as noise [24] and background information. Learning from irrelevant features can significantly reduce the overall performance of the model.

Inspired by human visual perception, which focuses on important regions within a complex scene, many researchers have proposed various plug-and-play modules that have been investigated in different domains [20,25]. Based on this, important regions representing the abnormality within the images are considered more important as compared to other regions. Attention-based mechanisms have been used alongside various deep CNN models, showing great potential in further enhancing CNN performance. These

attention networks suppress redundant information in the feature map and focus more on effective features. These networks have been widely used in different tasks and improve the representational power of the CNN networks by focusing on the dominant features while suppressing redundant and unnecessary one. The attention mechanism extracts better features effect and improve predictive power. Recent research focused on the development of more powerful attention mechanisms [26–29] to improve CNN model performance. The Convolution Block Attention Module (CBAM) proposed by [25] greatly improves the CNNs model performance by capturing the important features in the feature maps. CBAM contains two different types of attention modules: the Channel Attention Module (CAM) and the Spatial Attention Module (SAM), which make the model focus more on the important regions within the image. Multi-scale convolutional operations allow CNN-based models to focus on more relevant and concise features and allows the module to better capture the spatial relationships between features and improve performance of the module [30–32].

Therefore, in this article, we propose a model based on a modified attention mechanism for OCT image classification. We chose InceptionV3 after experimentation in our particular goal of OCT images analysis. After experimentation and ablation studies of other modules, InceptionV3 was found to be better feature extractor in our proposed networks. The inception modules with various kernel sizes provide effective multi scale feature extraction, which is essential for collecting the different characteristics in OCT images, ranging from complex retinal layers to larger structural abnormalities. Multi scale convolution operations at different levels are used to extract multi scale more enhanced features, resulting in more enhance performance. The proposed network improves the classification performance of the InceptionV3 model, achieving high and more robust performance. We call our proposed network "OCTNet", employing both CAM and multi scale SAM. The proposed module improves the discriminative power of the backbone. Using our proposed attention module, backbone features from the pre-trained InceptionV3 are refined resulting in more higher performance. The CAM module focuses on the inter-channel relationships among features the modified SAM module refine inter-spatial feature. This approach improves the discriminative power of the OCTNet module, achieving more robust performance in terms of classification and outperforming the existing approaches. In summary, we made the following five research contributions:

- This paper introduces a novel "OCTNet" model for the classification of OCT images. The proposed model utilizes both convolutional InceptionV3 modules and a multi-scale attention module, enabling it to extract features from the most relevant regions in complex OCT images.
- We introduce the multi-scale attention module utilizing both CAM and multi-scale SAM modules. This proposed attention-based module can be used with any backbone model to achieve robust classification results, thereby increasing the application domains of our research.
- To evaluate the effectiveness of the proposed OCTNet model, we used two state-of-the-art datasets. We also compared the classification performance with some previous studies to assess its strength, and the results show that the OCTNet model performs better than other previous research utilizing the same datasets.
- We improved accuracy for both binary and multi-class classification and achieved more robust results compared to SOTA methods.
- We performed different experiments and ablation studies and empirically proved the effectiveness of the proposed multi-scale model in achieving high classification accuracy.

We organized the remaining sections as follows: Section 2 reviews the most recent research on similar topics using the same datasets. Section 3 details the methodology used in the paper. Section 4 presents the experimental results and empirical findings of the proposed model, with an ablation study in Section 5. Finally, Section 6 contains a comprehensive discussion, followed by the conclusion in Section 7.

2. Related Work

In recent years, the research community has focused extensively on the domain of OCT image classification, which is mainly divided into two main domains, i.e., traditional ML and DL based methods. TML methods mostly rely on semiautomatic feature extraction methods and DL methods are based on fully automatic feature extraction. The main feature extraction techniques include histograms of oriented gradients, local binary patterns, and scale-invariant feature transform. The main classifiers include Support Vector Machine, Random Forest, and Multi layer Perceptron, which efficiently classify OCT images into different classes. Despite its high performance, these methods are based on manual feature extraction, which makes it time-consuming and also unable to perform well in complex tasks. In contrast, DL methods utilize fully automatic feature extraction, offering a more streamlined and potentially more accurate approach to OCT image classification. Recent research focused on developing different strategies for the detection and classification of abnormalities detection in OCT images. In the last few years, OCT image analysis and classification have been one of the trending research directions and several AI/DL models have already been proposed by different researchers for the automated detection and classification of retinal diseases. Among these methods, CSR, AMD, and DR were the most intensively studied diseases [33]. In this section, we will discuss the current state-of-the-art methods for the detection and classification of different diseases utilizing OCT images.

In [34], the authors present an automatic classification and detection method for AMD using OCT images. This method is based on DL for feature extraction and a local outlier factor algorithm. This model is based on the Res Net-50 architecture with Level 2-constrained soft-max loss, which is used to extract useful information and features from OCT images. However, the current research lacks a comparison between the binary classification of AMD and CSR images.

In [35], the authors proposed a method for detecting and segmenting three different retinal diseases, specifically cysts. They employed a deep ensemble-based architecture using a modified version of the UNET architecture, with a predictor block to combine the results of all three models. Despite achieving a classification accuracy of 79%, their model has potential for further improvement and optimization through enhanced ensemble techniques.

In [36], the author proposed multi-scale convolutional and expert-based ensemble models to classify and identify normal eyes and two different types of macular diseases, including AMD and DME. This approach achieved an accuracy of 95.9% using a public dataset of 45 OCT images. In [37], the author developed an enhanced model based on an automated DRNN for classifying and identifying different retinal OCT scans. This method utilized residual connections with cascaded multi-kernel convolutions for optimal training and classification. The model was trained on 108,312 OCT images and tested on 1000 OCT scans, demonstrating relatively high accuracy compared to other proposed methods.

In [38], the authors proposed a method to detect and classify OCT images into three different classes: DME, CNV, and Drusen, as well as normal retina. They developed an enhanced ensemble model called OCTx for this classification task. However, the model was trained and tested on a specific dataset, which may limit its classification performance on real-world datasets.

In [39], the authors proposed to used DL and ensemble learning for the classification and identification of different diseases, achieving high performance while utilizing more state-of-the-art datasets.

In [40], the authors proposed an end-to-end web-based graphical user interface along with an application that uses ML and DL techniques for efficient classification and identification of different eye diseases. The web-based application allows end users or patients to submit their OCT images, which are pre-processed and segmented using a pre-trained UNet custom model. The segmented images are then classified by using an ensemble model, utilizing Inception and XceptionV3 with a self-attention layer which achieve more robust results as compared to other methods.

In [41], the authors used an OCT dataset to classify AMD, DME, and normal images using CNN based model. The authors used pre-trained ImageNet weights, which further improved the classification performance from 68.17% to 92.89%. The proposed model is an ensemble of the three different CNN models with pre-trained weights and parameters, correctly classifies different classes of images to their respective classes.

In [42], the authors developed a DL-based intelligent system for the automatic categorization of OCT images. They utilized a dataset of 25,134 labeled OCT images, which were manually annotated by 17 experts. By employing 101-layer CNN architecture utilizing ResNet architecture and 10-fold cross-validation achieved high-performance results of accuracy of 95.9%. However, the proposed system depends on manual labeling, which makes the current system more complicated for reliable deployment. The Table 1 below summarizes recent approaches for ocular image classification using various methods and datasets.

Table 1. Summary of literature review highlighting different approaches to retinal eye disease analysis.

Dataset	Classes	Images Used	Method	Accuracy	Publication Year
OCT	4	84,495	Adaptive histogram equalization [43]	97.7%	2023
OCT	4	Entire dataset	Modified ResNet (50) [44]	96.93%	2023
OCT	4	Entire dataset	CNN [45]	96.93%	2023
OCT	4	Limited dataset	[46]	93%	2019
OCT	4	5000 for train/test	Modified CNN_VGG19 [47]	97.7%	2023
OCT – C8	4	24,000	VGG19 [48]	97%	2022
OCT –V2	4	V2-dataset	Deep CNN [49]	95.55%	2024
OCT	4	83,484	Alexnet [49]	97.1%	2018
OCT	3	108,312	Ensemble mode [36]	98.8%	2020
OCT	4	Entire dataset	Expert-based ensemble mode (DRRN) [37]	98.8%	2020
OCTX	3	84,484	VGG16, DenseNet, InceptionV3, and a Custom CNN [38]	98.53%	2024
OCT	4	108,309	ResNet152 [39]	98.8%	2020
OCT	4	108,312	InceptionV3 and Xception [40]	96.6%	2024
OCT and OCTID	4 and 5	3213 and entire dataset	deep ensemble convolutional neural network [50]	96.46% and 98.70%	2021
OCT	4	Dataset's Subset	ensemble model based on CNN, Efficientnet_v2 and Resnet [41]	97.89%	2024
OCT (Manully labeled)	4	25,134	CNN [42]	95.9%	2018

Overall, these recent approaches have made remarkable advancements in the field of retinal images classification, but there is still room for further improvements in terms of improving accuracy and robustness with diverse datasets. In this research, we propose a model based on a modified attention mechanism along with InceptionV3, which is chosen after experimentation, especially in our particular goal of OCT images analysis. After experimentation, we found InceptionV3 to be better feature extractor. The inception modules with various kernel sizes provide effective multi scale feature extraction, which is essential for collecting the different characteristics in OCT images, ranging from complex retinal layers to larger structural abnormalities. Multi scale convolution operations at different levels are used to extract multi scale more enhanced features, resulting in more enhance performance. The proposed network improves the classification performance of

the InceptionV3 model, achieving high and more robust performance. Our model improves the discriminative power of the backbone resulting in more higher performance. The CAM module focuses on the inter-channel relationships among features and the modified SAM module refines inter-spatial features. With the high discriminative power of OCTNet, our module results in more robust performance in term of classification and outperforms the existing approaches.

3. Materials and Methods

To enhance predictive performance, we developed a framework that includes InceptionV3 as a backbone feature extractor, followed by the fusion of dual attention modules comprising CAM and multi scale SAM. The input to system is randomly selected retinal OCT from two publicly available datasets. The proposed architecture effectively these images according to their corresponding classes. The proposed framework is shown in Figure 1 below, and explained in the next subsections.

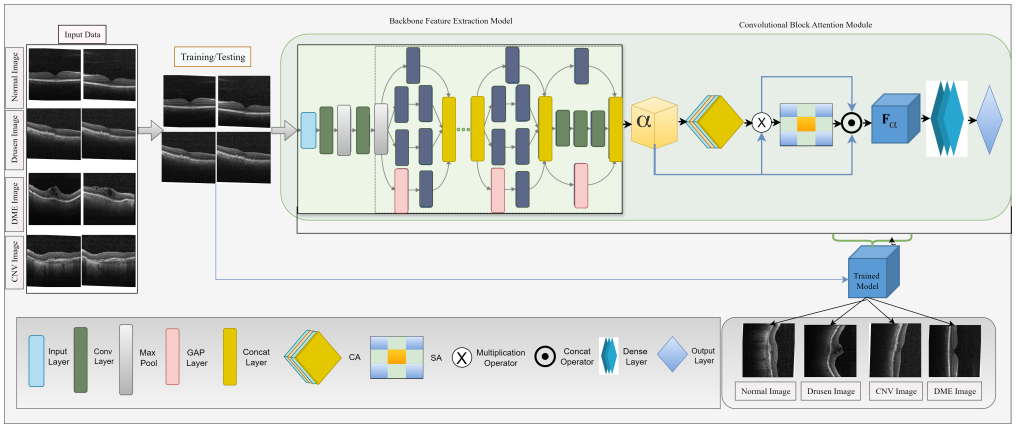


Figure 1. Proposed retinal eyes detection framework.

3.1. Deep Feature Extraction Method

Despite the growing interest in exploring attention mechanisms for image segmentation of natural scenes, there is still limited work for adopting medical image classification, especially OCT images. The existing research focuses on the integration of simple attention mechanisms into medical image classification and segmentation, and hence, there is room for further improvement. The ability of CNNs to extract useful features from input images makes them more suitable for various application domains related to computer vision and related tasks. The choice for selecting CNN architecture is a challenging task in domain-specific computer vision problems. Selecting a domain-specific CNNs architecture is necessary to perform accurate predictions for real-world problems. To solve this issue, much recent research focuses on the development of a problem-specific CNN architecture that performs better in that specific problem. Many researchers used the pre-trained CNN models in the feature extractors and achieved SOTA performance results for each specific problems. In this paper, we fine-tuned a pre-trained model that helps the InceptionV3 to effectively learn feature from the OCT images. The proposed "OCTNet" network is composed of dual attention mechanisms by the fusion of both channel attention and modified spatial attention to obtain more enhanced feature maps.

In recent years, SOTA networks based on attention mechanism achieved superior performance due to the integration of feasible attention models [51]. These mechanisms emphasized the most relevant and important elements of the input image via different learning strategies, and thus, become a very important component in DNN [25]. The integration of attention modules has made great progress in many vision related tasks such as image question-answering [52], video summarization [53] image captioning [54], and classification [30,55].

Our approach is to integrate the modified CA and SA modules within the proposed method, allowing it to concentrate on specific regions within the image to improve the classification performance specifically in OCT images. Specifically, we used different architectures including MobileNet, Xception, ResNet50, NASNetMobile, and InceptionV3 to find the best model as the backbone for the feature extractor. Our main objective is to find the most suitable feature selection mechanism for effective OCT image classification. From experimentation, we concluded that the InceptionV3 architecture performs better for OCT image classification. InceptionV3 includes more Inception modules and has been optimized to obtain better results as compared to their older versions [30,31]. The Inception modules featured in InceptionV3 are characterized by their multi-scale performance classification capability for different vision-related domains and achieve high classification results for different tasks. Within InceptionV3, these modules, denoted as (A), (B), and (C), integrate parallel convolutional and pooling layers, as shown in Figure 2.

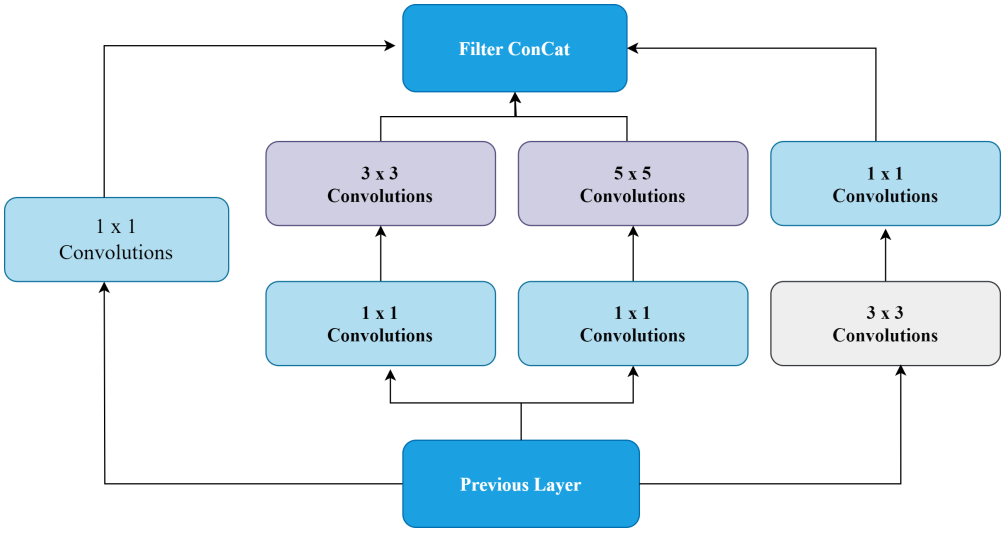


Figure 2. Schematic representation of the InceptionV3 architecture for retinal image classification.

These layers utilize compact filter sizes such as 1×1, 3×3, 1×3, and 3×1 to effectively reduce the number of trainable parameters. InceptionV3 operates on input images with a default size of $299 \times 299 \times 3$. Initial image processing involves the application of five convolutional layers, each employing multiple 3×3 kernels. In the proposed framework, we removed the final dense layers from the InceptionV3 model and obtain a feature vector with 8×8 and 2048 channels represented by α. Mathematically, it can be represented by the following Equation (1):

$$\alpha = \zeta(\zeta(\zeta((\mu(\chi))))) \quad (1)$$

where ζ represents each Inception modules, and μ represents the initial convolutional operation performed on the input χ.

As a result of the operation utilizing Equation (1), we obtained a feature vector presented by α, which contains different information from the images including its structure and other details, etc. To further improve classification accuracy and address misclassification, we incorporated a modified attention module that enhances the extraction of both channel and spatial details. These steps are further explained in the following subsections.

3.2. "Octnet" Dual Retinal Attention Networks

CNN-based architectures integrated with attention mechanisms have been extensively explored across various application domains [3,7,30,56], demonstrating promising performance in different tasks. Moreover, some recent studies have only incorporated CBAM into CNN architectures for diabetic retinopathy [57] without fully exploiting the analysis

of multi scale attention mechanisms. We propose the integration of CA and SAM into our OCTNet architecture, forming the core of the proposed model. By leveraging this dual attention scheme, OCTNet can effectively focus on the most critical regions for retinal image classification. The proposed module, along with the CA and multi scale SA mechanisms, enables the most important part within the retinal images by extracting multi scale features resulting in improved performance. The incorporation of multi-scale SA enhances the model's ability to capture most valuable features at different spatial resolutions, enabling a more comprehensive understanding of abnormal retinal images. Simultaneously, the CA module helps in prioritizing the most informative channels, further refining the feature representation. This approach empowers OCTNet to accurately detect abnormal retinal images with high accuracy compared to standard attention module. Our proposed OCTNet architecture achieves higher performance in retinal image classification by enabling the model to adaptively focus on the most discriminative features, resulting in improved accuracy and robustness for retinal image analysis.

Channel Attention Module

We employed CA to utilize the inter-channel relationship of related features. The feature of each channel in the images is evaluated as a feature map. The proposed module is shown in Figure 3, the CA consists of two different layers including a maximum convolution layers, pooling layer, global average pooling layer, and fully connected layers. During training, the importance of each convolution feature map of each channel does not represent a target object similarly, despite using either using max-pooling or average-pooling as proposed in [58–60] for different task. In this research, we used both the maximum-pooling and average-pooling techniques, which can generate more enhanced feature maps.

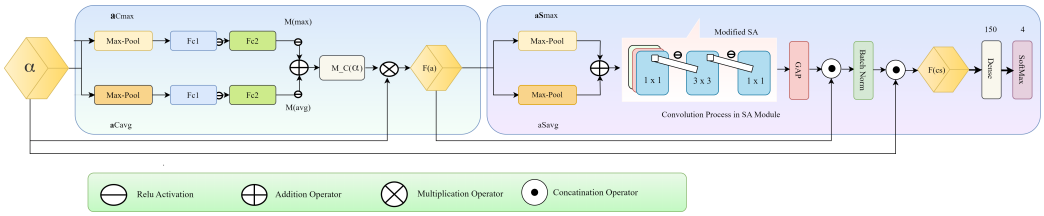


Figure 3. Proposed multi-scale model architecture for our model.

Therefore, these distinct pooling strategies operation were performed separately on the spatial dimension of feature map α, as a result, two different spatial background descriptions: $\alpha Cavg$ and $\alpha Cmax$, mathematically, they are calculated by using the following equation:

$$\alpha_{Cavg} = Avg - Pool(\alpha) \tag{2}$$

$$\alpha_{Cmax} = Max - Pool(\alpha) \tag{3}$$

After these operation, two Fully Connected (FC) layers, fc^1 and fc^2, generate the output of $\alpha Cavg$ and $\alpha Cmax$, which shared common parameters. $ReLU$ activation function is utilized on each FC layer to derive $Mmax$ and $Mavg$. By performing an addition operation on both feature maps, the weight coefficients $Mc(\alpha)$ are obtained as below.

$$M_{max} = \ominus(fc^2(\ominus(fc^1(\alpha_{Cmax})))) \tag{4}$$

$$M_{avg} = \ominus(fc^2(\ominus(fc^1\alpha_{Cavg})))) \tag{5}$$

$$M_c(\alpha) = \&(Mmax \oplus Mavg) \tag{6}$$

The fc^1 and fc^2 layers employed 1×1 convolution operations to expand and compress the features of each channel. Here, the $\ominus$ is ReLU function and $\oplus$ is addition operation.

We then the more concise features $Mc(\alpha)$ by adding $Mmax$ and $Mavg$. To maintain the feature flow a skip connection, $\otimes$ is introduced resulting in the CA map F_c:

$$F_c = M_c(\alpha) \otimes \alpha. \tag{7}$$

3.3. Multi-Scale Spatial Attention Module

We integrated the modified SA module to developed a dual attention networks. Spatial attention maps are generated by analyzing the inter-spatial relationship within the features. The SA module significantly focuses on the dominant elements of the image within spatial areas. This research consists of two different pooling operations, i.e., max pooling and average pooling, and is shown in Figure 1.

The outputs of these pooling layers are then integrated to generate an enhanced feature descriptor. These pooling operations give priority to critical regions within the features map, which results in the enhancement of the model generalization capability.

Integrating max pooling and average pooling operations concurrently offers an efficient means of highlighting essential information embedded within the features. The mathematical expressions for computing max pooling and average pooling are outlined in Equations (8) and (9) correspondingly.

$$\alpha_{Savg} = \&Avg - Pool(F_c), \tag{8}$$

$$\alpha_{Smax} = \&max - Pool(F_c). \tag{9}$$

The obtained feature vectors are then added and combined with an addition operator and the resultant feature maps are passed to the convolutional layer to generate a 2D SA feature map. In this paper, we modified the SA module and utilized three convolutional layers with ReLU activation function. In the first layer, we utilized convolution layer of 1×1 with 64 filters, the second convolution layer uses a 3×3 with 64 layer, and the final layer uses a 1×1 convolution layer.

$$Ms_{F_c} = (\ominus(f^{1\times1}(\ominus(f^{3\times3}(\ominus(f^{1\times1}(\alpha_{Savg} \oplus \alpha_{Smax}))))))), \tag{10}$$

In the above equation, f is the filter size. The SA map can be generated by utilizing global average pooling on the feature maps and then concatenating it with the F_c, as shown in the equation below:

$$Ms_{F_{cGAP}} = GAP(F_c) \tag{11}$$

$$F_s = \odot[Ms_{F_{cGAP}}, F_c]. \tag{12}$$

After the concatenation, batch normalization operation was performed on F_s. Finally, in this method, these feature maps are concatenated from batch normalization and α to generate the F_{CS}:

$$F_{CS} = \odot[Batchnorm, \alpha]. \tag{13}$$

Following the feature extraction technique from the F_{CS}, the resultant features were then passed into a dense layer consisting of 150 neurons, using a soft max layer to assign input images to their appropriate classes.

An integral aspect of our study involves the integration of two attention modules to compute complementary information, with a focus on discerning "what" information holds importance and its spatial distribution. Subsequently, the output of the SA module traverses through three additional fully connected layers, each comprising 64, 32, and 2 neurons, respectively. The utilization of the SoftMax activation function facilitates the classification of input images based on their respective classes, as depicted in Figure 1.

4. Experimentation and Result

This section presents an in-depth description of the experimental setup, datasets utilized, and other parameters. After that, we study both qualitative and quantitative evalua-

tion with the most advanced SOTA approaches. Ultimately, ablation research is undertaken to authenticate the performance and effectiveness of the modified attention modules.

4.1. Experimental Setup

The training setup was designed based on the Keras framework and we utilized the TensorFlow [61] back end. The training was conducted with the following configuration: Stochastic Gradient Descent (SGD) [62] with momentum was used for optimization, with a learning rate (LR) of 0.0001 and a momentum value of 0.9. The model was trained for 50 epochs using a batch size of 24 and the default input size of 299 × 299 × 3 for the Inception model. The experiments were carried out on an Intel® Core™ i9 CPU operating at 3.60 GHz, in conjunction with an NVIDIA GeForce RTX 3070 GPU (NVIDIA Corporation, Santa Clara, CA, USA). In addition to these configurations, extensive experimentation was performed on various parameter values to identify the best-performing setup. Table 2 presents the implementation details and experimental parameters for the development of the proposed model.

Table 2. Specifications of the proposed model implementation.

Implementation Details for the Development of the Proposed System	
Component	**Description**
CPU	Intel Core i9 operating at 3.60 GHz
GPU	NVIDIA GeForce RTX 3070
Framework	Keras with TensorFlow backend
Total Epochs	30
Input size	299 × 299 × 3
Batch Size	16
Optimization Algorithm (OA)	Stochastic Gradient Descent with momentum
Learning Rate	0.0001
Momentum	0.09

4.2. Datasets

In this paper, we used two publicly available datasets OCT [63]. The OCT2017 dataset contains 108,312 OCT images from 4686 patients, including age-related wet Macular degeneration (CNV), age-related dry Macular degeneration (Drusen) and diabetic macular degeneration (Drusen), normal images (NORMAL). The test set comprises 250 images each of CNV, Drusen, DME, and NORMAL. We also used the OCTID [9], which is another publicly available dataset to check the capability of our proposed model on different dataset. It consists of more than 458 for training and 115 images for testing. These images are further divided into 5 different classes. Details and the number of images in each dataset are presented in Table 3. Sample of images from the datasets are depicted in Figures 4 and 5.

Table 3. Number of Images in Each Dataset.

Dataset	**No. of Classes**	**Training Images**	**Testing Images**
Kermany et al. [63]	4	108,312	1000
OCTID [9]	5	458	115

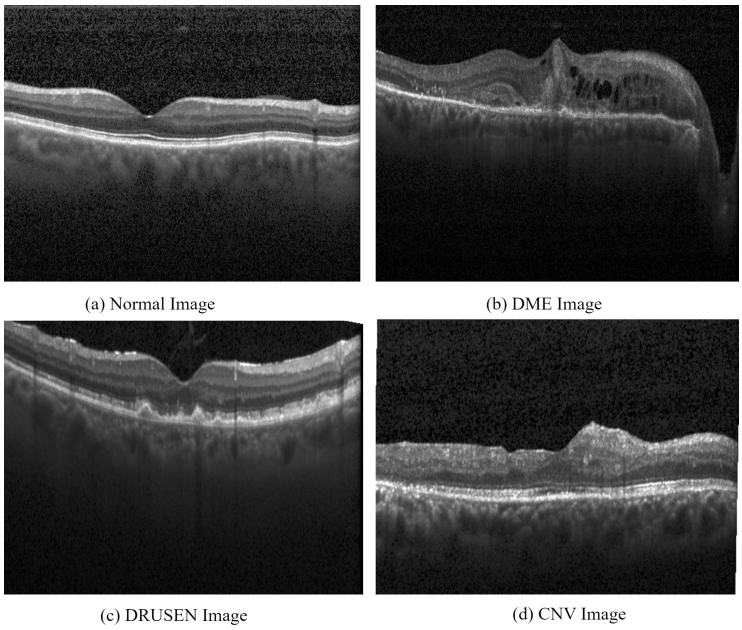


Figure 4. Sample of images from OCT [63]. Sub Figure (**a**) shows an image of a normal category; (**b**) shows an image having DME; (**c**) shows an image of Drusen; and (**d**) shows an image of CNV disease, respectively.

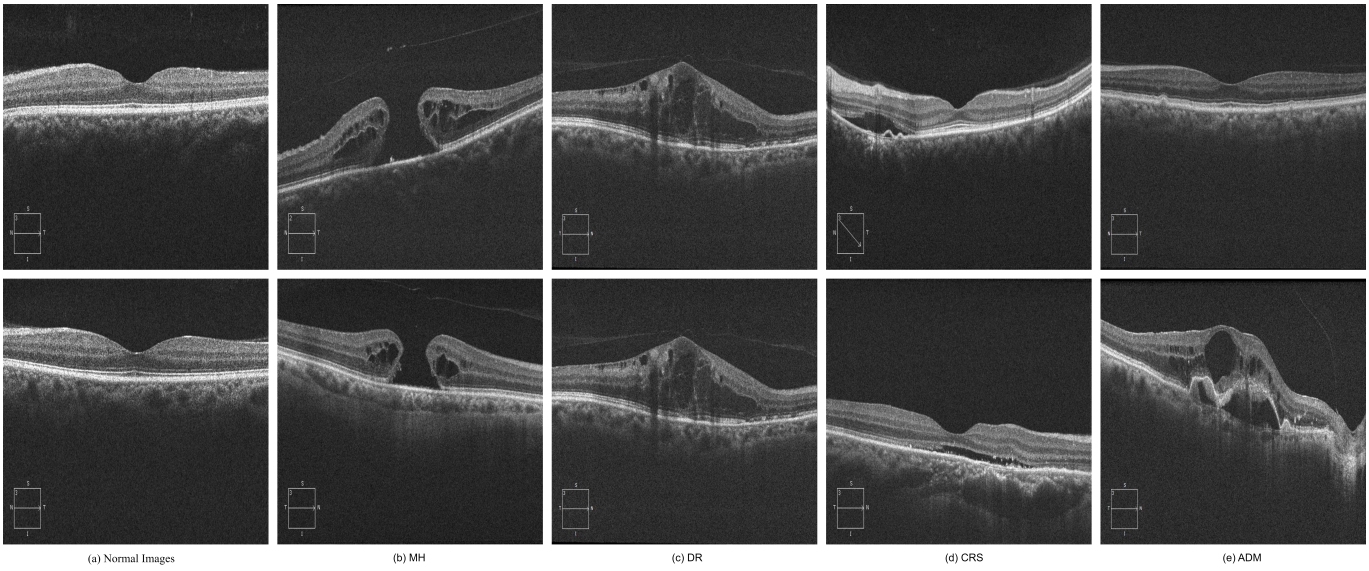


Figure 5. Sample images from the OCTID dataset [9]. Subfigure (**a**) presents a sample of normal images; (**b**) shows a sample from Macular Hole; (**c**) shows an image from Diabetic Retinopathy; (**d**) shows an image of Central Serous Retinopathy; and (**e**) shows a sample image of Age-related Macular Degeneration.

4.3. Statistical Analysis Method

To effectively analyze the robustness and generalization ability of the proposed network architecture, in this paper, we used performance evaluation metrics such as accuracy, precision, recall, and F1 score to compare the performance of different attention mechanisms. The specific definitions for accuracy, precision, recall, and F1 score are provided as follows.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + PN} \tag{14}$$

$$Precision = \frac{TP}{TP + FP} \tag{15}$$

$$Recall/Sensitivity = \frac{TP}{TP + FN} \tag{16}$$

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

where TP represents the number of positive samples that are correctly predicted; TN represents the number of positive samples that are incorrectly predicted as negative; FP represents the number of negative samples that are predicted as positive; FN represents the number of negative samples that are incorrectly predicted as positive. $TP + FP$ represents the total number of samples predicted as positive, while $TP + FN$ represents the actual number of positive samples in the original dataset.

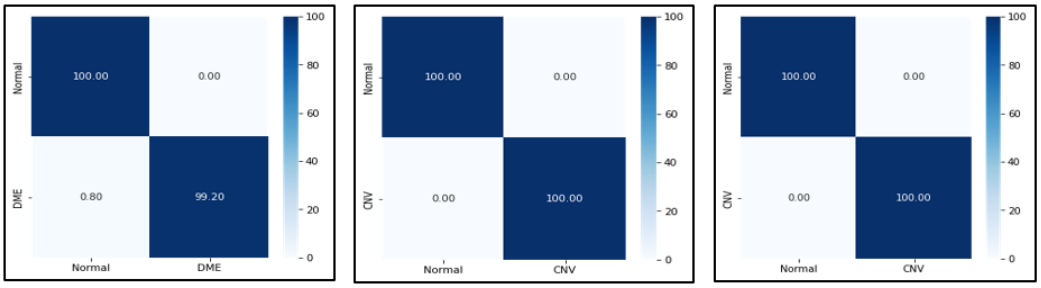
4.4. Quantitative Analysis

In this section, we give the details of experimental results analysis on two different OCT datasets. In this research, the proposed model was evaluated with using both binary and multi-class classification. The classification performance and different performance evaluation matrices values of our finding are listed in Table 4. It can be concluded that the proposed model performs well in terms of binary classification between normal and abnormal classes as well as multi-class classification.

Table 4. Classification performance of the proposed model across different datasets.

Datasets	Accuracy	Precision	Recall	F1-Score
OCTID	99.69	99.87	99.82	99.88
OCT	99.50	99.51	99.50	99.50

In this paper, we only considered the OCT data set [63] for checking the binary classification performance. Figure 6 illustrates the confusion matrices for the binary classifications, respectively. The analysis of these sub figures of Figure 6 of confusion matrices provide deeper information and insights about the model's classification accuracy. In terms of binary classification, very few images are misclassified as incorrect. Overall the proposed model achieved a high average accuracy of 99.5%, a precision of 99.73%, recall of 99.96%, and F1 score of 99.96%, respectively, as shown in Figure 7.



(a) (b) (c)

Figure 6. Confusion matrices: (**a**) Binary vs DME, (**b**) Normal vs. CNV and (**c**) (Binary vs Drusen) respectively.

Figures 8 and 9 illustrate the confusion matrices for the multi class classifications for two datasets, respectively. In terms of multi class classification, the proposed model achieves high accuracy at up to 99.50% and 99.68%, respectively, as shown in Figure 10.

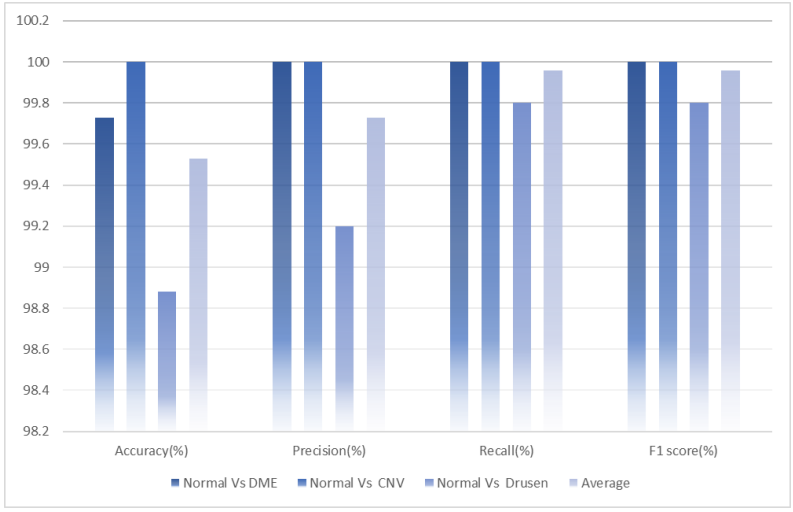


Figure 7. Qualitative analysis of the proposed model for binary classification.

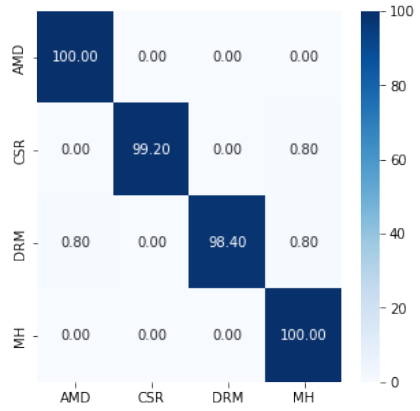


Figure 8. Confusion matrix of the proposed model for multi class classification for OCT dataset.

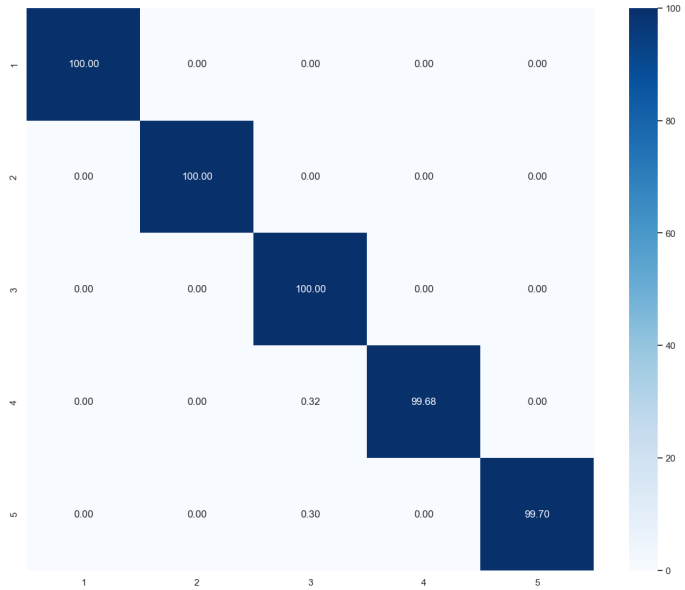


Figure 9. Confusion matrix of the proposed model for multi class classification for OCTID dataset.

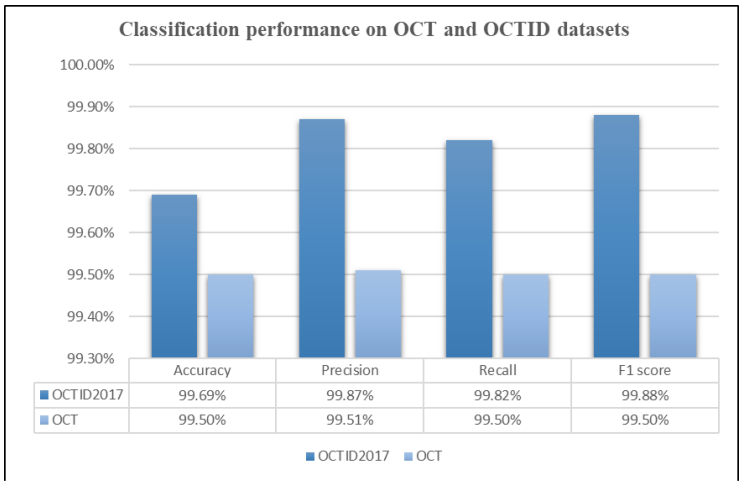


Figure 10. Results of the proposed model for multi class classification.

4.5. Comparative Analysis of the Proposed Method with Other Methods

In this section, we compared the classification performance of our model with other models utilizing the same datasets. Table 5 shows a comprehensive analysis and comparison of different methodologies for retinal image classification using OCT images from the dataset [63]. As shown, our proposed model achieves high performance in terms of accuracy as compared to other models utilizing the same datasets.

Noteworthy techniques include Adaptive Histogram Equalization [43], Modified ResNet [44], and CNN [45], with reported accuracies ranging from 93% to 98.8%, respectively. In addition, ensemble models, such as those presented by [37,39,50], showcase remarkable accuracy rates. Deep learning architectures [49] achieved accuracies ranging from 95.55% to 98.8%, respectively. Most notably, the proposed model, leveraging a Modified Attention-based module, presents promising results, boasting accuracy of 99.50% using [63]. Figure 11 shows an in-depth visual comparison of the proposed method and other techniques. This visual illustration shows the superior performance and robustness of our model in terms of classification accuracy. It highlights the significant advancements made by our approach, especially in handling complex and diverse image datasets.

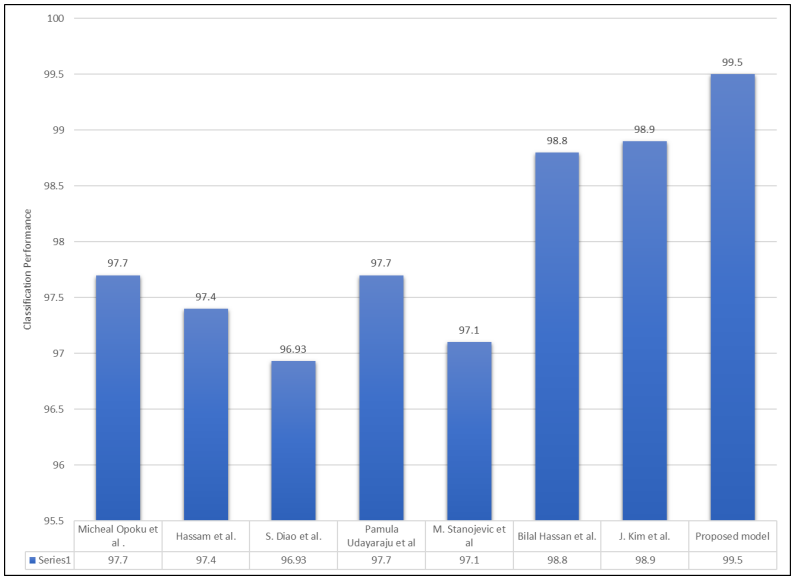


Figure 11. Comparative analysis of the proposed method with other state-of-the-art methods, including those by [37,39,43–45,47,49].

Table 5. Classification comparison of the proposed model with other state-of-the-art models.

Paper	Dataset (Kermany)	Method	Accuracy
Michael Opoku et al. [43]	Yes	Capsule network with contrast limited adaptive histogram equalization	97.7%
Hassan et al. [44]	Yes	Modified ResNet (50) with random forest algorithm	97.47%
S. Diao et al. [45]	Yes	dual guidance CNN	96.93%
Pamula Udayaraju et al. [47]	Yes	hybrid model based on Modified CNN_$VGG19$	97.7%
M. Stanojevic et al. [49]	Yes	Alexnet	97.1%
Bilal Hassan et al. [37]	Yes	Deep Recurrent Residual Inception Network	98.8%
J. Kim et al. [39]	Yes	Ensemble model based ResNet152 and other several CNN models	98.9%
Proposed model	Yes	InceptionV3 + Modified Attention based module	99.50%

5. Ablation Study

In this research, we conducted a series of ablation studies to determine the most effective configuration for our model. Specifically, we experimented with different backbone models in conjunction with Channel Attention (CA) and Spatial Attention (SA) modules to assess their impact on performance. The results of these ablation studies, which compare various deep feature extraction methods, are shown in Table 6.

The idea behind the selection of InceptionV3 is due to its high performance as compared to other DL models and we obtain feature map with more enhanced features. Given the limited robustness of other baseline CNNs, both with and without OCTNet, our focus was on evaluating the performance of OCTNet under the following configurations: (1) InceptionV3 model alone. (2) InceptionV3 fused with CA. (3) InceptionV3 fused with SA. (4) InceptionV3 fused with both CA and SA. (5) InceptionV3 fused with both SA and CA. (6) InceptionV3 fused with SA and CA attentions using (3×3) and (2×2) kernels. (7) InceptionV3 fused with the proposed attention-based method.

The experimental results are presented in Table 6. The standalone InceptionV3 model achieved a classification accuracy of 98.2%. However, integrating different attention modules altered its performance. Specifically, incorporating CA alone reduced accuracy by 0.2%, while incorporating SA alone increased accuracy by 0.4%. Combining both CA and SA modules, without modifying the kernel size, resulted in a 0.6% increase in accuracy. Interestingly, using (3×3) and (2×2) kernels with both attention modules led to a 0.5% decrease in accuracy. Ultimately, fusing the proposed attention-based method with InceptionV3 elevated the classification accuracy to 99.50%, marking a 1.3% improvement.

From this comparative analysis, it is evident that the proposed model achieves competitive accuracy compared to other state-of-the-art methods using the same datasets. The proposed attention-based model consistently outperformed all other ablation models, highlighting its efficiency and effectiveness.

Table 6. Ablation study of InceptionV3 with different fusion strategies.

Backbone Model	Accuracy	Difference	Precision	Recall	F1
InceptionV3	98.2%	N/A	98.23	98.20	98.15
InceptionV3 + CA	98.0	−0.2	98.28	98.6	98.88
InceptionV3 + SA	98.60	0.4	98.62	98.60	98.62
InceptionV3 + (SA + CA) (7×7)	98.80	0.6	98.81	98.81	98.80
InceptionV3 + (CA + SA) (7×7)	98.20	0	98.23	98.20	98.15
InceptionV3 + CBAM [(3×3), (2×2)]	97.70	−0.5	97.71	97.70	97.70
Proposed method	99.50	1.3	99.51	99.5	99.50

Figure 12 shows the visual difference by applying various fusion strategies alongside InceptionV3. The Figure demonstrates that our model significantly improves performance compared to other fusion methods. This visual comparison underscores the effectiveness of integrating the proposed attention mechanisms, highlighting the substantial improvements in classification accuracy and robustness achieved by our approach.

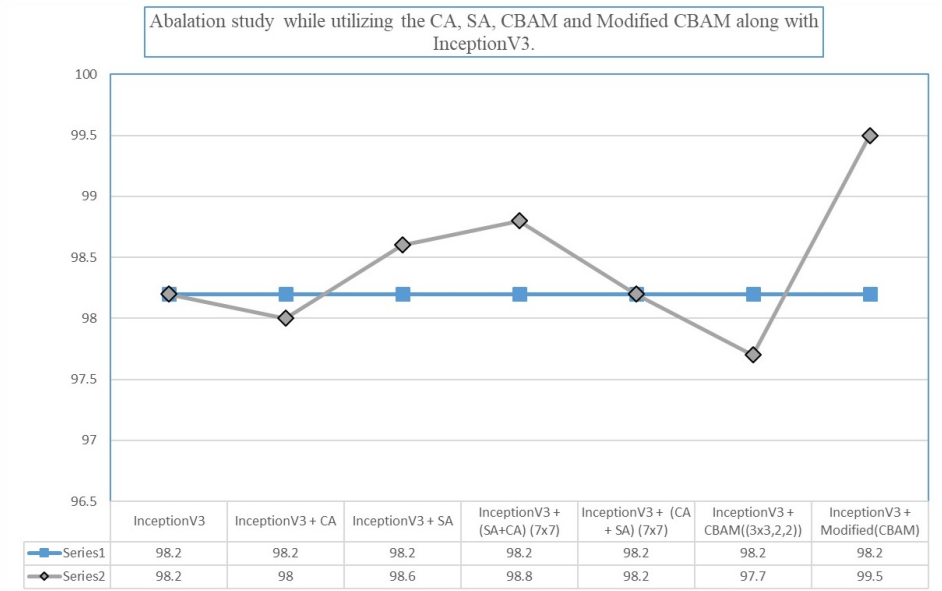


Figure 12. Visual difference of applying different fusion strategies along with the InceptionV3 backbone model.

6. Discussion

This section discusses the technical aspects and results of the proposed method for the detection and classification of different retinal disease using OCT images. In this research, two datasets of retinal diseases were used, consisting of four and five classes of images, respectively. An InceptionV3 was employed as a backbone for deep feature extraction with pre-trained weights. The extracted feature vectors were then passed to the dual attention mechanism to further optimized the extracted feature vectors. The proposed method achieved a classification accuracy of 99.51, and 99.55%, which is the highest among all studied DL models.

Notably, the InceptionV3 model outperformed other DL models, which achieved classification accuracy of 98.2%. The results obtained from conducting various experiments on both models indicate that InceptionV3 along with the proposed attention module improves overall accuracy by +1.3%, which is higher than other models. The proposed model for retinal eye disease classification method performed better than most recent models in the literature, as shown by the comparison in Table 6. Table 6 demonstrates the superior performance of our method compared to current models. Based on this discussion, it can be concluded that the proposed model can be used as an effective tools for the classification of retinal diseases.

7. Conclusions

Accurate classification and detection of types of diseases is essential for early diagnosis and prevention, and it is still a challenging task in image processing. This study proposed a novel method for the classification of retinal diseases using OCT images. The proposed model was tested using two datasets of retinal OCT images, consisting of four and five classes, respectively. The proposed method comprises two main steps. In the first step, deep features were extracted using InceptionV3. In the second step, the proposed dual attention module further improved the extracted features. The enhanced feature map was thereafter used for the classification of retinal eye diseases. The classification performance of the proposed method showed that the proposed model achieved the best classification performance, with overall accuracy of 99.50% and 99.65%, respectively. The potential impact of our method is highly significant, as it can aid in the early and accurate diagnosis of retinal eye diseases. A comparative analysis with existing approaches was also conducted and analyzed, and our results demonstrated that the proposed method outperformed other

methods with the highest accuracy. Thus, it can be concluded that the proposed method is reliable and can accurately classify retinal eye diseases.

Author Contributions: Conceptualization, I.K.; methodology, I.K. and H.K.; software, I.K.; validation, I.K. and H.K.; formal analysis, A.M. and H.K.; investigation, I.K. and H.K.; resources, J.K.; data curation, I.K.; writing—original draft preparation, I.K.; visualization, I.K.; project administration, H.K., J.K. and H.K.; funding acquisition, J.K. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by National Research Foundation of Korea grant (RS-2024-00419269) and (RS-2024-00449882).

Data Availability Statement: The data presented in this study are openly available in OCT and OCTID at https://doi.org/10.1016/j.cell.2018.02.010 and https://doi.org/10.1016/j.compeleceng.2019.106532.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Zrenner, E. Will retinal implants restore vision? *Science* **2002**, *295*, 1022–1025. [CrossRef] [PubMed]
2. Bhandari, M.; Shahi, T.B.; Neupane, A. Evaluating Retinal Disease Diagnosis with an Interpretable Lightweight CNN Model Resistant to Adversarial Attacks. *J. Imaging* **2023**, *9*, 219. [CrossRef] [PubMed]
3. Huang, X.; Ai, Z.; Wang, H.; She, C.; Feng, J.; Wei, Q.; Hao, B.; Tao, Y.; Lu, Y.; Zeng, F. GABNet: Global attention block for retinal OCT disease classification. *Front. Neurosci.* **2023**, *17*, 1143422. [CrossRef] [PubMed]
4. De Fauw, J.; Ledsam, J.R.; Romera-Paredes, B.; Nikolov, S.; Tomasev, N.; Blackwell, S.; Askham, H.; Glorot, X.; O'Donoghue, B.; Visentin, D.; et al. Clinically applicable deep learning for diagnosis and referral in retinal disease. *Nat. Med.* **2018**, *24*, 1342–1350. [CrossRef]
5. dos Santos Samagaio, F.G. Automatic Macular Edema Identification and Characterization Using OCT Images. *Comput. Methods Programs Biomed.* **2018**, *163*, 47–63. [CrossRef]
6. Wu, L.; Fernandez-Loaiza, P.; Sauma, J.; Hernandez-Bogantes, E.; Masis, M. Classification of diabetic retinopathy and diabetic macular edema. *World J. Diabetes* **2013**, *4*, 290. [CrossRef]
7. Das, V.; Dandapat, S.; Bora, P.K. Automated classification of retinal OCT images using a deep multi-scale fusion CNN. *IEEE Sens. J.* **2021**, *21*, 23256–23265. [CrossRef]
8. Hemalakshmi, G.; Murugappan, M.; Sikkandar, M.Y.; Begum, S.S.; Prakash, N. Automated retinal disease classification using hybrid transformer model (SViT) using optical coherence tomography images. *Neural Comput. Appl.* **2024**, *36*, 9171–9188 [CrossRef]
9. Gholami, P.; Roy, P.; Parthasarathy, M.K.; Lakshminarayanan, V. OCTID: Optical coherence tomography image database. *Comput. Electr. Eng.* **2020**, *81*, 106532. [CrossRef]
10. Hui, V.W.; Szeto, S.K.; Tang, F.; Yang, D.; Chen, H.; Lai, T.Y.; Rong, A.; Zhang, S.; Zhao, P.; Ruamviboonsuk, P.; et al. 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.* **2022**, *11*, 247–257. [CrossRef]
11. Darooei, R.; Nazari, M.; Kafieh, R.; Rabbani, H. Optimal deep learning architecture for automated segmentation of cysts in OCT images using X-let transforms. *Diagnostics* **2023**, *13*, 1994. [CrossRef] [PubMed]
12. Tsuji, T.; Hirose, Y.; Fujimori, K.; Hirose, T.; Oyama, A.; Saikawa, Y.; Mimura, T.; Shiraishi, K.; Kobayashi, T.; Mizota, A.; et al. Classification of optical coherence tomography images using a capsule network. *BMC Ophthalmol.* **2020**, *20*, 114. [CrossRef]
13. Srinivasan, P.P.; Kim, L.A.; Mettu, P.S.; Cousins, S.W.; Comer, G.M.; Izatt, J.A.; Farsiu, S. Fully automated detection of diabetic macular edema and dry age-related macular degeneration from optical coherence tomography images. *Biomed. Opt. Express* **2014**, *5*, 3568–3577. [CrossRef] [PubMed]
14. Sengupta, S.; Singh, A.; Leopold, H.A.; Gulati, T.; Lakshminarayanan, V. Ophthalmic diagnosis using deep learning with fundus images–A critical review. *Artif. Intell. Med.* **2020**, *102*, 101758. [CrossRef] [PubMed]
15. Iqbal, S.; Qureshi, A.N.; Li, J.; Mahmood, T. On the analyses of medical images using traditional machine learning techniques and convolutional neural networks. *Arch. Comput. Methods Eng.* **2023**, *30*, 3173–3233. [CrossRef]
16. Qureshi, I.; Yan, J.; Abbas, Q.; Shaheed, K.; Riaz, A.B.; Wahid, A.; Khan, M.W.J.; Szczuko, P. Medical image segmentation using deep semantic-based methods: A review of techniques, applications and emerging trends. *Inf. Fusion* **2023**, *90*, 316–352. [CrossRef]
17. Yadav, S.S.; Jadhav, S.M. Deep convolutional neural network based medical image classification for disease diagnosis. *J. Big Data* **2019**, *6*, 113. [CrossRef]
18. Kumar, R.; Kumbharkar, P.; Vanam, S.; Sharma, S. Medical images classification using deep learning: A survey. *Multimed. Tools Appl.* **2023**, *83*, 19683–19728. [CrossRef]

19. Tan, Z.; Zhang, Q.; Lan, G.; Xu, J.; Ou, C.; An, L.; Qin, J.; Huang, Y. OCT Retinopathy Classification via a Semi-Supervised Pseudo-Label Sub-Domain Adaptation and Fine-Tuning Method. *Mathematics* **2024**, *12*, 347. [CrossRef]
20. 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, Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
21. Huang, G.; Liu, Z.; Van Der Maaten, L.; Weinberger, K.Q. Densely connected convolutional networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 4700–4708.
22. Szegedy, C.; Vanhoucke, V.; Ioffe, S.; Shlens, J.; Wojna, Z. Rethinking the inception architecture for computer vision. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 2818–2826.
23. Karthik, M.; Thangavel, K.; Sasirekha, K. Novel deep CNN model based breast cancer classification. In Proceedings of the 2023 7th International Conference on Computing Methodologies and Communication (ICCMC), Erode, India, 23–25 February 2023; IEEE: Piscataway, NJ, USA, 2023; pp. 524–529.
24. Ferentinos, K.P. Deep learning models for plant disease detection and diagnosis. *Comput. Electron. Agric.* **2018**, *145*, 311–318. [CrossRef]
25. Woo, S.; Park, J.; Lee, J.Y.; Kweon, I.S. Cbam: Convolutional block attention module. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 3–19.
26. Hu, J.; Shen, L.; Sun, G. Squeeze-and-excitation networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 7132–7141.
27. Wang, Q.; Wu, B.; Zhu, P.; Li, P.; Zuo, W.; Hu, Q. ECA-Net: Efficient channel attention for deep convolutional neural networks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 14–19 June 2020; pp. 11534–11542.
28. Jaderberg, M.; Simonyan, K.; Zisserman, A. Spatial transformer networks. *Adv. Neural Inf. Process. Syst.* **2015**, *28*.
29. Hu, J.; Shen, L.; Albanie, S.; Sun, G.; Vedaldi, A. Gather-excite: Exploiting feature context in convolutional neural networks. *Adv. Neural Inf. Process. Syst.* **2018**, *31*.
30. Yar, H.; Hussain, T.; Agarwal, M.; Khan, Z.A.; Gupta, S.K.; Baik, S.W. Optimized dual fire attention network and medium-scale fire classification benchmark. *IEEE Trans. Image Process.* **2022**, *31*, 6331–6343. [CrossRef] [PubMed]
31. Sun, G.; Zhang, X.; Jia, X.; Ren, J.; Zhang, A.; Yao, Y.; Zhao, H. Deep fusion of localized spectral features and multi-scale spatial features for effective classification of hyperspectral images. *Int. J. Appl. Earth Obs. Geoinf.* **2020**, *91*, 102157. [CrossRef]
32. Liu, S.; Guo, S.; Cong, J.; Yang, Y.; Guo, Z.; Gu, B. One-Shot Learning for Optical Coherence Tomography Angiography Vessel Segmentation Based on Multi-Scale U2-Net. *Mathematics* **2023**, *11*, 4890. [CrossRef]
33. Rahimy, E. Deep learning applications in ophthalmology. *Curr. Opin. Ophthalmol.* **2018**, *29*, 254–260. [CrossRef]
34. Kaymak, S.; Serener, A. Automated age-related macular degeneration and diabetic macular edema detection on oct images using deep learning. In Proceedings of the 2018 IEEE 14th International Conference on Intelligent Computer Communication and Processing (ICCP), Cluj-Napoca, Romania, 6–8 September 2018; IEEE: Piscataway, NJ, USA, 2018; pp. 265–269.
35. Rahil, M.; Anoop, B.; Girish, G.; Kothari, A.R.; Koolagudi, S.G.; Rajan, J. A deep ensemble learning-based CNN architecture for multiclass retinal fluid segmentation in oct images. *IEEE Access* **2023**, *11*, 17241–17251. [CrossRef]
36. Rasti, R.; Rabbani, H.; Mehridehnavi, A.; Hajizadeh, F. Macular OCT classification using a multi-scale convolutional neural network ensemble. *IEEE Trans. Med. Imaging* **2017**, *37*, 1024–1034. [CrossRef]
37. Hassan, B.; Qin, S.; Ahmed, R. RRI-Net: Classification of multi-class retinal diseases with deep recurrent residual inception network using OCT scans. In Proceedings of the 2020 IEEE International Symposium on Signal Processing and Information Technology (ISSPIT), Louisville, KY, USA, 9–11 December 2020; IEEE: Piscataway, NJ, USA, 2020; pp. 1–6.
38. Paul, D.; Tewari, A.; Ghosh, S.; Santosh, K. Octx: Ensembled deep learning model to detect retinal disorders. In Proceedings of the 2020 IEEE 33rd International Symposium on Computer-Based Medical Systems (CBMS), Rochester, MN, USA, 28–30 July 2020; IEEE: Piscataway, NJ, USA, 2020; pp. 526–531.
39. Kim, J.; Tran, L. Ensemble learning based on convolutional neural networks for the classification of retinal diseases from optical coherence tomography images. In Proceedings of the 2020 IEEE 33rd International Symposium on Computer-Based Medical Systems (CBMS), Rochester, MN, USA, 28–30 July 2020; IEEE: Piscataway, NJ, USA, 2020; pp. 532–537.
40. Naik, G.; Narvekar, N.; Agarwal, D.; Nandanwar, N.; Pande, H. Eye Disease Prediction Using Ensemble Learning and Attention on OCT Scans. In *Future of Information and Communication Conference*; Springer: Cham, Switzerland, 2024; pp. 21–36.
41. Yang, J.; Wang, G.; Xiao, X.; Bao, M.; Tian, G. Explainable ensemble learning method for OCT detection with transfer learning. *PLoS ONE* **2024**, *19*, e0296175. [CrossRef]
42. Lu, W.; Tong, Y.; Yu, Y.; Xing, Y.; Chen, C.; Shen, Y. Deep learning-based automated classification of multi-categorical abnormalities from optical coherence tomography images. *Transl. Vis. Sci. Technol.* **2018**, *7*, 41. [CrossRef]
43. Opoku, M.; Weyori, B.A.; Adekoya, A.F.; Adu, K. CLAHE-CapsNet: Efficient retina optical coherence tomography classification using capsule networks with contrast limited adaptive histogram equalization. *PLoS ONE* **2023**, *18*, e0288663. [CrossRef] [PubMed]
44. Hassan, E.; Elmougy, S.; Ibraheem, M.R.; Hossain, M.S.; AlMutib, K.; Ghoneim, A.; AlQahtani, S.A.; Talaat, F.M. Enhanced deep learning model for classification of retinal optical coherence tomography images. *Sensors* **2023**, *23*, 5393. [CrossRef] [PubMed]
45. Diao, S.; Su, J.; Yang, C.; Zhu, W.; Xiang, D.; Chen, X.; Peng, Q.; Shi, F. Classification and segmentation of OCT images for age-related macular degeneration based on dual guidance networks. *Biomed. Signal Process. Control* **2023**, *84*, 104810. [CrossRef]

46. Díaz, M.; Novo, J.; Cutrín, P.; Gómez-Ulla, F.; Penedo, M.G.; Ortega, M. Automatic segmentation of the foveal avascular zone in ophthalmological OCT-A images. *PLoS ONE* **2019**, *14*, e0212364. [CrossRef] [PubMed]
47. Udayaraju, P.; Jeyanthi, P.; Sekhar, B. A hybrid multilayered classification model with VGG-19 net for retinal diseases using optical coherence tomography images. *Soft Comput.* **2023**, *27*, 12559–12570. [CrossRef]
48. Subramanian, M.; Shanmugavadivel, K.; Naren, O.S.; Premkumar, K.; Rankish, K. Classification of retinal oct images using deep learning. In Proceedings of the 2022 International Conference on Computer Communication and Informatics (ICCCI), Coimbatore, India, 25–27 January 2022; IEEE: Piscataway, NJ, USA, 2022; pp. 1–7.
49. Stanojević, M.; Drašković, D.; Nikolić, B. Retinal disease classification based on optical coherence tomography images using convolutional neural networks. *J. Electron. Imaging* **2023**, *32*, 032004. [CrossRef]
50. Rahimzadeh, M.; Mohammadi, M.R. ROCT-Net: A new ensemble deep convolutional model with improved spatial resolution learning for detecting common diseases from retinal OCT images. In Proceedings of the 2021 11th International Conference on Computer Engineering and Knowledge (ICCKE), Mashhad, Iran, 28–29 October 2021; IEEE: Piscataway, NJ, USA, 2021; pp. 85–91.
51. Khan, H.; Ullah, I.; Shabaz, M.; Omer, M.F.; Usman, M.T.; Guellil, M.S.; Koo, J. Visionary vigilance: Optimized YOLOV8 for fallen person detection with large-scale benchmark dataset. *Image Vis. Comput.* **2024**, *149*, 105195. [CrossRef]
52. Yang, Z.; He, X.; Gao, J.; Deng, L.; Smola, A. Stacked attention networks for image question answering. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 21–29.
53. Khan, H.; Hussain, T.; Khan, S.U.; Khan, Z.A.; Baik, S.W. Deep multi-scale pyramidal features network for supervised video summarization. *Expert Syst. Appl.* **2024**, *237*, 121288. [CrossRef]
54. Pedersoli, M.; Lucas, T.; Schmid, C.; Verbeek, J. Areas of attention for image captioning. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 1242–1250.
55. Deeb, A.; Debow, A.; Mansour, S.; Shkodyrev, V. Covid-19 diagnosis with deep learning: Adjacent-pooling ctscan-Covid-19 classifier based on resnet and cbam. *Biomed. Signal Process. Control* **2023**, *86*, 105285. [CrossRef]
56. Munsif, M.; Khan, H.; Khan, Z.A.; Hussain, A.; Ullah, F.U.M.; Lee, M.Y.; Baik, S.W. Pv-anet: Attention-based network for short-term photovoltaic power forecasting. In Proceedings of the 8th International Conference on Next Generation Computing 2022, Jeju, Republic of Korea, 6–8 October 2022; pp. 133–135.
57. Farag, M.M.; Fouad, M.; Abdel-Hamid, A.T. Automatic severity classification of diabetic retinopathy based on densenet and convolutional block attention module. *IEEE Access* **2022**, *10*, 38299–38308. [CrossRef]
58. Wang, Q.; Teng, Z.; Xing, J.; Gao, J.; Hu, W.; Maybank, S. Learning attentions: Residual attentional siamese network for high performance online visual tracking. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 4854–4863.
59. Shu, X.; Yang, J.; Yan, R.; Song, Y. Expansion-squeeze-excitation fusion network for elderly activity recognition. *IEEE Trans. Circuits Syst. Video Technol.* **2022**, *32*, 5281–5292. [CrossRef]
60. Li, D.; Wen, G.; Kuai, Y.; Porikli, F. End-to-end feature integration for correlation filter tracking with channel attention. *IEEE Signal Process. Lett.* **2018**, *25*, 1815–1819. [CrossRef]
61. Paszke, A.; Gross, S.; Massa, F.; Lerer, A.; Bradbury, J.; Chanan, G.; Killeen, T.; Lin, Z.; Gimelshein, N.; Antiga, L.; et al. Pytorch: An imperative style, high-performance deep learning library. *Adv. Neural Inf. Process. Syst.* **2019**, *32*.
62. Kiefer, J.; Wolfowitz, J. Stochastic estimation of the maximum of a regression function. *Ann. Math. Stat.* **1952**, *23*, 462–466.
63. Kermany, D.S.; Goldbaum, M.; Cai, W.; Valentim, C.C.; Liang, H.; Baxter, S.L.; McKeown, A.; Yang, G.; Wu, X.; Yan, F.; et al. Identifying medical diagnoses and treatable diseases by image-based deep learning. *Cell* **2018**, *172*, 1122–1131. [CrossRef] [PubMed]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.