Cureus Journal of Computer Science

Open Access Original Article

A Lightweight Multi-Scale Dilated Convolutional Network for Retinal Disease Diagnosis Using Optical Coherence Tomography Images

Received 03/18/2025
Review began 04/01/2025
Review ended 07/02/2025
Published 07/22/2025

DOI: https://doi.org/10.7759/s44389-025-03919-4

Rajesh S. Aouti [1] , Sangram Redkar [1] , Prabha Dwivedi [2]

1. Robotics and Autonomous Systems, Arizona State University, Tempe, USA 2. Omics, Naval Surface Warfare Center Indian Head Division, Charles County, USA

Corresponding author: Rajesh S. Aouti, raouti@asu.edu

Abstract

Optical coherence tomography (OCT) is a critical imaging modality for diagnosing and monitoring retinal diseases. Although deep learning-based computer-aided diagnostic (CAD) systems have shown promise in automating OCT image analysis, conventional convolutional neural networks (CNNs) often fail to capture global context and long-range dependencies, limiting their overall performance. We propose a novel lightweight multi-scale dilated convolutional network to address these challenges and enhance OCT image classification. The proposed model employs a ResNet-18 backbone, replacing standard convolutional blocks with depth-wise separable convolutions to reduce model complexity, resulting in 1.97M parameters. To capture broader contextual information, dilated convolutions with layer-specific dilation rates are incorporated into parallel paths within each residual block. The model was evaluated on four OCT datasets (OCT2017, OCT-C8, OCTDL, and OCTID). On the OCT2017 dataset, the proposed model achieved an average accuracy of 99.85%, with precision, recall, specificity, and F1-score of 99.69%, 99.69%, 99.90%, and 99.69%, respectively. The OCT-C8 dataset attained an average accuracy of 98.90%, along with precision, recall, specificity, and F1-score values of 96.91%, 96.20%, 99.57%, and 96.55%, respectively. The model also demonstrated robust performance on the more challenging OCTDL and OCTID datasets, confirming its effectiveness in handling diverse and complex data. The proposed model addresses the inherent limitations of traditional CNNs in OCT image classification by integrating depth-wise separable convolutions and multi-scale dilated convolutions. Its high accuracy and strong performance across various datasets underscore its potential to improve diagnostic accuracy and efficiency in clinical retinal disease assessment.

Categories: Health Informatics, Image Processing and Analysis, Deep Learning
Keywords: retinal oct imaging, deep learning, dense residual blocks, convolution neural networks, dilation convolution, depth-wise separable convolution.

Introduction

Optical coherence tomography (OCT), a non-contact, non-invasive, and rapid imaging technique, has become an indispensable tool for ophthalmologists to visualize cross-sectional retinal layers, analyze retinal biomarkers, and design treatments [1]. It employs low-coherence interferometry to capture microscopic-resolution cross-sectional images of biological tissues. By analyzing the light reflected and backscattered from the retinal tissue, the distinct layers of the retina can be visualized in the resulting OCT image. OCT images contain various retinal biomarkers that serve as essential indicators for diagnosing and monitoring retinal diseases. Retinal diseases encompass a variety of abnormalities, many of which can lead to significant vision loss if left untreated. Age-related macular degeneration (AMD) [2] presents in two forms: dry AMD, characterized by drusen deposits - yellowish extracellular material - on the retina, which gradually thins the macula and causes central vision loss, and wet AMD, associated with choroidal neovascularization (CNV) [3], where abnormal blood vessels grow beneath the retina, leading to blood leakage and accelerating vision loss. CNV can also occur independently, resulting in fluid leakage, retinal detachment, and scarring. Central serous retinopathy (CSR) is another abnormality caused by fluid accumulation under the retina due to a retinal pigment epithelium defect, leading to macular detachment and temporary or recurrent vision distortion [4]. Diabetic macular edema (DME) [5], a common diabetic complication, occurs when retinal blood vessels leak fluid and protein into the macula, causing swelling and structural disruption. Closely related is diabetic retinopathy (DR), which progresses from microaneurysms and retinal hemorrhages to proliferative diabetic retinopathy (PDR), marked by abnormal blood vessel growth that can lead to blindness. Macular holes (MH) are full-thickness defects in the macula, often caused by vitreous traction, leading to central blind spots and blurred vision [6]. Key biomarkers include intraretinal fluid and subretinal hyperreflective material, often associated with advanced stages of neovascular AMD, drusen deposits, and outer retinal tabulation that characterize different forms of AMD. Disruption within the inner retinal layers has been associated with visual acuity results, especially in conditions like DR and DME. These biomarkers not only aid in identifying retinal pathologies but also play a significant role in monitoring disease progression and guiding treatment

How to cite this article
Aouti R S, Redkar S, Dwivedi P (July 22, 2025) A Lightweight Multi-Scale Dilated Convolutional Network for Retinal Disease Diagnosis Using Optical Coherence Tomography Images. Cureus J Comput Sci 2 : es44389-025-03919-4. DOI https://doi.org/10.7759/s44389-025-03919-4

strategies. Understanding these conditions and their distinct biomarkers is essential for early detection, personalized treatment planning, and improving patient outcomes. Despite the critical clinical value of OCT imaging, the manual interpretation of these images presents significant challenges. The subtle characteristics of specific lesions can be complex, leading to potential misinterpretations or missed diagnoses.

Computer-aided diagnostic (CAD) algorithms are increasingly being employed to overcome the challenges associated with the manual interpretation of OCT images. These systems utilize advanced computational methods to automate the detection of retinal abnormalities, significantly improving diagnostic accuracy and efficiency. Deep learning, especially through convolutional neural networks (CNNs) [7,8], has transformed this domain by automatically extracting hierarchical features from OCT scans. These features range from simple edge detection to intricate retinal biomarkers detection, eliminating the need for manual feature extraction. Advanced techniques, such as attention mechanisms, hybrid architectures integrating CNNs and Transformers, and lesion-aware networks, further enhance the robustness and interpretability of these models. While CNNs have revolutionized medical image analysis, including the classification and segmentation of retinal OCT images, they come with several limitations. One significant disadvantage is their inability to capture long-range dependencies and global context in images effectively. CNNs primarily focus on local features due to their convolutional kernel operations, which may result in the loss of holistic spatial relationships essential for understanding complex patterns in medical images. Additionally, CNNs can be computationally expensive and require substantial processing power, particularly for large-scale or high-resolution datasets, which can be a barrier to deployment in resource-constrained settings.

This study introduces a lightweight, multi-scale dilated convolutional network for OCT image classification. Built on the ResNet-18 framework, the model replaces traditional convolutional layers with depth-wise separable convolution blocks, significantly decreasing model complexity while maintaining robust performance. The main contributions of this work are outlined below.

1) We replaced standard convolution blocks with depth-wise separable blocks to reduce the model size having 1.97 M parameters.

2) Alongside the standard and residual paths, we incorporate parallel paths using dilated convolutions to expand the receptive field. The number of parallel paths and their dilation rates are layer-specific, with the 1st layer having four paths with dilation rates of [1, 3, 5, 7], gradually reducing to a single path with a dilation rate of [1] at the final layer.

Materials And Methods

Related work

The classification of retinal OCT images has garnered significant attention in recent years, driven by the need for accurate and automated diagnostic systems to assist in the detection and monitoring of retinal diseases. Traditional approaches to OCT image classification relied on handcrafted feature extraction methods combined with classical machine learning algorithms, such as Support Vector Machines (SVMs) and Random Forests. While these methods provided some success, they were limited by their dependency on manually engineered features, which often failed to capture the intricate patterns and biomarkers present in OCT images. The advent of deep learning revolutionized this field, mainly using CNNs. Kermany et al. [9] pioneered the use of CNNs for OCT image classification, developing a model capable of diagnosing common retinal diseases such as CNV, DME, and AMD. Their approach utilized a transfer learning-based InceptionV3 architecture. De Fauw et al. [10] introduced an end-to-end deep learning system combining segmentation and classification tasks for OCT image analysis. The system employed a U-Net for segmenting retinal layers and a CNN-based classifier for predicting disease categories. Their model demonstrated superior performance in detecting AMD and diabetic retinopathy (DR) with high precision and recall rates. This work underscored the importance of integrating segmentation with classification for improved diagnostic accuracy. Khan et al. [11] proposed a deep learning-based method for classifying retinal diseases using OCT images, which combined CNN with Ant Colony Optimization (ACO) for feature selection. Inspired by the foraging behavior of ants, ACO efficiently identified the most relevant features from OCT images, reducing input dimensionality while preserving diagnostic information. This integration enhanced computational efficiency and minimized overfitting, outperforming traditional techniques. Ma et al. [12] introduced a hybrid ConvNet-Transformer architecture to classify retinal OCT images. This model includes a low-level feature extraction component utilizing residual dense blocks to improve training efficiency. It features parallel branches-one based on a transformer to capture long-range dependencies and the other on a convolutional network to extract local contextual information. These features are combined using an adaptive re-weighting strategy to perform the final classification. He et al. [13] proposed an interpretable Swin-Poly Transformer network for automatic retinal OCT image classification. This architecture employs a Swin Transformer to construct multi-scale feature connections by shifting the window partition, enabling flexibility in modeling neighboring non-overlapping regions. Additionally, the Swin-Poly Transformer enhances cross-entropy

loss optimization by refining the importance of polynomial bases, further improving classification accuracy. Huang et al. [14] introduced GABNet, a novel architecture incorporating a global attention block for retinal OCT disease classification. The model integrates global attention mechanisms to enhance feature representation by focusing on global and local retinal image contexts, improving classification accuracy. The attention block ensures the effective modeling of long-range dependencies, complementing the localized processing of convolutional layers. GABNet demonstrated superior performance on multiple datasets, achieving high classification accuracy and robustness in identifying various retinal diseases.

Our proposed model uses the ResNet-18 [15] architecture as the backbone. It incorporates multi-scale feature extraction through dilated convolution, which increases the receptive field. Arbitrarily modifying receptive field sizes to capture multi-scale features can lead to a significant increase in both model parameters and computational cost. We employed depth-wise separable convolutions to mitigate this, reducing the model size to just 1.97 million parameters. This design choice allows our multi-scale architecture to boost performance while keeping the model lightweight.

Material

We assess the effectiveness of our proposed model using four publicly accessible OCT datasets: OCT2017 [9], OCT-C8 [16], OCTDL [17], and OCTID [18]. The OCT2017 dataset includes 84,484 OCT B-scans, gathered and provided by institutions such as the Shiley Eye Institute at the University of California San Diego, the California Retinal Research Foundation, Medical Center Ophthalmology Associates, Shanghai First People's Hospital, and Beijing Tongren Eye Center. The training comprises 37,205 images with CNV, 11,348 with DME, 8,616 with DRUSEN, and 26,315 normal cases. The test set contains 242 images for each class: CNV, DME, DRUSEN, and normal. The OCT-C8 dataset contains a total of 24,000 images categorized into eight distinct classes: AMD, CNV, CSR, DME, DR, Drusen, Macular Hole (MH), and Normal (healthy) retina. It consists of 18,400 training images, with 2,300 images per class and 2,800 images each for the validation and testing sets, maintaining 350 images per class. The detailed distribution of images across categories is presented in Table *1*.

Database	Class	Training Images	Validation Images	Testing Images	Total Images
OCT2017 [9]	CNV	37,205	8	242	37,455
	DME	11,348	8	242	11,598
	Drusen	8,616	8	242	8,866
	Normal	26,315	8	242	26,565
OCT-C8 [16]	AMD	2,300	350	350	3,000
	CNV	2,300	350	350	3,000
	CSR	2,300	350	350	3,000
	DME	2,300	350	350	3,000
	DR	2,300	350	350	3,000
	Drusen	2,300	350	350	3,000
	MH	2,300	350	350	3,000
	Normal	2,300	350	350	3,000

TABLE 1: Retinal OCT2017 and OCT-C8 datasets

CNV, Choroidal Neovascularization; DME, Diabetic Macular Edema; AMD, Age-related Macular Degeneration; CSR, Central Serous Retinopathy; DR, Diabetic Retinopathy; MH, Macular Hole

The OCTDL dataset comprises 2,064 B-scan OCT images, categorized into seven classes: AMD, DME, Epiretinal Membrane (ERM), Normal (NO), Retinal Vein Occlusion (RVO), Retinal Artery Occlusion (RAO), and Vitreomacular Interface Disease (VID). Meanwhile, the OCTID dataset includes 572 spectral-domain OCT volumetric scans divided into five groups: Normal (NO), MH, AMD, CSR, and DR. As neither OCTDL nor OCTID datasets come with a predefined training and testing split, we adopted an 80-20 split for training and testing purposes. The detailed distribution of images across the categories is provided in Table *2*.

Database	Class	Training Images	Testing Images	Total Images
OCTDL [17]	AMD	991	240	1,231
	DME	107	40	147
	ERM	117	38	155
	NO	264	68	332
	RVO	19	3	22
	RAO	87	14	101
	VID	66	10	76
OCTID [18]	AMD	48	7	55
	CSR	80	22	102
	DR	79	28	107
	MH	86	16	102
	Normal	164	42	206

TABLE 2: Retinal OCTDL and OCTID datasets

AMD, Age-related Macular Degeneration; DME, Diabetic Macular Edema; ERM, Epiretinal Membrane; NO, Normal; RVO, Retinal Vein Occlusion; RAO, Retinal Artery Occlusion; VID, Vitreomacular Interface Disease; CSR, Central Serous Retinopathy; DR, Diabetic Retinopathy; MH, Macular Hole

Method

The proposed lightweight model builds upon the ResNet-18 architecture (Figure *1*) to enhance the model's efficiency, contextual understanding, and feature representation for OCT image classification. The ResNet-18 model has demonstrated significant success in various computer vision tasks due to its simplicity, efficiency, and robust feature extraction capabilities. It is designed with residual connections that enable the learning of residual mappings, addressing the vanishing gradient problem commonly encountered in deep neural networks. These residual connections facilitate the flow of gradients through the network, effectively training deeper layers while maintaining computational efficiency. Convolution blocks with depth-wise separable convolutions significantly reduce parameters without compromising accuracy. We modify the convolution blocks by introducing parallel paths with dilated convolutions, which expands the receptive field and enables multi-scale feature extraction. These enhancements collectively improve the model's ability to classify retinal diseases while maintaining computational efficiency accurately.

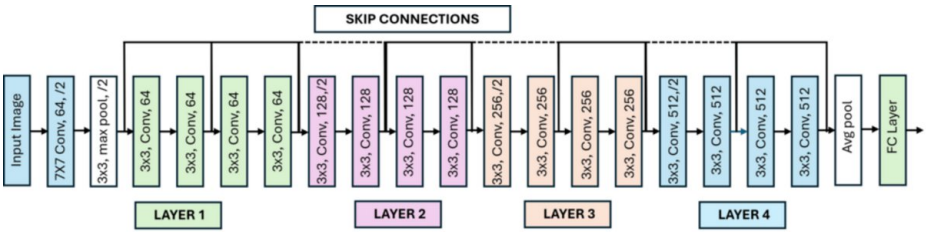

FIGURE 1: ResNet-18 architecture

Depth-Wise Separable Convolution

Depth-wise separable convolution [19] is a modification of the standard convolution operation that significantly reduces computational complexity while preserving the model's ability to extract meaningful features. Unlike standard convolution, where each filter is applied across all input channels simultaneously, followed by summation to produce the output feature map, depth-wise separable convolution decomposes the operation into two distinct steps. The first step, depth-wise convolution, applies a single filter independently to each input channel, focusing exclusively on spatial filtering and capturing spatial patterns within each channel without combining information. The second step, pointwise convolution, performs a 1 × 1 convolution across all channels to integrate the outputs of the

depth-wise convolution, combining channel-wise information to produce the final feature map. The proposed model employs depth-wise convolutions instead of the standard convolution blocks to ensure the network remains lightweight and fast, making it deployable on resource-constrained devices.

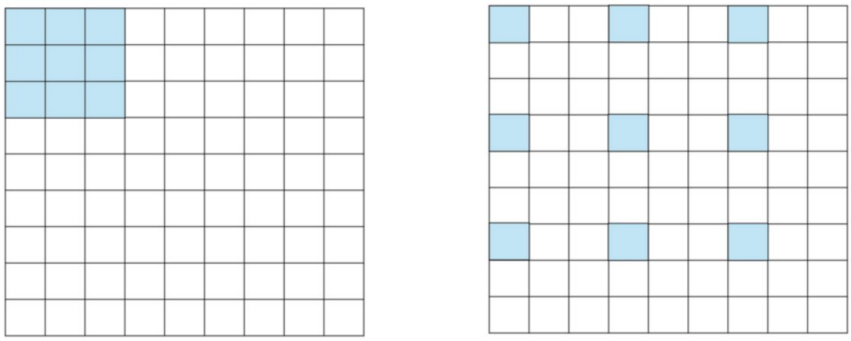

FIGURE 2: Pixel representation with a dilation convolution for 3 × 3 kernel size with different dilation rates

Parallel Paths

To expand the receptive field and improve the model's ability to capture long-range dependencies, we integrate parallel paths with dilated convolutions into each layer, which consists of two convolution blocks, followed by a residual connection, as shown in Figure *1*. Dilated convolutions modify standard convolution operations where the filter kernels are expanded by inserting spaces between kernel elements. This increases the receptive field without increasing the parameters or computation required, enabling the model to process more significant contextual regions within an image, as shown in Figure *2*. In the initial layers of the network, where fine-grained details such as edges and textures are crucial, the parallel paths use multiple dilation rates of [1, 3, 5, 7]. This setup allows the network to capture features at varying spatial scales simultaneously. As the network progresses to deeper layers, where the focus shifts from capturing fine-grained details to extracting high-level abstract features, the number of parallel paths gradually reduces. In the final layers, only a single path with a dilation rate of [1] is used, as shown in Figure *3*. This hierarchical design ensures that the network efficiently balances local feature extraction in the early layers with broader contextual analysis in the deeper layers. Including parallel paths with dilated convolutions also helps mitigate one of the key limitations of standard convolutional layers, their inability to effectively capture global dependencies due to their localized receptive field. By expanding the receptive field at different scales, the model can identify relationships and structures spanning wider portions of the image.

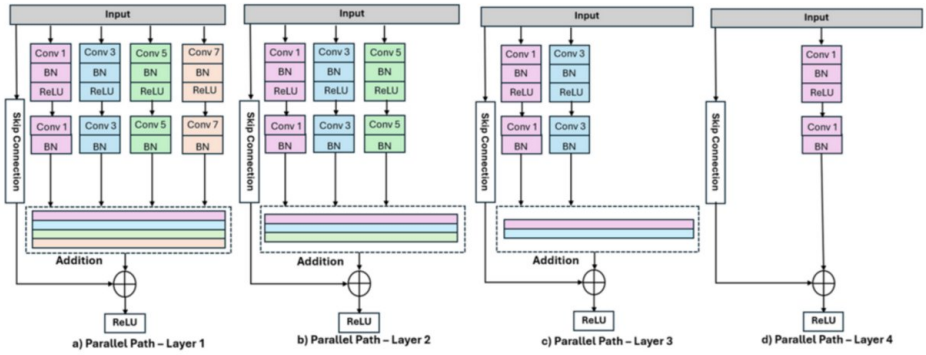

FIGURE 3: Proposed multi-scale residual blocks at each layer of the ResNet-18 architecture

Training Parameters

The input images are resized to 256 × 256 and then center-cropped to 224 × 224. Data augmentation is employed to introduce variability in the dataset. Transformations such as vertical and horizontal flips and

affine transformations such as translation and rotation of 15 degrees are randomly applied with a 30% probability within each training batch. The training was performed using the PyTorch framework, utilizing the ADAM optimizer with parameters β1 = 0.9, β2 = 0.999, and an initial learning rate of 0.001. We used a reduce-on-plateau learning rate scheduler where the learning rate value is reduced by a factor of 0.1 with patience of 10. A batch size of 64 was trained for 100 epochs, using the cross-entropy (BCE) loss function for training.

Results

Performance evaluation

We evaluate the performance of the proposed model using standard metrics, including accuracy, precision, recall, specificity, and F1-score, to assess its classification capabilities comprehensively. Accuracy measures the overall correctness of predictions, while precision evaluates the proportion of correctly identified positive predictions among all positive predictions. Recall indicates the model's ability to identify true positives, and specificity measures its effectiveness in identifying true negatives. The F1-score is the harmonic average of precision, and recall ranges from 0 to 1. A confusion matrix is used to further analyze the distribution of predictions, offering detailed insights into the model's performance. Additionally, Grad-CAM visualizations are employed to interpret the model's decision-making process, highlighting the key image regions that influenced its predictions and improving its transparency for clinical applications. The confusion matrices for all datasets are presented in Figure *4*. The corresponding Grad-CAM visualizations are illustrated in Figure *5*.

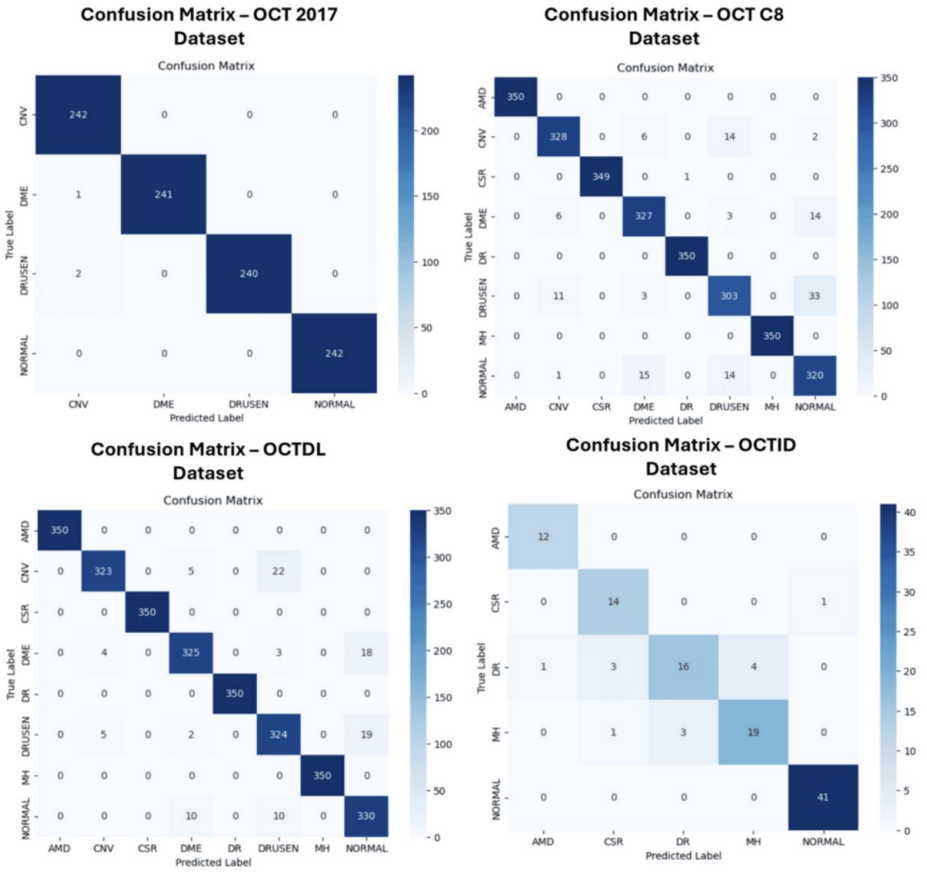

FIGURE 4: Confusion matrix obtained for the retinal OCT datasets

OCT, Optical Coherence Tomography; CNV, Choroidal Neovascularization; DME, Diabetic Macular Edema; AMD, Age-related Macular Degeneration; CSR, Central Serous Retinopathy; DR, Diabetic Retinopathy; MH, Macular Hole

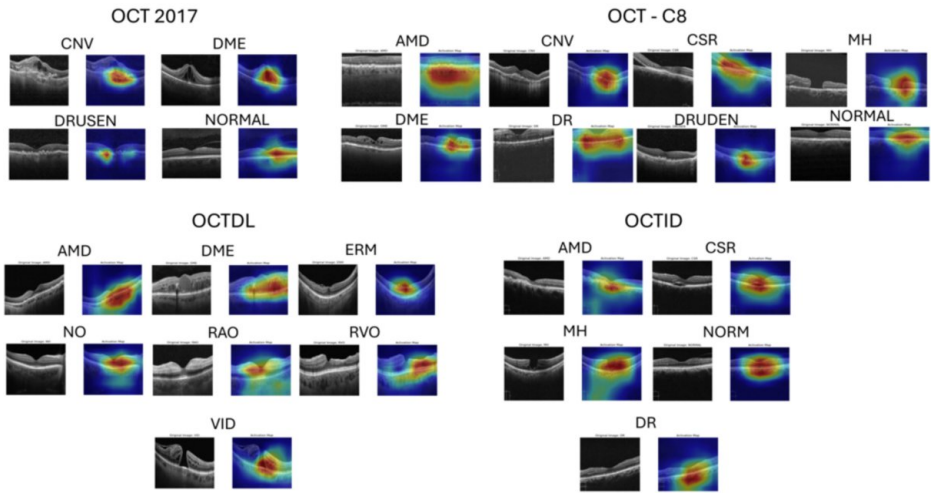

FIGURE 5: Retinal OCT images and their class activation maps obtained for the OCT datasets

OCT, Optical Coherence Tomography; CNV, Choroidal Neovascularization; DME, Diabetic Macular Edema; AMD, Age-related Macular Degeneration; CSR, Central Serous Retinopathy; DR, Diabetic Retinopathy; MH, Macular Hole; ERM, Epiretinal Membrane; NO, Normal; RVO, Retinal Vein Occlusion; RAO, Retinal Artery Occlusion; VID, Vitreomacular Interface Disease

Performance results

The proposed lightweight multi-scale dilated convolutional network demonstrates reliable performance across multiple OCT datasets, achieving high classification accuracy and consistent evaluation metrics. On the OCT2017 dataset, the model recorded an average accuracy of 99.85%, with precision, recall, specificity, and F1-score values of 99.69%, 99.69%, 99.90%, and 99.69%, respectively. On the OCT-C8 dataset, the model achieved an average accuracy of 98.90%, precision of 96.91%, recall of 96.20%, specificity of 99.57%, and F1-score of 96.55%. Notably, it achieved perfect classification metrics for AMD, CSR, DR, and MH, demonstrating its capability to handle a broad spectrum of retinal diseases. However, slightly reduced recall for CNV and DME suggests areas for improvement, potentially due to overlapping features with other retinal conditions. These findings are detailed in Table *3*, which presents the class-wise evaluation metrics.

Dataset	Class	Accuracy	Precision	Recall	Specificity	F1-Score
OCT2017	CNV	0.9969	0.9878	1.0000	0.9959	0.9938
	DME	0.9990	1.0000	0.9959	1.0000	0.9979
	Drusen	0.9979	1.0000	0.9917	1.0000	0.9959
	Normal	1.0000	1.0000	1.0000	1.0000	1.0000
	Average	0.9985	0.9969	0.9969	0.9990	0.9969
OCT-C8	AMD	1.0000	1.0000	1.0000	1.0000	1.0000
	CNV	0.9857	0.9480	0.9371	0.9927	0.9425
	CSR	0.9996	1.0000	0.9971	1.0000	0.9986
	DME	0.9832	0.9316	0.9343	0.9902	0.9330
	DR	0.9996	0.9972	1.0000	0.9996	0.9986
	Drusen	0.9721	0.9072	0.8657	0.9873	0.8860
	MH	1.0000	1.0000	1.0000	1.0000	1.0000
	Normal	0.9718	0.8672	0.9143	0.9800	0.8901
	Average	0.9890	0.9691	0.9620	0.9957	0.9655

TABLE 3: Class-wise performance of the proposed model on the OCT2017 and OCT-C8 datasets

CNV, Choroidal Neovascularization; DME, Diabetic Macular Edema; AMD, Age-related Macular Degeneration; CSR, Central Serous Retinopathy; DR, Diabetic Retinopathy; MH, Macular Hole

For the OCTDL dataset, the average accuracy was 99.58%, with a precision of 98.09%, recall of 97.74%, specificity of 99.76%, and F1-score of 97.88%. While the model performed exceptionally for ERM and RAO, the classification of conditions like RVO and NO presented challenges, indicated by lower precision and F1 score. This suggests a need for additional training data and fine-tuning to improve the detection of rarer conditions. On the OCTID dataset, the model achieved an average accuracy of 95.48%, precision of 86.90%, recall of 88.52%, specificity of 97.21%, and F1-score of 87.33%. Performance for Normal images was nearly perfect, but reduced recall for DR and MH suggests challenges in distinguishing these categories from others due to overlapping features and dataset-specific variations. The results indicate that the proposed model effectively balances computational efficiency with high classification performance across diverse datasets. The class-wise evaluation metrics for OCTDL and OCTID datasets are tabulated in Table *4*.

Dataset	Class	Accuracy	Precision	Recall	Specificity	F1-Score
OCTDL	AMD	0.9927	1.0000	0.9878	1.0000	0.9939
	DME	0.9952	0.9730	0.9730	0.9973	0.9730
	ERM	1.0000	1.0000	1.0000	1.0000	1.0000
	NO	0.9903	0.9412	1.0000	0.9885	0.9697
	RAO	1.0000	1.0000	1.0000	1.0000	1.0000
	RVO	0.9952	0.9524	0.9524	0.9974	0.9524
	VID	0.9976	1.0000	0.9286	1.0000	0.9630
	Average	0.9958	0.9809	0.9774	0.9976	0.9788
OCTID	AMD	0.9913	0.9231	1.0000	0.9903	0.9600
	CSR	0.9565	0.7778	0.9333	0.9600	0.8485
	DR	0.9043	0.8421	0.6667	0.9670	0.7442
	MH	0.9304	0.8261	0.8261	0.9565	0.8261
	Normal	0.9913	0.9762	1.0000	0.9865	0.9880
	Average	0.9548	0.8690	0.8852	0.9721	0.8733

TABLE 4: Class-wise performance of the proposed model on the OCTDL and OCTID datasets

AMD, Age-related Macular Degeneration; DME, Diabetic Macular Edema; ERM, Epiretinal Membrane; NO, Normal; RVO, Retinal Vein Occlusion; RAO, Retinal Artery Occlusion; VID, Vitreomacular Interface Disease; CSR, Central Serous Retinopathy; DR, Diabetic Retinopathy; MH, Macular Hole

On the OCT2017 dataset, the proposed model results are competitive with or surpass those of several state-of-the-art models, including those of Karthik et al. [20], Choudhary et al. [21], and Salaheldin et al. [22]. While the model of He et al. [13] slightly outperforms our model by a margin of 0.001 across precision, recall, and F1-score, this comes at the cost of significantly higher computational complexity. Specifically, He et al. [13] model comprises 27.5M parameters, whereas our proposed model requires only 1.97M parameters. These results highlight the proposed model's efficiency in achieving nearly identical classification performance while significantly more lightweight and computationally feasible. On the OCT-C8 dataset, the proposed model results are competitive with or outperforms several state-of-the-art models, including those of Karthik et al. [20] and Subramanian et al. [16]. These results underscore the effectiveness of the proposed model in handling complex datasets while maintaining high classification performance. A comparison of the performance of the proposed model with the state-of-the-art models is tabulated in Table *5*.

Dataset	Method	Accuracy	Precision	Recall	Specificity	F1-Score
OCT2017	He et al. [13]	0.9980	0.9980	0.9980	-	0.9980
	Karthik et al. [20]	0.9820	0.9900	0.9900	-	0.9900
	Choudhary et al. [21]	0.9917	-	0.9900	0.9950	-
	Salaheldin et al. [22]	0.9956	0.9893	0.9911	0.9970	0.9902
	Proposed Method	0.9985	0.9969	0.9969	0.9990	0.9969
OCT-C8	He et al. [13]	0.9711	0.9713	0.9711	-	0.9962
	Karthik et al. [22]	0.9240	0.9300	0.9200	-	0.9200
	Subramanian et al. [16]	0.9720	0.9725	0.9713	-	0.9725
	Proposed Method	0.9890	0.9691	0.9620	0.9957	0.9655

TABLE 5: Comparison of the performance of the proposed model with state-of-the-art models

Discussion

The proposed lightweight multi-scale dilated convolutional network addresses key challenges in retinal OCT image analysis by balancing computational efficiency with robust diagnostic performance. Replacing standard convolutional layers with depth-wise separable convolutions effectively reduces the model's size and computational cost without a significant drop in accuracy, making it suitable for deployment in resource-constrained devices. Additionally, incorporating dilated convolutions in the residual blocks expands the receptive field, enabling the model to capture broader contextual information for identifying subtle or complex retinal patterns. Notably, it performs well on the OCT2017 and OCT-C8 datasets, surpassing state-of-the-art techniques in accuracy and specificity.

However, the model encounters challenges when applied to datasets with substantial class imbalance and limited sample diversity, as observed in OCTDL and OCTID. These datasets vary significantly in terms of class distribution, image count, disease types, and clinical origin, which can affect the model’s ability to generalize across different real-world scenarios. For instance, OCT2017 contains a large number of high-quality images predominantly representing four common retinal conditions, with relatively balanced and extensive training data. In contrast, OCT-C8 includes a broader range of eight disease classes, but with fewer samples per class, especially in the test set. OCTDL and OCTID introduce further variability as they include rarer and clinically complex conditions and suffer from extreme class imbalance and limited sample size, particularly in OCTID, which includes fewer than 60 samples per class in some cases. This heterogeneity in disease types, annotation consistency, image acquisition devices, and sample sizes across datasets underscores the risk of dataset-specific bias when relying solely on benchmark performance. Additionally, they include rare and clinically complex conditions that exhibit minimal visual differences and suffer from inadequate representation, which leads to a drop in the model’s sensitivity and precision. This indicates the need for better feature learning strategies and enhanced representation of minority classes. Addressing this concern, future work could involve cross-domain evaluation training on one dataset and testing on another to assess generalizability. Additionally, domain adaptation techniques or multi-domain training strategies could be leveraged to build models that are robust across diverse clinical environments.

Another critical limitation arises from the model’s difficulty in accurately identifying rare retinal diseases. In datasets like OCTDL and OCTID, conditions such as RAO and VID have significantly fewer samples compared to more prevalent diseases like AMD or DME. This imbalance results in biased learning, where the model tends to favor majority classes, leading to lower detection rates for rare but clinically significant conditions. To mitigate this, future efforts could focus on incorporating few-shot learning methods that allow the model to generalize from very limited examples. Synthetic data generation using generative adversarial networks (GANs) is another promising avenue to augment underrepresented classes, improve diversity, and reduce overfitting. These approaches would enhance the model’s capacity to recognize rare patterns and contribute to improved performance.

Interpretability remains a vital concern for real-world clinical deployment, as medical practitioners must understand and trust the decisions made by AI systems. To promote transparency, Grad-CAM was employed to generate visual explanations of model predictions, highlighting salient regions in OCT scans that influence the output. While this improves the model's interpretability to an extent, Grad-CAM offers only coarse localization and does not guarantee alignment with the model's internal reasoning. Moreover,

the effectiveness of these visualizations can vary depending on the disease category and the visual clarity of features. Future work should include clinician-in-the-loop evaluations to assess the utility of these explanations in a real diagnostic workflow. Additionally, exploring more advanced interpretability techniques that offer fine-grained and class-specific explanations could further bridge the gap between automated predictions and clinical reasoning.

Consideration for real-world deployment is the model's robustness to noise and imaging artifacts, which are commonly encountered in clinical OCT scans due to motion blur, illumination inconsistencies, patient movement, and variations in imaging devices. Although the model performs well under clean benchmark conditions, its performance may degrade in the presence of such artifacts. To address this, future iterations should integrate artifact-aware augmentation strategies during training, simulate real-world degradations, and incorporate adversarial training schemes to improve resilience. Domain generalization techniques could also be leveraged to enhance the model's ability to generalize across diverse acquisition settings, ensuring consistent reliability when applied across different clinical facilities and imaging devices.

Ablation study

To evaluate the effectiveness of different architectural modifications, an ablation study was conducted on variants of the ResNet-18 backbone, progressively integrating depth-wise separable convolutions (DW), dilated convolutions (DIA), and the convolutional block attention module (CBAM) on the OCTDL dataset. The results, summarized in Table *6*, demonstrate the trade-offs between model complexity and performance across key metrics, including accuracy, precision, recall, specificity, and F1 score. The baseline ResNet-18 model, with 11.18 million parameters, achieved an accuracy of 0.9931 and an F1 score of 0.9622. Replacing standard convolutions with depth-wise separable convolutions drastically reduced the model size to 1.44 million parameters, an approximate 8× reduction in complexity. While the accuracy and specificity were largely preserved or slightly improved, we observed a moderate drop of 0.5%, 3.7% and 2.5% for precision, recall, and F1-score, respectively. This performance trade-off is expected due to the reduced representational capacity of the lightweight architecture but is acceptable given the substantial gains in computational efficiency, especially in resource-constrained environments. Further enhancement with dilated convolutions (DW + DIA) marginally increased the parameter count to 1.86 million but yielded consistent performance improvements. Compared to the baseline, this configuration improved accuracy, precision, recall, specificity, and F1 score by 0.2%, 0.8%, 2.4%, 0.3%, and 1.7%, respectively. These results confirm that the incorporation of dilated convolutions enhances the network's ability to capture multi-scale contextual features, which is particularly beneficial in distinguishing subtle variations in retinal OCT images.

Model	Parameters (in Millions)	Accuracy	Precision	Recall	Specificity	F1-Score
ResNet-18	11.18	0.9931	0.9724	0.9538	0.9939	0.9622
ResNet-18 + DW	1.44	0.9931	0.9675	0.9184	0.9956	0.9379
ResNet-18 + DW + DIA (Proposed Architecture)	1.86	0.9958	0.9809	0.9774	0.9976	0.9788
ResNet-18 + DW + DIA + CBAM	1.97	0.9626	0.7798	0.7873	0.9748	0.7723

TABLE 6: Ablation study of the proposed method on OCTDL dataset

DW, Depth-Wise Convolution; DIA, Multi-Path Dilation Blocks; CBAM, Convolutional Block Attention Module

To explore whether additional performance gains could be achieved, we integrated the CBAM) [23] into each parallel convolutional path of the proposed model. CBAM is a lightweight attention mechanism that applies sequential channel and spatial attention to emphasize informative features. The channel attention module recalibrates feature responses across different channels based on their global importance, while the spatial attention module emphasizes disease-relevant regions within each feature map. Although CBAM is effective in many vision tasks, its application in this context led to a decline in performance across all metrics. Accuracy dropped from 0.9958 to 0.9626, while F1-score fell sharply from 0.9788 to 0.7723. The sequential application of channel and spatial attention may have overemphasized localized features while suppressing globally distributed patterns, undermining the benefit of the multi-scale receptive fields introduced by dilated convolutions. This imbalance likely impaired the model's capacity to generalize across complex retinal structures. Additionally, the slight increase in parameter count and architectural complexity introduced by CBAM may have led to overfitting.

Conclusions

The proposed lightweight multi-scale dilated convolutional network advances retinal OCT image analysis, combining computational efficiency with robust diagnostic performance. By integrating depth-wise separable convolutions and multipath dilated convolutions, the model effectively addresses key challenges such as capturing global context, handling subtle retinal patterns, and maintaining scalability for deployment in resource-constrained settings. Extensive evaluation across multiple datasets demonstrates its ability to achieve high accuracy, precision, and recall, often outperforming existing methods. Despite its strengths, there remain opportunities for further enhancement, particularly in improving performance for rare or overlapping retinal abnormalities and increasing its generalizability to diverse datasets. Its contributions to automated OCT image analysis have the potential to significantly enhance retinal disease management and support the broader vision of precision healthcare.

Additional Information

Author Contributions

All authors have reviewed the final version to be published and agreed to be accountable for all aspects of the work.

Concept and design: Rajesh S. Aouti, Sangram Redkar, Prabha Dwivedi

Acquisition, analysis, or interpretation of data: Rajesh S. Aouti, Prabha Dwivedi

Drafting of the manuscript: Rajesh S. Aouti, Sangram Redkar

Critical review of the manuscript for important intellectual content: Rajesh S. Aouti, Sangram Redkar, Prabha Dwivedi

Supervision: Sangram Redkar, Prabha Dwivedi

Disclosures

Human subjects: All authors have confirmed that this study did not involve human participants or tissue. **Animal subjects:** All authors have confirmed that this study did not involve animal subjects or tissue. **Conflicts of interest:** In compliance with the ICMJE uniform disclosure form, all authors declare the following: **Payment/services info:** All authors have declared that no financial support was received from any organization for the submitted work. **Financial relationships:** All authors have declared that they have no financial relationships at present or within the previous three years with any organizations that might have an interest in the submitted work. **Other relationships:** All authors have declared that there are no other relationships or activities that could appear to have influenced the submitted work.

Acknowledgements

We want to extend our gratitude to the Research Computing Facility service at Arizona State University.

References

1. Huang D, Swanson EA, Lin CP, et al.: Optical coherence tomography. Science. 1991, 254:1178-1181. 10.1126/science.1957169
2. Rasti RR, Rabbani H, Mehridehnavi A, Hajizadeh F: Macular OCT classification using a multi-scale convolutional neural network ensemble. IEEE Transactions on Medical Imaging. 2018, 37:1024-1034. 10.1109/tmi.2017.2780115
3. Thomas CJ, Mirza RG, Gill MK: Age-related macular degeneration. Medical Clinics of North America. 2021, 105:473-491. 10.1016/j.mcna.2021.01.003
4. Nicholson B, Noble J, Forooghian F, Meyerle C: Central serous chorioretinopathy: Update on pathophysiology and treatment. Survey of Ophthalmology. 2013, 58:103-126. 10.1016/j.survophthal.2012.07.004
5. Ciulla TA, Amador AG, Zinman B: Diabetic retinopathy and diabetic macular edema: pathophysiology, screening, and novel therapies. Diabetes Care. 2003, 26:2653-2664. 10.2337/diacare.26.9.2653
6. Hwang S, Kang SW, Kim SJ, et al.: Risk factors for the development of idiopathic macular hole: a nationwide population-based cohort study. Scientific Reports. 2022, 12:21778. 10.1038/s41598-022-25791-1
7. Lee CS, Baughman DM, Lee AY: Deep learning is effective for classifying normal versus age-related macular degeneration OCT images. Ophthalmology Retina. 2017, 1:322-327. 10.1016/j.oret.2016.12.009
8. Choi KJ, Choi JE, Roh HC, et al.: Deep learning models for screening of high myopia using optical coherence tomography. Scientific Reports. 2021, 11:21663. 10.1038/s41598-021-00622-x
9. Kermany DS, Goldbaum M, Cai W, et al.: Identifying medical diagnoses and treatable diseases by image-based deep learning. Cell. 2018, 172:1122-1131. 10.1016/j.cell.2018.02.010
10. De Fauw J, Ledsam JR, Romera-Paredes B, et al.: Clinically applicable deep learning for diagnosis and referral in retinal disease. Nature Medicine. 2018, 24:1342-1350. 10.1038/s41591-018-0107-6
11. Khan A, Pin K, Aziz A, Han JW, Nam Y: Optical coherence tomography image classification using hybrid deep learning and ant colony optimization. Sensors. 2023, 23:6706. 10.3390/s23156706
12. Ma Z, Xie Q, Xie P, Fan F, Gao X, Zhu J: HCTNet: a hybrid ConvNet-transformer network for retinal optical coherence tomography image classification. Biosensors. 2022, 12:542. 10.3390/bios12070542

13. He J, Wang J, Han Z, Ma J, Wang C, Qi M: An interpretable transformer network for the retinal disease classification using optical coherence tomography. Scientific Reports. 2023, 13:3637. 10.1038/s41598-023-30853-z
14. Huang X, Ai Z, Wang H, et al.: GABNet: global attention block for retinal OCT disease classification. Frontiers in Neuroscience. 2023, 17:1143422. 10.3389/fnins.2023.1143422
15. He K, Zhang X, Ren S, Sun J: Deep residual learning for image recognition. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA. 2016, 770-778. 10.1109/CVPR.2016.90
16. Subramanian M, Shanmugavadivel K, Naren OS, Premkumar K, Rankish K: Classification of retinal oct images using deep learning. 2022 International Conference on Computer Communication and Informatics (ICCCI), Coimbatore, India. 2022, 1-7. 10.1109/ICCCI54379.2022.9740985
17. Kulyabin M, Zhdanov A, Nikiforova A, et al.: OCTDL: Optical coherence tomography dataset for image-based deep learning methods. Scientific Data. 2024, 11:365. 10.1038/s41597-024-03182-7
18. Gholami P, Roy P, Parthasarathy MK, Lakshminarayanan V: OCTID: Optical coherence tomography image database. Computers & Electrical Engineering. 2020, 81:106532. 10.1016/j.compeleceng.2019.106532
19. Chollet F: Xception: Deep learning with depthwise separable convolutions. 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA. 2017, 1800-1807. 10.1109/CVPR.2017.195
20. Karthik K, Mahadevappa M: Convolution neural networks for optical coherence tomography (OCT) image classification. Biomedical Signal Processing and Control. 2023, 79:104176. 10.1016/j.bspc.2022.104176
21. Choudhary A, Ahlawat S, Urooj S, Pathak N, Lay-Ekuakille A, Sharma N: A deep learning-based framework for retinal disease classification. Healthcare. 2023, 11:212. 10.3390/healthcare11020212
22. Salaheldin AM, Abdel Wahed M, Saleh N: Machine learning-based platform for classification of retinal disorders using optical coherence tomography images. Artificial Intelligence and Sustainable Computing. Algorithms for Intelligent Systems . Pandit M, Gaur MK, Rana PS, Tiwari A (ed): Springer, Singapore; 2022. 269-283. 10.1007/978-981-19-1653-3_21
23. Woo S, Park J, Lee JY, Kweon IS: CBAM: Convolutional block attention module. 2018, 10.48550/arXiv.1807.06521