

Lightweight deep learning approach for retinal OCT image classification: A CNN with hybrid pooling and optimized learning

Parth R. Dave^{1,2}, Nikunj H. Domadiya²

¹Department of Computer Engineering, Gujarat Technological University-Ahmedabad, Gujarat, India

²Department of Computer Engineering, L.D. College of Engineering, Ahmedabad, Gujarat, India

Article Info

Article history:

Received Apr 16, 2025

Revised Oct 16, 2025

Accepted Nov 5, 2025

Keywords:

Artificial intelligence

Convolutional neural network

Machine learning

Optical coherence tomography

Deep learning

ABSTRACT

Optical coherence tomography (OCT) is a non-invasive technique through which a retina specialist can see the structure behind the eye. This technology offers a key role to identify various abnormalities in the retina: Drusen, diabetic macular edema (DME) and choroidal neovascularization (CNV). However, manual analysis of OCT scans can be time-consuming and prone to variability among clinicians. To address this challenge, we present a lightweight and explainable deep learning-based approach for automatic classification of retinal OCT images. The primary goal of this research is a model that delivers high diagnostic accuracy. A computer-aided suggestive method can help retinal doctors automatically classify the anomalies with more confidence and precision. In this paper, we proposed a novel approach based on deep learning: a six-layer convolutional neural network (CNN) integrated with hybrid pooling for effective feature extraction. Data augmentation and exponential learning rate is implemented to handle data imbalance between classes and for stabilized learning consecutively. Our proposed approach achieved 98.75% of accuracy while testing on the dataset. To further enhance the interpretability of the model, we also integrate explainable AI (XAI) using class activation mapping (CAM) to visualize the critical regions in the retina that contribute to the classification decisions.

This is an open access article under the [CC BY-SA](#) license.



Corresponding Author:

Nikunj H. Domadiya

Department of Computer Engineering, L.D. College of Engineering

Ahmedabad, Gujarat, India

Email: domadiyanikunj002@gmail.com, nikunjdomadiya.ce@ldce.ac.in

1. INTRODUCTION

Retinal optical coherence tomography (OCT) imaging plays a vital role in analyzing the layers of the retina without any surgical treatment in the eye. It helps identify different anomalies and their progression, if any. In ophthalmology, the use of retinal OCT imaging plays a crucial role in shaping treatment decisions and keeping a watch on the success of therapies such as laser treatment, eye injections, and surgeries. Classifying retinal OCT images with the help of modern computerized techniques can be leveraged in terms of suggestive systems to the retina specialist. The ability of automated image categorization is to provide reliable and impartial evaluations of retinal images, minimizing human error and diagnostic inconsistencies.

Conventional machine learning algorithms take the texture features of the images as an input through which local or global information of the image can be represented. Once the feature vector is implemented by different feature extraction techniques: Local binary pattern (LBP) [1], histogram of oriented

gradients (HOG) [2], gray-level co-occurrence matrix (GLCM) [3], edge detection, and shape detection, the vector is provided to an artificial neural network (ANN), support vector machine (SVM) [4] or other supervised algorithm which eventually classify the image in the required class. The lagging part of conventional machine learning algorithms is that the images are described by some feature and it is not learning the features itself, hence may lead to lossy information representation and end up with less accuracy of the algorithm.

Deep learning has attracted significant interest from researchers in computer vision because of its capability to independently learn intricate details and features from input images. CNN has changed the modus operandi of shallow and conventional networks, and it exceeds their performance over them in pots of aspects. Deep learning isn't confined to just classification tasks—it extends its applications to a wide range of areas, including object detection, medical imaging, video analysis, speech recognition, natural language processing, and beyond.

In recent time, for the field of computer vision, many pretrained models: AlexNet [5], VGG16 [6], VGG19 [6], Resnet50 [7], and GoogleNet [8]. are available, which are actually trained on very large image datasets to learn the diversity. These models often contain a large and diverse set of trainable parameters, which can be utilized for feature extraction from a specific image dataset by eliminating the final fully connected layer(s).

CNN learns the spatial information from input image or data with the help of a layered convolutional and pooling architecture which then finally gets connected to the dense layer for the end point classification. The dense section may consist of one or more fully connected hidden units preceding the final output component. This output component contains neurons corresponding to the number of classes in the dataset, enabling classification. There are many activation functions, parameters, hyperparameters and mechanisms available to handle the underfitting or overfitting of the model and make it more generalised towards the test dataset. To further enhance the interpretability of the proposed CNN model through explainable AI, class activation mapping (CAM) [9] was employed. This method enables visualization of the regions within retinal OCT images that contribute most significantly to classification outcomes. By generating CAMs at different convolutional layers, the feature learning process of the model can be analyzed in a clinically meaningful manner.

This paper is divided into five sections. Section 2 covers the Related Work. The Proposed Approach is covered in section 3. Section 4 contains Experimental Setup and Results and finally the last section covers the Conclusion and Future Work.

2. RELATED WORK

Medical image processing has become a very good suggestive system in recent era due to its capacity to generate more accurate and less error prone results. Over time, a variety of traditional as well as cutting-edge methods have been applied for the classification of retinal OCT images.

Srinivasan *et al.* [10] applied image denoising, retinal curvature correction, and region-focused cropping before extracting feature vectors using HOG descriptors [2] for retinal OCT images. Each block's descriptor vector normalized as with a small constant. The values of the vectors were capped and renormalized. The final feature vector comprised normalized histograms from all the blocks. For multi class classification, they have used SVM [4] classifier in one vs one method. It consists of such three linear SVMs as the dataset contains three classes: AMD, DME and Normal. The retinal image dataset for the said purpose was created locally by taking the images of various patients with retinal disorders.

Liu *et al.* [11] proposed another classical machine learning method to diagnose macular pathologies in retinal OCT images. They used image alignment and construction of global and local features of the images. The global descriptor utilizes a multiscale spatial pyramid [12], while the local descriptor employs a PCA-based reduced LBP histogram [1]. They recorded good accuracy with the help of a SVM and locally created dataset for anomalies: MH, macular edema (ME) and AMD.

Alsaih *et al.* [13], developed yet another conventional approach using feature extraction techniques: HOG [2] and LBP [1]. To reduce the dimensionality of the feature vector, principal component analysis (PCA) was applied and after that SVM [4] was applied for end classification on local dataset.

On the other side, many cutting-edge techniques have been implemented to classify the retinal OCT images. Huang *et al.* [14] implemented a layer-guided convolutional neural network (CNN). They proposed different networks to extract the information from the retinal layers and then it is provided to the final deep learning network to classify the images. By concentrating on the retinal layer-specific information with the help of some transfer learning approach, they achieved better accuracy on OCT2017 [15] and HUCM [14] retinal image datasets with four classes: CNV, DME, DRUSEN, and normal.

Another approach with deep learning was implemented by Kim and Tran [16], which proposes the implementation of two binary models. Before applying the actual classification, they performed

preprocessing, area of interest segmentation using U-net, and rotation via histogram orientation to extract the deep information from the images of OCT2017 [15] dataset. They achieved better accuracy compared to the approach derived [14].

Diao *et al.* [17] proposed deep learning models to classify AMD from retinal OCT images. This approach encompasses two novel models: CM-CNN and CAM-Unet. CM-CNN improves the classification process by performing segmentation via some means, and CAM-UNet enhances the segmentation task by integrating class activation maps. The approach achieved good accuracy for the targeted work.

Hassan *et al.* [18] implemented a blended approach of deep learning and classical machine learning to perform automated classification of retinal OCT images using OCT2017 [15] dataset. The enhanced optical coherence tomography (EOCT) model presented in this study demonstrates a significant advancement by combining deep learning (ResNet-50 [7]) with machine learning (Random Forest) and optimizing with dual SGD and Adam optimizers. This model achieves state-of-the-art accuracy.

Paul *et al.* [19], the researchers introduced a model called OCTx to classify retinal OCT images into four categories: diabetic macular edema (DME), choroidal neovascularization (CNV), Drusen, and Normal Retina. Their approach was an enhanced ensemble model which is combining multiple deep learning techniques to improve classification accuracy. However, since the model was developed and evaluated in a specific or limited dataset, its performance may not be effectively carried over to real-world clinical settings. This limitation in generalization could affect its dependability when used on diverse datasets.

Yang *et al.* [20], researchers built a CNN-based model to classify age-related macular degeneration (AMD), DME, and Normal Retina using an OCT dataset. By incorporating pre-trained ImageNet weights, they significantly improved the model's accuracy significantly from 68.17% to 92.89%. Their method utilized an ensemble of three distinct CNN models, each enhanced with pre-trained weights and fine-tuned parameters, ensuring more precise classification of retinal images into their respective categories.

Stanojević *et al.* [21], the researchers proposed a deep learning-based classification of retinal diseases using OCT images. It evaluates CNN architectures which includes AlexNet, VGG, Inception, and ResNet. The Inception1 model that trained with RMSprop optimizer, achieved the highest accuracy of 95.53%, along with an F1-score of 0.93687. The study shows that Inception-based models outperform others in accurately classifying the images. The comprehensive comparison of the machine learning based and deep learning based approaches for the OCT image classification is depicted in Table 1.

Table 1. Summary of retinal OCT image classification techniques

Author(s)	Method type	Dataset	Approaches	Drawback
Liu <i>et al.</i> [11]	Classical ML	Local	SVM with global + local features	Sensitive to feature extraction and alignment errors
Alsaih <i>et al.</i> [13]	Classical ML	Local	HOG, LBP, PCA with SVM	Dependence on manual feature engineering
Huang <i>et al.</i> [14]	Deep Learning	OCT2017, HUCM	Layer-guided CNN with transfer learning	Requires pre-segmentation of retinal layers
Kim and Tran [16]	Deep Learning	OCT2017	U-Net combined with image rotation and binary classifiers	High computational cost, complex pre-processing
Diao <i>et al.</i> [17]	Deep Learning	Not specified	CM-CNN and CAM-Unet architectures	Separate models for segmentation and classification
Hassan <i>et al.</i> [18]	Hybrid (DL + ML)	OCT2017	ResNet-50 features + Random Forest (EOCT approach)	Complex hybrid framework, higher training complexity
Paul <i>et al.</i> [19]	Ensemble DL	OCTx, OCT2017 (limited)	Multiple DL models; good but limited generalization	Limited generalization due to small dataset
Yang <i>et al.</i> [20]	CNN Ensemble	OCT2017	Accuracy boosting through ensemble learning	High memory and computational requirements
Stanojević <i>et al.</i> [21]	Deep Learning	OCT2017	Inception1 model performed best among tested models	Computationally intensive, less suited for real-time deployment

We proposed a lightweight CNN model designed to balance diagnostic accuracy with computational efficiency in retinal OCT classification. Unlike traditional methods that rely on manual feature extraction, our end-to-end trainable model learns important patterns directly from raw images, reducing human bias. The architecture uses just six layers and incorporates hybrid pooling-max pooling to capture key features and average pooling to retain spatial details-resulting in a compact yet powerful model. To improve training, we applied exponential learning rate decay, which allows faster convergence and finer tuning, unlike fixed learning rates. Data augmentation further enhances model performance by addressing class imbalance and reducing overfitting. Without relying on complex preprocessing or heavy transfer learning, our approach offers strong generalization across varied imaging conditions. This makes it well-suited for practical clinical use, especially in low-resource settings or mobile diagnostic tools, where speed, efficiency, and accuracy are essential.

3. PROPOSED WORK

The proposed approach explores data augmentation, exponential learning rate decay and CNN techniques. The use of exponential learning rate decay helps in making the learning faster in the initial stages while in the later stages helps to get near minima by taking small steps or jumps. Image expansion techniques allow a single image to be represented in various forms by modifying attributes like orientation, shear range, and zoom ratio. It helps to get greater details for the training purpose as the training dataset gets evolved by it.

The augmentation may also help to identify and differentiate some overlapping features of the images of different classes; potentially increasing the accuracy of the model. The expanded dataset is fed into the CNN for training, allowing it to capture fine-tuned image features through its convolutional layers. The final stage of the model includes a dense layer containing 1,024 neurons, which functions as a hidden unit, which is ultimately linked to 4 output neurons corresponding to the four distinct classes in the OCT2017 [15] dataset. The complete workflow of the proposed retinal OCT classification model is outlined in Algorithm 1. A detailed interpretation of each step of Algorithm 1 is explained in the subsequent sections to ensure the complete and clear understanding of our methodology.

Algorithm 1: Proposed CNN model for retinal OCT classification

Require: OCT2017 dataset with four retinal classes

Ensure: Classified disease label

Step 1: Data Preprocessing

- Resize images to 200×200 , normalize pixel values

Step 2: Data Augmentation

- Perform shear, zoom, and horizontal flip transformations

Step 3: CNN Model Training

- Initialize convolutional layers with hybrid pooling
- Use ReLU activation and dropout to prevent over fitting
- Optimize model using Adam with exponential learning rate decay

Step 4: Model Evaluation

- Test on unseen images and compute accuracy, confusion matrix
- Generate ROC curves for performance validation

Step 5: Deployment

- Save lightweight model for real-time clinical usage

3.1. Data preparation

The OCT2017 dataset [15] consists of 4 different classes: CNV, DME, Drusen and Normal. The training dataset contains 83,484 images divided such that CNV contains 37,205 images, DME has 11,348, Drusen has 8,616 and finally Normal contains 26,315 images. The images in the training dataset are having varying pixel dimensionality and also contain some noise as well. The DPI (dots per inch) of all the images is 96. On the other hand, the test dataset contains a total of 968 images, 242 images each per class. For the validation purpose, a total of 32 images, 8 each per class is given. The Figure 1 represents the sample images for each class in the dataset.

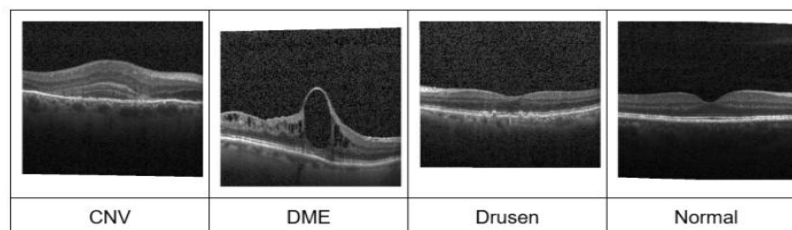


Figure 1. Sample OCT images from OCT2017 dataset

3.2. Data augmentation

Data augmentation is a technique through which new images can be generated by applying some attributes to the training dataset images. The overlapping features of different classes must be learnt by the CNN model and that can be enhanced up to some level by the data augmentation mechanism as it provides more information in different ways. Therefore, the data augmentation is applied to all the images into the training dataset. As mentioned by Dave and Pandya [22], data augmentation can lead to improved accuracy of the model. The attributes which are used for data augmentation in the proposed approach are described in Table 2.

Medical images often have class imbalances and limited variation in orientation, brightness, or structure-factors critical for accurate classification. Techniques like shearing, zooming, and flipping introduce synthetic diversity, helping the model recognize key features, especially in underrepresented classes like Drusen. This reduces overfitting and boosts the model's robustness when applied to real-world clinical data. Augmentation ensures the model learns from a broader range of examples, improving its reliability and diagnostic performance.

Table 2. Data augmentation attributes

Attribute name	Attribute value
rescale	1./255
shear range	0.3
zoom range	0.3
horizontal flip	True
Target Size (px)	200x200

3.3. Exponential learning rate decay

Learning rate is a crucial hyper-parameter in deep learning or machine learning. It handles how fast or gradual learning happens while training. In mathematical terms it derives the size of the jump on the error surface to be taken to reach up to the minima. There can be a trade off if we put the learning rate so high; leading to fast learning in the initial stages while greater oscillation in later on stages generating chances to miss the actual minima. On the other hand, if we put the learning rate very small then very gradual steps would be carried out towards minima on the error surface; increasing the time of learning at a greater space.

$$\text{New Learning Rate} = \text{Initial Learning Rate} \times (\text{Decay Rate})^{(\text{step} / \text{decay steps})} \quad (1)$$

To combat this situation, the proposed work has used exponential learning rate decay which uses larger jumps in the initial set of training and then gradually decreasing the size of the jump towards the actual minima, making very stable updates in later stages of learning. The mathematical interpolation of the same is given by (1). This implementation ultimately helps in faster learning in initial stages and avoids overshooting and oscillations around actual minima of the error surface. Table 3 lists the parameters used for the exponential decrease in the learning rate. Based on these values and (1), Table 4 shows how the learning rate changes step by step during training.

Table 3. Parameter values for exponential learning rate decay

Parameter name	Parameter value
initial learning rate	0.001
decay steps	10000
decay rate	0.9

Table 4. Updates in learning rate using parameters of Table 3

Steps	Equation	New Value of LR
10000	LR=0.001x(0.9) ¹	0.0009
20000	LR=0.001x(0.9) ²	0.00081
30000	LR=0.001x(0.9) ³	0.000729

3.4. Convolutional neural network structure

The structure of the CNN model used in our approach is described in Figure 2. The model comprises the combination of max and average pooling by which robust features of the images can be extracted. The first convolutional layer applies 32 filters of 3×3 size on the original input image of size 200×200 pixels with a rectified linear unit (ReLU) [23] as an activation function. The first layer uses the max pooling of size 2×2 to grab the maximum pixel value and to reduce the feature map size. The second layer consists of 16 filters of size 3×3 and average pooling of size 2×2. Average pooling gathers the average pixel value of its surrounding to not miss detailed information. The essence of maximum and average pooling is represented in Figure 2. The second, third, fourth, and fifth layers contain 32, 64, 64, and 128 kernels, respectively, each with a size of 3×3. These are followed by pooling operations in the order: average (2×2), max (2×2), max (2×2), and average (2×2).

$$\text{MaxPool}(x) = \max(x_{ij})_{1 \leq i \leq k, 1 \leq j \leq k} \quad (2)$$

$$\text{AvgPool}(x) = \frac{1}{k^2} \sum_{i=1}^k \sum_{j=1}^k x_{ij} \quad (3)$$

The mathematical interpolation for the maximum and average pooling is described by (2) and (3), respectively, where x_{ij} represents the values of the input feature map within the pooling window and K is the size of the pooling window. Hybrid pooling balances specificity (by max pooling) and robustness (by average pooling), ensuring that the model captures both discriminative features and structural coherence. In the proposed CNN model, a combination of hybrid pooling is utilized to enhance performance. Mathematically, hybrid pooling combines max-pooling and average-pooling outputs as (4), where $\alpha \in [0,1]$ controls the trade-off between feature sharpness and smoothing. This mechanism preserves salient features while maintaining global context, offering a more balanced feature representation than individual pooling techniques.

$$\text{HybridPool}(x) = \alpha \times \text{MaxPool}(x) + (1 - \alpha) \times \text{AvgPool}(x) \quad (4)$$

The CNN structure utilizes the ReLU [23] activation function across all feature extraction layers. ReLU [23] has the properties of not being saturated over the input data points and is activated on a positive set of inputs or neurons. In addition to these properties, ReLU [23] also helps solve the problem of vanishing gradients at large. The function can be described by (5), where x denotes input data to the function. Research and results have shown that ReLU generates faster output in large and complex networks.

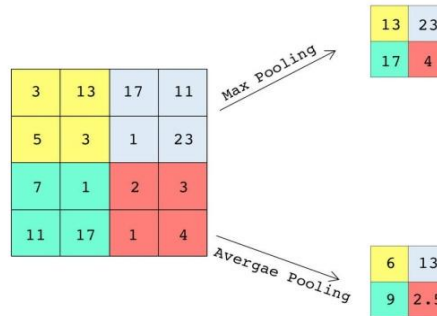


Figure 2. Illustration of 2x2 max and average pooling

The second, third, fourth, and fifth layers contain 32, 64, 64, and 128 kernels, respectively, each with a size of 3x3. These are followed by pooling operations in the order: average (2x2), max (2x2), max (2x2), and average (2x2). The CNN structure utilizes the ReLU [22] activation function across all feature extraction layers. Relu [22] has the properties of not being saturated over the input data points and is activated on a positive set of inputs or neurons. In addition to these properties, Relu [22] also helps solve the problem of vanishing gradients at large. The function can be described by (2), where x denotes input data to the function. Research and results have shown that Relu generates faster output in large and complex networks.

$$F(x) = \max(x, 0) \quad (5)$$

The last pooling layer in the structure is an average pooling of 2x2 on 128 filters of the 3x3 convolutional layer. Hence, performing the flattening of the features at this stage would have $1 \times 1 \times 128$ features which are connected to the 1024 dense neurons. This dense unit acts as a hidden unit placed just before the final output stage. To help prevent overfitting and improve the model's ability to generalize, a dropout rate of 0.3 is applied. Finally, the output layer includes 4 neurons, each corresponding to one of the four classes in the dataset. To classify the images into proper classes, the last layer uses the softmax function, which works on probability distribution. Categorical-cross-entropy is used with Adam optimizer and exponential learning rate decay to illustrate multiclass classification. The proposed CNN architecture can be visualized by Figure 3.

$$\sigma(z)_j = \frac{e^{(z)_j}}{\sum_{k=1}^K e^{(z)_k}} \quad (6)$$

The softmax function is mathematically represented by (6), with its graphical interpretation shown in Figure 4. This function scales the output values between 0 and 1, ensuring that the total sum of all class probabilities equals 1. In (6), Z represents the input vector to the output layer, and the index j refers to each output unit, where $j = 1, 2, \dots, K$.

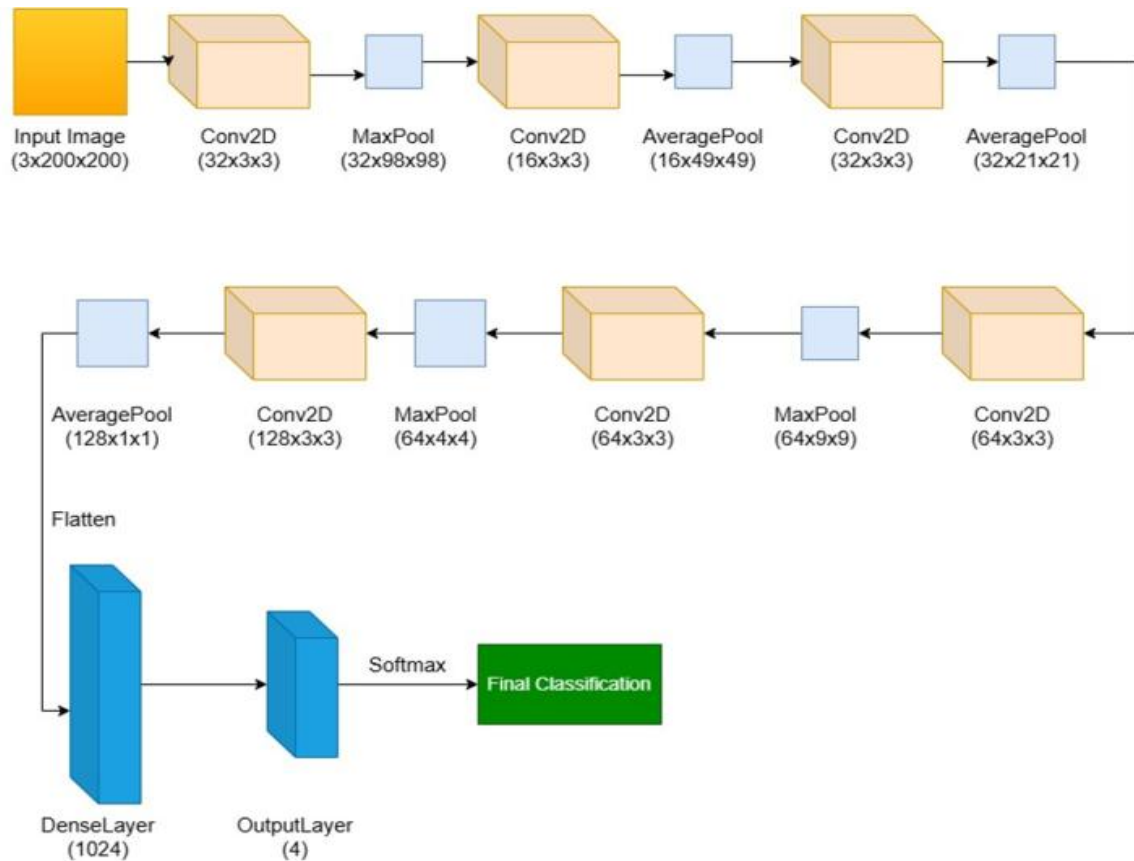


Figure 3. Proposed CNN architecture

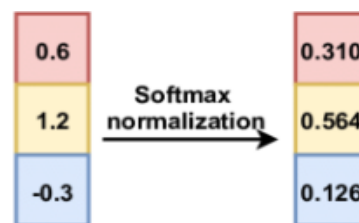


Figure 4. Softmax output as class probability distribution

4. EXPERIMENTAL SETUP AND RESULTS

The implementation was carried out on a local system equipped with an AMD Ryzen 7 5700G 64-bit processor, Radeon Graphics at 3.80 GHz, and 16 GB of RAM. The proposed approach was developed in Python, utilizing libraries such as Keras, TensorFlow, scikit-learn, and Matplotlib. The approach uses callback as the early stopping mechanism for the CNN training based on validation update loss values and a patience value kept at 5. We store the best weights for the model that has been trained over 11 epochs. As mentioned in the previous section, a total of 83,484 input instances were used for training, 32 instances served as the validation set, and 968 instances were allocated for testing.

4.1. Evaluation of proposed approach

The proposed deep neural network architecture was carefully trained and tested using the OCT2017 image dataset to ensure a reliable and comprehensive evaluation of its performance. Training was conducted over 11 epochs, utilizing an early stopping mechanism to prevent overfitting and optimize generalization. The entire training process took approximately 170 minutes, while testing required just 121 milliseconds, demonstrating the efficiency of our lightweight approach. By incorporating data augmentation and an

exponential learning rate decay strategy, the model successfully learned intricate retinal patterns while maintaining stability during training.

The training and validation errors changed over 11 epochs for the proposed CNN model is depicted by Figure 5. As the figure illustrates, the training error consistently goes down, which suggests that the model is learning well from the data. This smooth and steady trend highlights the effectiveness of using exponential learning rate decay and data augmentation, both of which helped the model train efficiently and reliably. The close match between the training and validation curves shows that the model performs well even on unseen data, demonstrating its overall robustness. The following equations represent common metrics such as precision, recall and F1 score, used in evaluation of classification models:

$$\text{Precision} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Positives}} \quad (7)$$

$$\text{Recall} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}} \quad (8)$$

$$\text{F1 Score} = 2 * \frac{\text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}} \quad (9)$$

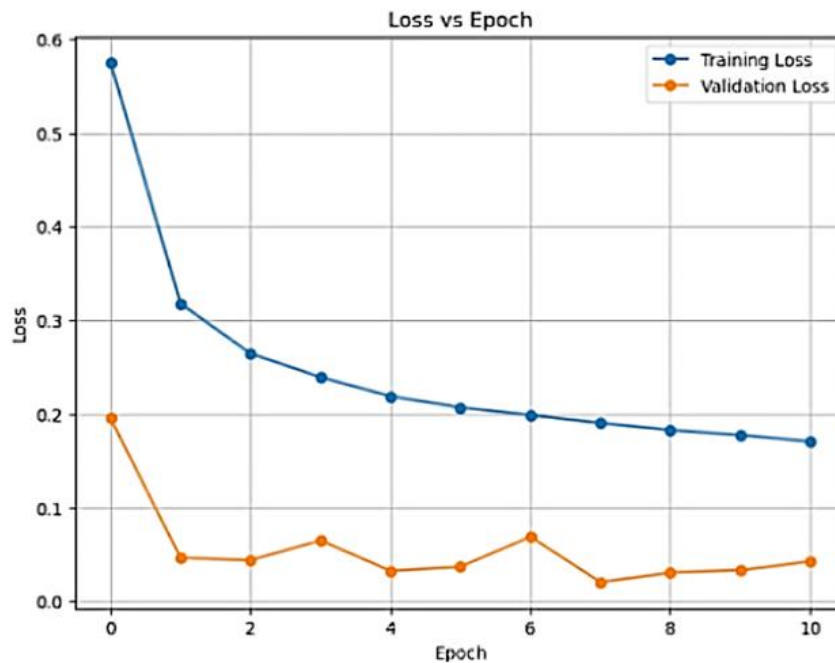


Figure 5. Model error progression across epochs for both training and validation sets

The confusion matrix shown in Figure 6 further illustrates the precision of the model, revealing minimal misclassifications in the four categories of the retina: CNV, DME, Drusen, and Normal. The proficiency of the model in classifying Drusen and DME, two conditions that often share overlapping visual features, is particularly notable. The high precision, recall, and F1 score values depicted in Table 5 reaffirm the reliability of the model. In fact, it achieved an F1 score of 1.00 for the Normal and Drusen classes, indicating flawless classification in those cases.

Despite class imbalance, especially with Drusen having fewer samples, augmentation techniques were applied uniformly to all classes, increasing sample diversity. Each sample, (x, y) , was transformed into $(T(x), y)$, where T represents random transformations like shearing and flipping. This led to improved Drusen class performance with an F1-score of 0.99 and a recall of 0.98, highlighting the augmentation's role in addressing imbalance and enhancing model generalization. These figures and tables correspond to the numerical labels 0, 1, 2, and 3, representing the four classes: CNV, DME, Drusen, and Normal, respectively.

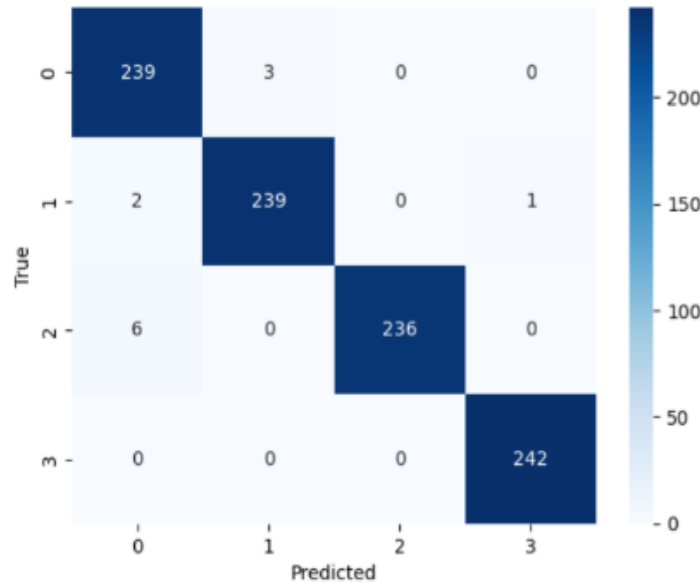


Figure 6. Confusion matrix evaluation on proposed model for the test dataset

Table 5. Performance metrics of the proposed model on test dataset

Class	Precision	Recall	F1-score	Support
CNV	0.97	0.99	0.98	242
DME	0.99	0.99	0.99	242
Drusen	1.00	0.98	0.99	242
Normal	1.00	1.00	1.00	242
Accuracy			0.99	968
Macro Avg	0.99	0.99	0.99	968
Weighted Avg	0.99	0.99	0.99	968

To further validate its robustness, the receiver operating characteristic (ROC) curve of the model shown in Figure 7 was plotted, revealing an AUC close to 1, which indicates strong predictive capacity and confidence in its classifications. Compared to previous methods, including Layer-Guided CNN, Binary CNN, and EOCT models, our approach consistently outperformed them in terms of accuracy. The comparison of the existing approaches with our proposed approach is shown in Table 6.

The comparative overview of four progressively refined CNN architectures developed for retinal OCT image classification is depicted in Table 7. The initial model with same number of convolutional layers depicted by Figure 4 (Model A) uses max pooling after each convolutional layer, a fixed learning rate of 0.001, and was trained for 11 epochs. It gained an accuracy of 94.3%, but struggled to accurately classify minority classes such as Drusen. In Model B, the application of data augmentation significantly enhanced the model's ability to generalize across various samples that improving overall class balance and increasing the accuracy to 96.12%. Model C introduced hybrid pooling as the proposed model (a combination of Max and Average pooling), which allowed for more effective retention of spatial and textural details, thereby gaining the accuracy to 97.3%. Building on these improvements, the proposed Model D brought together all previous enhancements and introduced an exponential learning rate decay, which helped the model learn more steadily and converge more smoothly during training and finally achieving the benchmark accuracy of 98.75%.

4.2. Explainability analysis using CAM

To get a clearer picture of how our CNN “works”, CAM [9] is imposed as an explainable AI tool (XAI) to peek inside its layers. CAM turns each convolutional layer's feature maps into heatmaps, showing exactly which parts of an OCT image, the model is focusing on at that depth. While comparing CAMs output from the first, middle, and deepest layers, a smooth progression is seen: early on the network spots simple edges and textures, then it homes in on the indicative signs of disease-like fluid pockets or structural breaks, before making its final call.

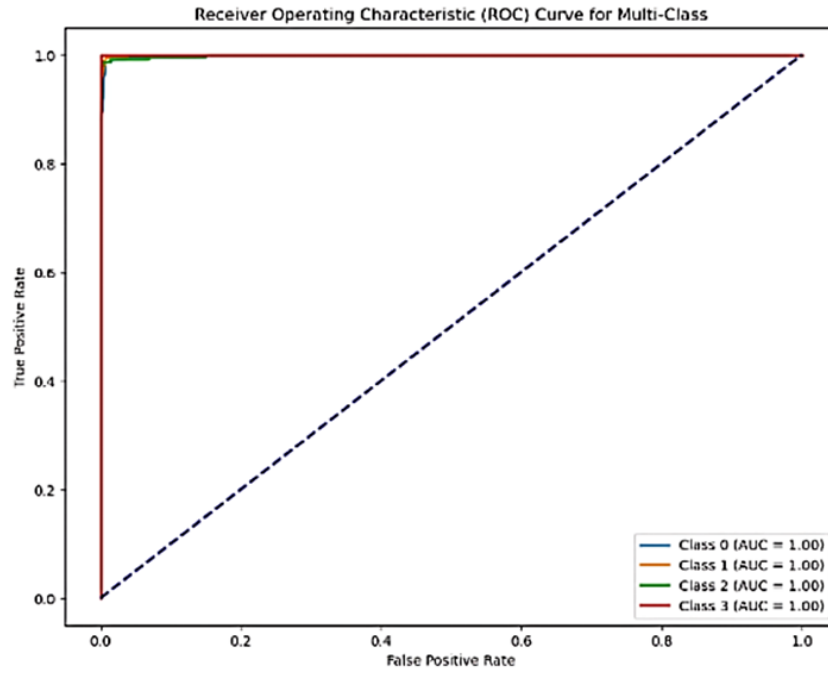


Figure 7. ROC curve for the proposed approach

Table 6. Comparison with existing approaches

Year	References	Method	Dataset	Acc
2019	Diaz <i>et al.</i> [24]	CNN	OCT2017	93
2019	Huang <i>et al.</i> [14]	LayerGuided CNN	OCT2017	88.4
2020	Saraiva <i>et al.</i> [25]	CNN	OCT2017	93.3
2021	Kim and Tran [16]	Binary CNN Model 1 / Model 2 (Heavy models)	OCT2017	98.1/98.7
2023	Hassan <i>et al.</i> [18]	EOCT Model	OCT2017	97.47
2023	Diao <i>et al.</i> [17]	CNN	OCT2017	96.93
2023	Opoku <i>et al.</i> [26]	Capsule network with contrast limited adaptive histogram equalization	OCT2017	97.7
2024	Stanojević <i>et al.</i> [21]	Deep CNN	OCT2017	95.55
2024	Yang <i>et al.</i> [20]	Ensemble Model based on CNN, Efficientnet v2 and Resnet	OCT2017	97.89
2025	Proposed approach	CNN + Data Augmentation + Exponential Learning Rate Decay + mixture of max and average pooling	OCT2017	98.75

Table 7. Ablation study of the proposed approach

Model	Data augmentation	Pooling	Learning rate	Acc	Observation
Model A (Baseline CNN)	NO	Max Pooling	Fixed LR	94.3%	Simple CNN with standard pooling and fixed LR shows limited learning, especially on minority classes like Drusen
Model B	Yes	Max Pooling	Fixed LR	96.12%	Augmentation improves generalization and class balance, especially for underrepresented classes.
Model C	Yes	Hybrid Pooling (Max + Avg)	Fixed LR	97.3%	Addition of hybrid pooling boosts performance by better preserving spatial and edge details.
Model D (Proposed Model)	Yes	Hybrid Pooling (Max + Avg)	Exponential LR	98.75%	Final model; exponential LR decay stabilizes learning, leading to optimal convergence and best accuracy.

CAM provides visual insights into the regions of the retinal OCT images that most significantly contribute to the classification decisions made by the CNN model. This visualization helps ensure that the model is focusing on clinically meaningful regions, thus promoting transparency, trust, and potential clinical applicability. Mathematically, for a given class c , the CAM heatmap $M_c(x, y)$ is defined by (10).

$$M_c(x, y) = \sum_k w_k^c f_k(x, y) \quad (10)$$

Where w_k^c is the weight associated with the k -th feature map for class c , and $f_k(x, y)$ denotes the activation at spatial location (x, y) in that feature map.

The layer-wise class activation map (CAM) visualizations obtained from the proposed CNN model are presented Figure 8. Specifically, Figure 8(a) displays the layer-wise CAM for CNV, Figure 8(b) for DME, Figure 8(c) for DRUSEN, and Figure 8(d) for NORMAL. In this figure, the first-layer CAM representation starts from the left-hand side and proceeds consecutively to the last layer on the right-hand side for each subfigure. It is observed that as the depth of the network increases, the model's focus shifts from broad structural patterns to specific clinical features, confirming the hierarchical learning capability of the CNN. The results of CAM strongly affirm that the proposed CNN model not only achieves high classification accuracy but also learns clinically relevant features in a hierarchical manner, thus enhancing the transparency, reliability, and readiness of the model for real-world clinical deployment.

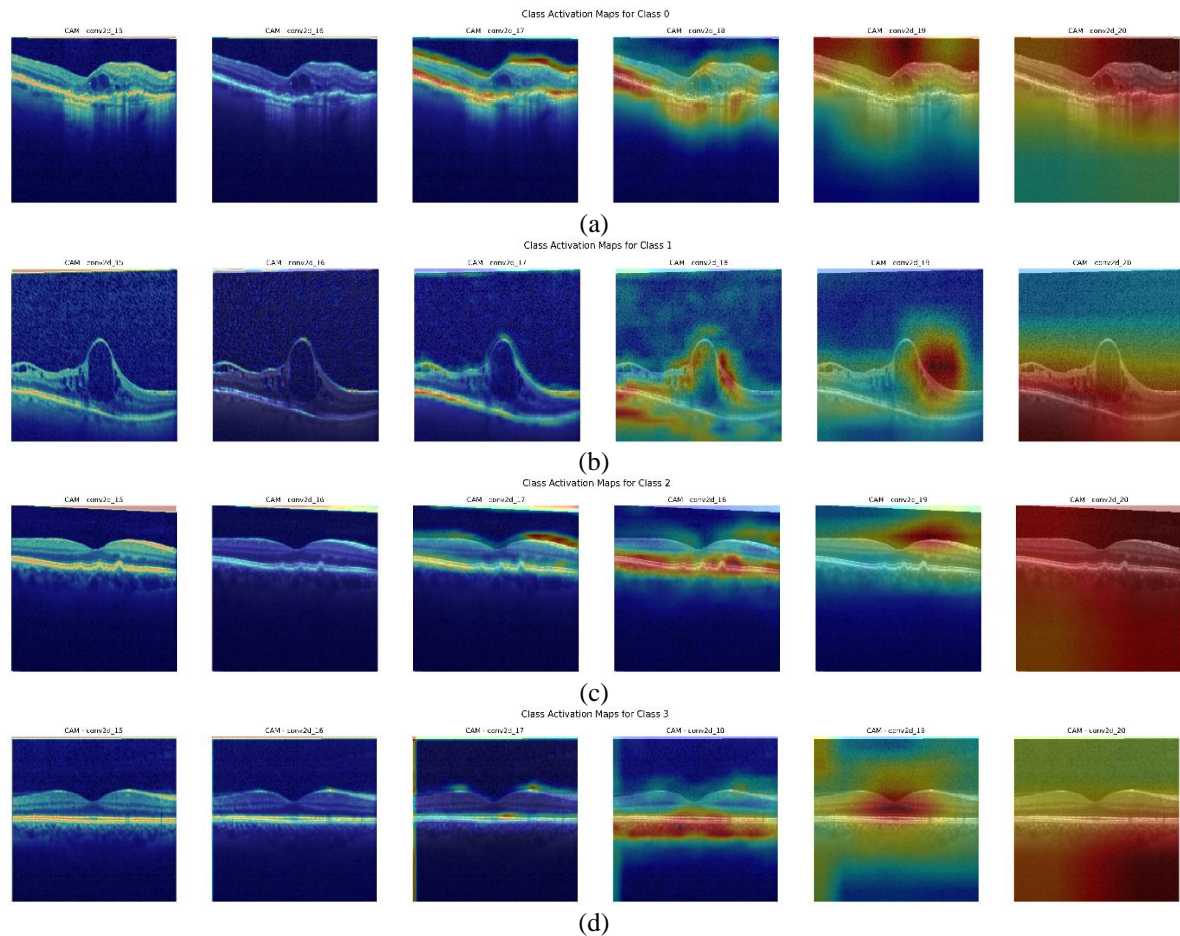


Figure 8. Layer-wise (left to right) class activation maps (CAMs) from the proposed CNN model: (a) layer-wise CAM for CNV, (b) layer-wise CAM for DME, (c) layer-wise CAM for DRUSEN, and (d) layer-wise CAM for NORMAL

4.3. Efficiency and deployment suitability of the proposed model

A significant advantage of this model, beyond its high accuracy, is that it contains six layers with only 275,636 trainable parameters, which makes it highly efficient with minimal memory requirements that ensure faster inference times. The comparative analysis of state of the art approaches with approximate number of trainable parameters used in various deep learning models applied to the OCT2017 dataset is stated in Table 8.

The proposed lightweight architecture requires only 1.05 MB of memory, making it highly efficient and suitable for real-time clinical use, especially in settings with limited processing power and components. Unlike traditional CNN models, which demand high processing power and extensive memory, our model offers a fast, efficient, and scalable solution for automated retinal disease diagnosis. Table 9 summarizes the potential platforms and devices suitable for deploying the proposed lightweight CNN model. The table highlights the flexibility of the model for mobile, embedded, desktop, and web-based clinical applications.

Table 8. Comparison of trainable parameters with proposed model

Year	References	Trainable parameters (Millions-Approximately)
2019	Huang <i>et al.</i> [14]	24.88
2020	Saraiva <i>et al.</i> [25]	0.95
2021	Kim and Tran [16]	Model 1:- 235.6 / Model 2:- 452.1
2023	Hassan <i>et al.</i> [18]	25.64
2023	Opoku <i>et al.</i> [26]	8.2
2024	Stanojević, <i>et al.</i> [21]	6.6
2024	Yang <i>et al.</i> [20]	21.4
2025	Proposed approach	0.28 (275636)

Table 9. Potential deployment platforms and devices for the proposed lightweight CNN model

Platform	Device examples	Notes
Mobile Phones	Android phones, iPhones	Requires TFLite/CoreML conversion for deployment on mobile devices.
Edge Devices	Raspberry Pi 4, NVIDIA Jetson Nano	Ideal for offline deployment in remote clinics with limited resources.
Desktop Computers	Basic Windows/Linux machines with minimal GPU	High-speed inference possible due to small model size and low computational load.
Web Browsers	Chrome, Firefox, Safari	Can be deployed using TensorFlow.js for easy access via browsers.
Tablets	Android Tablets, iPads	Suitable for portable, mobile health screening applications.

5. COLCLUSION AND FUTURE WORK

Our proposed work introduces a lightweight deep learning model for the automated classification of retinal OCT images for four classes of OCT2017 dataset: CNV, DME, Drusen, and Normal. Leveraging CNNs, data augmentation, and exponential learning rate decay, the model achieved an outstanding accuracy of 98.75% which surpasses numerous existing benchmark models. F1-score for CNV, DME, Drusen and Normal are 0.98, 0.99, 0.99 and 1.00 respectively. The proposed model integrates a hybrid pooling mechanism, using max and average pooling layers that ensure robust feature extraction while preserving essential spatial information. In addition, the integration of synthetic data generation techniques significantly improved the generalization of the model in a wide range of retinal images. A six-layer CNN architecture having 275,636 trainable parameters makes it highly efficient with a minimal memory requirement, which ensures faster inference times. Despite the remarkable performance of the proposed approach, there are several areas where further improvements can be carried out such as more detailed explainable AI (XAI) for clinical adoption, finding more anomalies than just focusing on only four classes, clinical validation with diverse datasets, scalability and transferability of the implementation.

ACKNOWLEDGEMENTS

The authors sincerely appreciate the support provided by Department of Computer Engineering and L. D. College of Engineering -Ahmedabad.

FUNDING INFORMATION

The authors state no funding involved.

AUTHOR CONTRIBUTIONS STATEMENT

This journal uses the Contributor Roles Taxonomy (CRediT) to recognize individual author contributions, reduce authorship disputes, and facilitate collaboration.

Name of Author	C	M	So	Va	Fo	I	R	D	O	E	Vi	Su	P	Fu
Parth R. Dave	✓	✓	✓	✓	✓	✓	✓	✓	✓		✓		✓	
Nikunj H. Domadiya	✓	✓		✓	✓		✓			✓	✓	✓		

C : Conceptualization

I : Investigation

Vi : Visualization

M : Methodology

R : Resources

Su : Supervision

So : Software

D : Data Curation

P : Project administration

Va : Validation

O : Writing - Original Draft

Fu : Funding acquisition

Fo : Formal analysis

E : Writing - Review & Editing

CONFLICT OF INTEREST STATEMENT

Authors state no conflict of interest.

DATA AVAILABILITY

Data availability does not apply to this paper as no new data were created or analyzed in this study.




REFERENCES

- [1] T. Ojala, M. Pietikäinen, and T. Mäenpää, "Multiresolution gray-scale and rotation invariant texture classification with local binary patterns," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, no. 7, pp. 971–987, 2002, doi: 10.1109/TPAMI.2002.1017623.
- [2] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, vol. 1, pp. 886–893, 2005, doi: 10.1109/CVPR.2005.177.
- [3] R. M. Haralick, K. Shanmugam, and I. H. Dinstein, "Textural features for image classification," *IEEE Transactions on Systems, Man, and Cybernetics*, no. 6, pp. 610–621, 1973, doi: 10.1109/TSMC.1973.4309314.
- [4] M. A. Hearst, S. T. Dumais, E. Osuna, J. Platt, and B. Schölkopf, "Support vector machines," *IEEE Intelligent Systems and Their Applications*, vol. 13, no. 4, pp. 18–28, 1998, doi: 10.1109/5254.708428.
- [5] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," *Advances in Neural Information Processing Systems*, vol. 25, 2012.
- [6] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.
- [7] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778, 2016, doi: 10.1109/CVPR.2016.90.
- [8] C. Szegedy et al., "Going deeper with convolutions," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1–9, 2015, doi: 10.1109/CVPR.2015.7298594.
- [9] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, "Learning deep features for discriminative localization," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2921–2929, 2016, doi: 10.1109/CVPR.2016.319.
- [10] P. P. Srinivasan et al., "Fully automated detection of diabetic macular edema and dry age-related macular degeneration from optical coherence tomography images," *Biomedical Optics Express*, vol. 5, no. 10, pp. 3568–3577, 2014, doi: 10.1364/BOE.5.003568.
- [11] Y.-Y. Liu et al., "Computerized macular pathology diagnosis in spectral domain optical coherence tomography scans based on multiscale texture and shape features," *Investigative Ophthalmology and Visual Science*, vol. 52, no. 11, pp. 8316–8322, 2011, doi: 10.1167/iovs.10-7012.
- [12] J. Wu and J. M. Rehg, "Where am I: Place instance and category recognition using spatial pact," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1–8, 2008, doi: 10.1109/CVPR.2008.4587627.
- [13] K. Alsaih et al., "Classification of SD-OCT volumes with multi-pyramids, LBP and HOG descriptors: Application to DME detections," in *Proceedings of the 38th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pp. 1344–1347, 2016, doi: 10.1109/EMBC.2016.7590956.
- [14] L. Huang, X. He, L. Fang, H. Rabbani, and X. Chen, "Automatic classification of retinal optical coherence tomography images with layer-guided convolutional neural network," *IEEE Signal Processing Letters*, vol. 26, no. 7, pp. 1026–1030, 2019, doi: 10.1109/LSP.2019.2917779.
- [15] D. S. Kermany et al., "Identifying medical diagnoses and treatable diseases by image-based deep learning," *Cell*, vol. 172, no. 5, pp. 1122–1131, 2018.
- [16] J. Kim and L. Tran, "Retinal disease classification from OCT images using deep learning algorithms," in *Proceedings of the IEEE Conference on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB)*, pp. 1–6, 2021, doi: 10.1109/CIBCB49929.2021.9562919.
- [17] S. Diao et al., "Classification and segmentation of OCT images for age-related macular degeneration based on dual guidance networks," *Biomedical Signal Processing and Control*, vol. 84, p. 104810, 2023, doi: 10.1016/j.bspc.2023.104810.
- [18] E. Hassan et al., "Enhanced deep learning model for classification of retinal optical coherence tomography images," *Sensors*, vol. 23, no. 12, p. 5393, 2023, doi: 10.3390/s23125393.
- [19] D. Paul, A. Tewari, S. Ghosh, and K. Santosh, "OCTX: Ensembled deep learning model to detect retinal disorders," in *Proceedings of the IEEE 33rd International Symposium on Computer-Based Medical Systems (CBMS)*, pp. 526–531, 2020, doi: 10.1109/CBMS49503.2020.00105.
- [20] J. Yang, G. Wang, X. Xiao, M. Bao, and G. Tian, "Explainable ensemble learning method for OCT detection with transfer learning," *PLOS One*, vol. 19, no. 3, p. e0296175, 2024, doi: 10.1371/journal.pone.0296175.
- [21] M. Stanojević, D. Drašković, and B. Nikolić, "Retinal disease classification based on optical coherence tomography images using convolutional neural networks," *Journal of Electronic Imaging*, vol. 32, no. 3, pp. 032004–032004, 2023, doi: 10.1117/1.JEL.32.3.032004.




- [22] P. R. Dave and H. A. Pandya, "Satellite image classification with data augmentation and convolutional neural network," in *Advances in Electrical and Computer Technologies: Select Proceedings of ICAECT 2019*, Springer, pp. 83–92, 2020, doi: 10.1007/978-981-15-5558-9_9.
- [23] V. Nair and G. E. Hinton, "Rectified linear units improve restricted Boltzmann machines," in *Proceedings of the 27th International Conference on Machine Learning (ICML-10)*, pp. 807–814, 2010.
- [24] M. Díaz, J. Novo, P. Cutrín, F. Gómez-Ulla, M. G. Penedo, and M. Ortega, "Automatic segmentation of the foveal avascular zone in ophthalmological OCT-A images," *PLOS One*, vol. 14, no. 2, p. e0212364, 2019, doi: 10.1371/journal.pone.0212364.
- [25] A. A. Saraiva et al., "Classification of optical coherence tomography using convolutional neural networks," in *Proceedings of the 13th International Joint Conference on Biomedical Engineering Systems and Technologies (BIOSTEC 2020)*, pp. 168–175, 2020, doi: 10.5220/0009091001680175.
- [26] M. Opoku, B. A. Weyori, A. F. Adekoya, and K. Adu, "CLAHE-CapsNet: Efficient retina optical coherence tomography classification using capsule networks with contrast limited adaptive histogram equalization," *PLOS One*, vol. 18, no. 11, p. e0288663, 2023, doi: 10.1371/journal.pone.0288663.

BIOGRAPHIES OF AUTHORS



Prof. Parth R. Dave    was born at Halol in Gujarat, India on 7th August, 1989. Prof. Dave completed B.E. and M. Tech with specialization in Computer Engineering from Dharmsinh Desai University, Nadiad, Gujarat, India in the years 2011 and 2013 respectively. The author's major field of study covers machine learning, deep learning, image processing and multimedia information retrieval. He had worked as an Assistant Professor in the Computer Engineering Department at Dharmsinh Desai University, Nadiad, Gujarat, India from 2013 to 2020. He is currently working as an Assistant Professor in the Computer Engineering Department at L. D. College of Engineering, Ahmedabad, Gujarat, India since 2020. He delivers expert talks and serves as reviewer in international conferences and journals. He has also presented papers in International Journals, IEEE Conferences. He can be contacted at email: prd7889@gmail.com.



Nikunj H. Domadiya    completed M. Tech and Ph.D. from NIT-Surat, Gujarat, India in the years 2013 and 2020 respectively. The author's major field of research covers Data Privacy and Security, Machine learning, Deep learning, and multimedia information retrieval. He is currently working as an Assistant Professor in the Computer Engineering Department at L. D. College of Engineering, Ahmedabad, Gujarat, India since 2020. He can be contacted at email: domadiyanikunj002@gmail.com.