International Journal of Computing and Digital Systems
2025, VOL. 17, NO. 1, 1–12

http://dx.doi.org/10.12785/ijcds/1571024157

New Ensemble Model for Diagnosing Retinal Diseases from Optical Coherence Tomography Images

Shibly Hameed Al-Amiry[1] **and Ali Mohsin Al-juboori**[2]

[1, 2]*Department of Computer Science, University of Al-Qadisiyah, Diwaniyah, Iraq*

Received 24 April 2024, Revised 15 September 2024, Accepted 19 September 2024

Abstract: The vision depends greatly on the retina. Unfortunately, it may be exposed to many diseases that lead to poor vision or blindness. This research aims to diagnose retinal diseases through OCT images, focusing on Drusen, diabetic macular edema (DME), and choroidal neovascularization (CNV). A new ensemble approach is proposed that combines a specialization approach, hard voting method and highest probability method. It is based on three sub-models (Custom-model, Xception, and MobileNet). Because we noticed that some sub-models are better than others at classifying a particular category, each sub-model was specialized to the category it classifies best. If the specialized sub-model does not exist, the final classification will be based on the category with the highest votes, and if it does not exist, the category with the highest probability will be chosen. We also used a way to correct final misclassification through a list of negative predictions created to contain categories to which the sub-model is somewhat certain that an image does not belong. The proposed ensemble model achieved state-of-the-art accuracies of (100%, 96.03%, and 95.85%, respectively, on the splits (original split, 80:20, 70:30) of the UCSD-v2 dataset. The Duke and OCTID datasets were also employed to verify the performance efficiency of the model, with the ensemble model achieving accuracies of 100% and 95.73%, respectively. The ensemble model outperformed all sub-models and the results of previous studies. The results of this research emphasize the effectiveness of ensemble learning techniques in analyzing medical images, especially in diagnosing retinal diseases. Therefore, this research can help in the correct diagnosis and rapid referral of patients.

Keywords: Ensemble Learning, Deep Learning, OCT Images, Retinal Diseases, Drusen, DME, CNV

1. INTRODUCTION

The retina is an important component of the human eye due to its location near the optic nerve and its sensitivity to light. Its role is to transform light into neural signals, a fundamental process for sight [1]. The macula is an extremely important part of the retina, as it is responsible for central vision and detects the color and intensity of light [2]. The retina processes light and sends it to the brain via the optic nerve, enabling vision [3]. Several retinal diseases can weaken the macula, posing major health concerns that often develop over time, including CNV, DME, and Drusen [2] (as shown in Figure 1).

The objectives of this research are building a custom CNN model and using well-known DL models as sub-models, developing a new ensemble model characterized by classification accuracy and maximum utilization of sub-models, and introducing a new mechanism to correct misclassifications of sub-models.

The key contributions of this study are enumerated as follows:

- It can be observed models are better than each other in classifying certain categories. Hence, this study adopts a specialized strategy: when every model achieves higher accuracy for a specific category, being solely responsible for its classification.
- proposing a novel mechanism to correct misclassification. This ensures that when a model is dedicated to a specific category, the contributions of other models are not ignored. Rather, they help supplement the negative prediction list (NP list) of categories, as these models somewhat confidently a given image does not belong to the categories in this list.
- Achieving optimal accuracy: This study presents a new level of accuracy, reaching 100% for the first time in the UCSD-v2 dataset.
- This study introduces a novel approach within the ensemble learning framework, underscoring the significance of this approach and the need to highlight it further to maximize the utilization of multiple models.

E-mail address: shibly.hameed@qu.edu.iq, ali.mohsin@qu.edu.iq

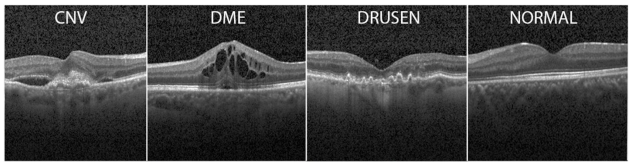

Figure 1. OCT Retina Diseases Images

Optical coherence tomography (OCT), since its introduction in 1991, has revolutionized ophthalmology because it is a non-invasive way to perform a detailed examination of the retina and choroid [4]. High-resolution OCT imaging is pivotal in diagnosing various retinal diseases [2]. It is essential for detecting and assessing macular lesions within the retina's layered structure, offering sensitive and quantitative analysis [5]. OCT effectively identifies early-stage cystic and sub-retinal swelling, often undetectable in standard retinal fundus photographs [4].

The introduction of Deep Learning (DL) techniques, especially Convolutional Neural Networks (CNNs), has initiated a new era in healthcare, revolutionizing medical diagnostics with precise and rapid decision-making [1]. In ophthalmology, these technologies have been particularly impactful, transforming automated diagnosis systems with their robust algorithms for fast and accurate disease classification [6]. The use of CNNs for retinal OCT image processing has been extensively explored, enabling these models to learn hierarchical abstract features from large datasets [2]. Research has focused on applications such as the segmentation of retinal layers [7] and the classification of OCT images [8], with some studies utilizing ensemble models for enhanced performance [9]. CNN models are preferred in many scenarios for their accuracy and efficiency in processing complex image data.

The motivation for this research is driven by the alarming statistics on retinal diseases impacting millions worldwide annually [6]: over 2.2 billion people worldwide suffer from eye illnesses, leading to significant visual impairment and, in extreme cases, complete blindness [10], with approximately 2 million CNV cases [11], 7.5 million DME cases in those over 40 [12], and more than 7 million Drusen cases annually in the USA [10].

This paper is structured as follows: Section 2 related works. Section 3 elaborates on the detailed methodologies used to build our ensemble framework, the proposed Custom sub-model, and the preprocessing steps used. Experimental results, showing the unprecedented accuracy levels achieved by our approach, are discussed in Section 4. Finally, the conclusions in Section 5.

2. RELATED WORKS

In 2020, D. Paul et al. [3] Obtained refined and high-quality images in the pre-processing, and they developed a novel framework called OCTx, that utilized an ensemble of four models: VGG16, InceptionV3, DenseNet, and a custom model. This ensemble approach effectively addressed overfitting and achieved 98.53% accuracy on the UCSD-V2 dataset. However, the study used a large number of epochs (250).

Also, in 2020, M. Berrimi and A. Moussaoui [13] proposed a new DL classification framework with transfer learning (TL), comparing a custom CNN architecture against pre-trained models like Inception-V3 and VGG-16. Using the UCSD-V2 dataset over 15 epochs, their custom CNN achieved 98.5% accuracy, while Inception-V3 reached 99.27%. Enhancements to the VGG-16 model, including additional convolution layers and regularization, increased its accuracy from 53% to 93.5%. This study did not balance the dataset, and image enhancement and noise removal techniques were absent.

In 2021, H. A. Nugroho and R. Nurfauzi [14] utilized several models (MnasNet0.5, Inception-V3, SuffleNet-v2, ResNet18, ResNet50, GoogleNet, MobileNet-v2, and DenseNet121) to diagnose retinal diseases in OCT images. MobileNet-V2 emerged as the most effective, with an accuracy of 99.64% on the UCSD-V2 dataset. However, the study did not address the dataset's imbalance.

Also, in 2021, P. Barua et al. [15] proposed a new framework for classifying retinal diseases using OCT images. The framework is based on generating deep multi-level features using 18 convolutional neural networks. The final features are selected from the best five neural networks, and classification is done using the Support Vector Machine (SVM) classifier. The framework was tested on the Duke dataset, with a split of 90% training and 10% testing, to achieve 100% accuracy. However, the dataset is incomplete; only 3194 images were used.

Moreover, in 2021, A. Singh et al. [16] used the DL model to diagnose diseases in OCT images from the OCTID dataset. It achieved an accuracy of 88.5%, and by removing samples with a high degree of uncertainty and referring them to human experts, the accuracy was 93.7%. The results support the idea that incorporating uncertainty and interpretations improves model confidence and reduces diagnostic error rates.

In 2022, S. Asif et al. [17] Employed TL in the pre-trained ResNet50 CNN to improve the model's precision, incorporated a new block "fully connected" and over 20 epochs achieved an accuracy of 99.48% on the UCSD-V2 dataset. The study overlooked the imbalance in the dataset.

In 2023, V. Latha et al. [18] Presented a method for detecting macular diseases in OCT images by merging the feature vectors of VGG16 and InceptionV3 models, using TL for enhanced local and global feature recognition. Their model, applied to the UCSD dataset (versions 2 and 3), with 50 epochs, achieved accuracies of 99.7% and 98.1%, respectively, with image augmentation as a pre-processing step.

Also, in 2023, P. Elena-Anca [19] evaluated five DL models, including a 12-layer convolutional model, InceptionResNet, DenseNet201, DenseNet121, and DenseNet169. The study highlighted the pre-trained DenseNet169 model's superior performance, achieving a 97% accuracy rate on the UCSD-V2 dataset for retinal disease diagnosis in 25 epochs. Notably, this study did not

incorporate any pre-processing procedures.

Furthermore, in 2023, İ. Kayadibi and G. Güraksın [20] suggested using FD-CNN with dual pre-processing for retinal disease identification. D-KNN and D-SVM were used to reclassify. D-SVM outperformed both in the UCSD-v2 dataset, recording an accuracy of 99.60%, whereas accuracy was 97.50% in the Duke dataset. Number of epochs was 5. However, the imbalanced dataset issue was overlooked.

Moreover, in 2023, O. Akinniyi et al. [21] Proposed a multi-stage classification network built on a pyramidal feature ensemble framework, using the pre-trained DenseNet model as the foundational network. The system demonstrated an accuracy of 94.26% for the comprehensive four-class classification by using the UCSD-V3 dataset and 99.69% on the Duke dataset over 50 epochs. There isn't noise removal in images, which could result in misclassification accuracy.

Continuing in 2023, P. Jayanthi et al. [22] Applied a transfer learning approach with VGG19, ResNet50, and a custom-built sequential model. They reported classification accuracies of 97.2%, 95.8%, and 99.6% on the UCSD-V2 dataset over 25 epochs. The custom model demonstrated superior accuracy compared to the pre-trained models. Despite the high accuracy, the dataset required balancing.

In 2024, J. Yang et al. [23] addressed the problem of diagnosing retinal diseases in OCT images from the Duke dataset with a split of 80% for training and 20% for testing. The ensemble approach was used based on three sub-models, namely AlexNet, EfficientNetv2, and ResNet34, whose results were combined using the soft voting method with the application of TL to improve performance. The proposed model achieved an accuracy of 97.89% over 100 epochs and showed a good ability to distinguish between different cases with a clear interpretation of the results.

In this paper, we propose an innovative approach for ensemble learning, emphasizing a novel approach to model specialization and misclassification correction. It was noted that all previous studies did not use a mechanism to correct classification errors, and we considered this point to be one of the most important points that we focused on in our study and suggested. To add further challenge to our approach, we have trained all models from scratch, deliberately avoiding using transfer learning techniques. Moreover, we not only used the UCSD-v2 dataset [24] but also applied our model to the Duke dataset [25] and the OCTID dataset. We also have implemented multi-step pre-processing to eliminate noise and accurately delineate the area of interest in the data.

3. PROPOSED METHODOLOGY

The proposed study will be detailed from pre-processing to classification below.

A. Image Pre-processing

The OCT images used suffer from many problems, such as differences in size and quality, shapes (square or rectangular), and zoom ratios. They also contain noise, such as salt and pepper noise and white background pixels, affecting the image analysis process used to train CNN networks.

The following steps, as shown in Figure 2, are performed to solve the most important problems mentioned above. First, the white background pixels of the image are colored black. In the second step, the pixel values in the image are normalized to enhance contrast and detail. In the third step, Gaussian filtering and adaptive thresholding are applied to the image to identify and extract the contour coordinates in a rectangular shape of the largest object, which represents the retina, and then used to crop the region of interest (ROI) from the OCT image. In the fourth step, the image contrast is intensified, binary thresholding is applied, and median blurring is used to isolate and extract the largest contour, replacing points outside the object with black points. In the final step, the image is resized to (200*80) pixels while maintaining its height ratio, centring it within the new dimensions. Because the retina is rectangular, and the weights of models are not used, rectangular images are accepted.

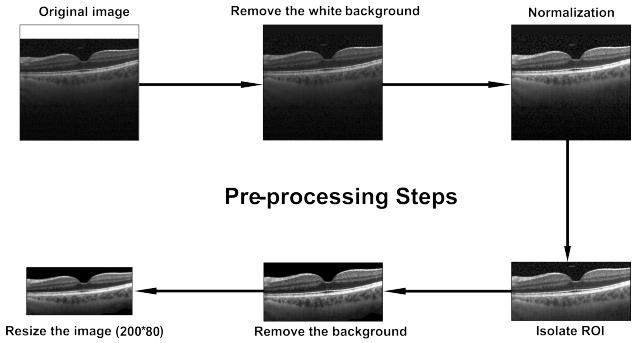

Figure 2. Pre-processing of OCT Image

B. Chosen Algorithms

The proposed ensemble model is composed of three distinct algorithms:

1) Custom model. The proposed CNN sub-model was created by using the Keras toolbox. There are (703,821) parameters that can be learned, while those that can't be learned are (1,152). The model comprises 60 layers, starting with a separable convolution, batch normalization, and a "Relu" activation layer. It ends at the "Softmax" activation function layer, which gives the probability of each class. The whole architecture is depicted in Table I.

2) Pre-trained models. The models used are Xception and MobileNet, with some layers that include several dense layers with L2 regularization set at 0.001 with some dropout layers to reduce overfitting, batch normalization layers to enhance performance, and activation layers that facilitate the learning of complex patterns by introducing non-linearity, also in the end the "Softmax" activation function layer is used. These additions

bolster the overall effectiveness of the models, which are trained from scratch without using transfer learning.

These algorithms were carefully selected for their efficiency in extracting features from retinal OCT images and their diversity in the number of trainable parameters (Custom model: 704,973, MobileNet: 16,509,444, Xception: 65,599,340) that are particularly effective for our ensemble model. Their combination ensures robustness and enhances the ensemble model's overall performance. As a result of an extensive evaluation process that included many deep neural network algorithms, this was the choice, as it showed superior performance compared to others in this study and others, as in [26].

C. Proposed Ensemble Learning Model

Ensemble learning in machine learning integrates outcomes from multiple algorithms, thereby enhancing performance beyond what individual algorithms can achieve [27]. The three main ensemble learning techniques are noteworthy: stacking, boosting, and bagging. Bagging, which stands for Bootstrap Aggregating, combines the predictions of several models trained on different subsets of data. A series of models known as "boosting" are trained to gradually improve performance by fixing the mistakes of the previous model. In addition, the hard voting method aggregates predictions by majority votes to derive the final decision, and the soft voting method selects the vote with the highest probability among the average probabilities from the sub-models. Voting methods can be used independently or as a component of main methods.

In this research, we introduce a novel ensemble model (as shown in Figures 3, 4, 5, and 6, in addition to Algorithm 1) that can be called "Negative Prediction-Based Specialization Ensemble Model". This model integrates the strengths of multiple sub-models to enhance classification accuracy. It is noted that most research utilizing ensemble learning for diagnosing retinal diseases neglects sub-models with less fortunate accuracy.

Issue is addressed by the proposed model that incorporates two key elements:

- Firstly, determining the best sub-model in classifying each category in the training set. The model that achieves the highest accuracy for a particular class becomes specialized in that class and is given priority in the final data classification. In the absence of a specialized model, the hard voting method is used, or the highest probability method is used.

- Secondly, creating a negative prediction list, supplemented with categories by each sub-model, to identify categories to which it is somewhat confident that a given image does not belong. This means that not only high-accuracy sub-models have strengths, but less successful models also have strengths that can be exploited.

D. Specializing Each Sub-model

We noticed that sub-models may be better than each other in classifying a particular category, and this feature was not exploited in previous studies in diagnosing retina diseases, so we added the character of specialization to the models, so after the training process, weights are used to predict each class of training set separately. This approach ensures that each sub-model specializes in the category or categories that it classifies better than other sub-models, thus enhancing the overall accuracy of the ensemble model, as shown in Figures 3. Note that when a particular model specializes in a specific category, its predictions are not limited to that category alone, but it assumes priority.

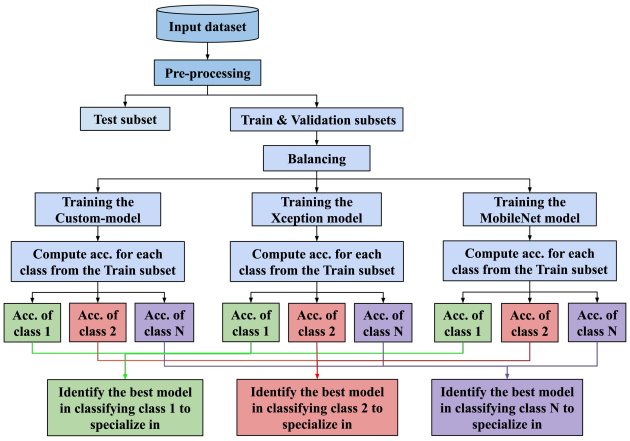

Figure 3. Specializing Each Sub-model

E. Negative Predictions List (NP-list)

The proposed ensemble model amalgamates the advantages of all sub-models, where each one specializes in the category it classifies most effectively. The rest of the models are not neglected but contribute by identifying categories that they are somewhat certain the specific image does not belong to (as shown in Figure 4). This strategy enhances the accuracy by enabling models to correct each other's misclassification.

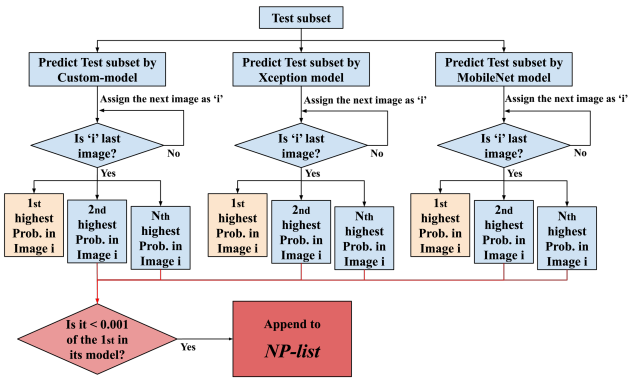

Figure 4. Negative Predictions List. (created by each sub-model)

TABLE I. Architecture of Custom-Model

Layers (type)	Param#	Layers (type)	Param#
Input Layer (1, 200*80)	0	Dropout (32, 8*3)	0
SeparableConv2D (32, 200*80)	73	MaxPooling2D (32, 3*1)	0
BatchNormalization (32, 200*80)	128	Conv2D (64, 3*1)	2112
Activation (32, 200*80)	0	Conv2D (64, 3*1)	2112
SeparableConv2D (32, 200*80)	1344	Conv2D (64, 3*1)	2112
BatchNormalization (32, 200*80)	128	Conv2D (64, 3*1)	36928
Activation (32, 200*80)	0	Conv2D (64, 3*1)	102464
MaxPooling2D (32, 67*27)	0	Concatenate (192, 3*1)	0
SeparableConv2D (64, 67*27)	2400	Conv2D (64, 3*1)	12352
BatchNormalization (64, 67*27)	256	BatchNormalization (64, 3*1)	256
Activation (64, 67*27)	0	Activation (64, 3*1)	0
SeparableConv2D (64, 67*27)	4736	Dropout (64, 3*1)	0
BatchNormalization (64, 67*27)	256	MaxPooling2D (64, 1*1)	0
Activation (64, 67*27)	0	Conv2D (96, 1*1)	6240
MaxPooling2D (64, 23*9)	0	Conv2D (96, 1*1)	6240
SeparableConv2D (96, 23*9)	6816	Conv2D (96, 1*1)	6240
BatchNormalization (96, 23*9)	384	Conv2D (96, 1*1)	83040
Activation (96, 23*9)	0	Conv2D (96, 1*1)	230496
SeparableConv2D (96, 23*9)	10176	Concatenate (288, 1*1)	0
BatchNormalization (96, 23*9)	384	Conv2D (96, 1*1)	27744
Activation (96, 23*9)	0	BatchNormalization (96, 1*1)	384
MaxPooling2D (96, 8*3)	0	Activation (96, 1*1)	0
Conv2D (32, 8*3)	3104	Dropout (96, 1*1)	0
Conv2D (32, 8*3)	3104	MaxPooling2D (96, 1*1)	0
Conv2D (32, 8*3)	3104	Conv2D (128, 1*1)	110720
Conv2D (32, 8*3)	9248	Attention (128, 1*1)	0
Conv2D (32, 8*3)	25632	Concatenate (256, 1*1)	0
Concatenate (96, 8*3)	0	Conv2D (4, 1*1)	1028
Conv2D (32, 8*3)	3104	GlobalAveragePooling2D (4)	0
BatchNormalization (32, 8*3)	128	Activation (4)	0
Activation (32, 8*3)	0		

In all models, the "Softmax" activation function layer was used to perform the final classification, and since it gives the probabilities of all classes, we took advantage of this feature. The class whose probability is less than one in a thousand from the highest probability in that model, where satisfies the condition in the following equation, is added to the NP list.

$$\text{prob. (class I)} < 0.001 \times \text{highest-prob.} \quad (1)$$

This list helps by correcting any misclassification of the specialized model, if any, or the misclassification of the majority voting method and the highest probability method, which are used in the absence of a specialized model among the highest predictions of the three models.

F. Final Classification of The Proposed Model

If the top prediction of each sub-model is for a class that is not specialized in it, we use the majority voting method. However, if the votes are equal or the majority voting category belongs to an NP list, the prediction with the highest probability is adopted. If the highest probability class also belongs to the NP list, we choose the next highest probability prediction, and so on (as shown in Figure 5 and Figure 6).

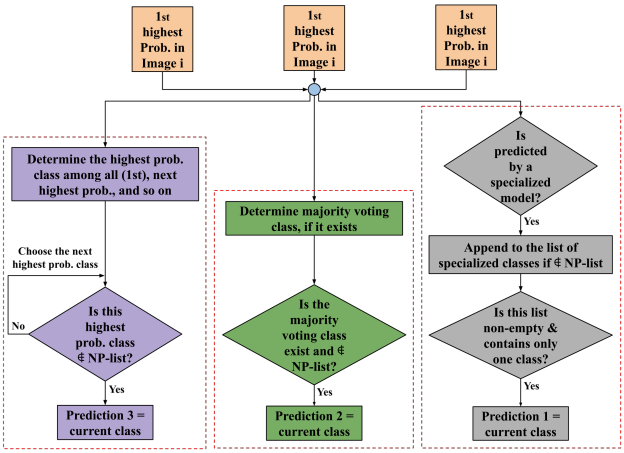

Figure 5. The classes resulting from the three methods (specialization, highest vote, highest probability)

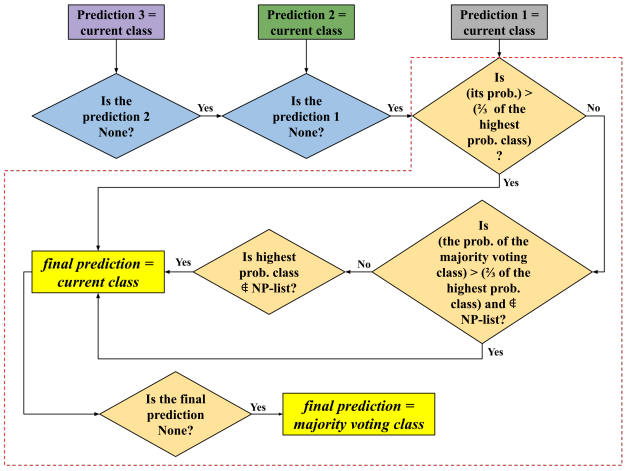

Figure 6. Final classification

Suppose the final prediction, whether from a specialized or non-specialized model, is less than two-thirds of the highest probability prediction. The prediction with the most votes or the highest probability is chosen in this case. The process of calculating the final prediction in this ensemble model is summarized in the following pseudo-code:

Algorithm 1 : Algorithm to Calculate Final Prediction

Initializing the required variables.
for *each sub-model in the models:*
    Train the model on the training set.
    Use weights of the current model to compute its accuracy on each class in the training set separately.
end
Specialize each sub-model to a specific class (or classes) in which it outperforms other models in accuracy.
for *each image in the testing set:* ▷ *//Forming the Negative Prediction list (NP-list)*
    Compute the predicted probability (prob.) for each class by each model for that image and identify the max predicted probability (max-prob.).
    Append the prob. and its class to the predictions list.
    Identify classes that have very low probabilities for each model if the condition is met:
prob. $< 0.001 *$ max-prob. ▷ *// less than one in a thousand from the max-prob. in that model.*
    Append these low-probability classes to the NP-list.
end
for *all max-prob. from all models of each image in the testing set:* ▷ *//Final Classification.*
    Determine which class has the majority of votes and which class has the highest-probability among all models for the current image.
    Calculate the number of models that show specialization in their max-prob., when the classes ∉ NP-list.
    if *there is one specialized model:*
        Final prediction = predicted class of specialized model.
    else if *there is more than one specialized model*
        Final prediction = majority voting class if it ∉ NP-list.
    else if *there is no specialized model*
        ▷ *//Each max-prob. class, its predicted model did not specialize in.*
        if *the majority voting class exists and ∉ NP-list:*
            Final prediction = majority voting class.
        else if *the highest-probability class ∉ NP-list*
            Final prediction = highest-probability class.

        else
            Choose the next highest-probability class that is ∉ NP-list, and so on.
        end
    end
    if *the prob. of the final prediction < (2/3 * highest-probability):*
        ▷ *//When the final prediction probability is less than two-thirds of the highest probability.*
        if *(majority voting class ∉ NP-list) & (its prob. > 2/3 * highest-probability):*
            Final prediction = majority voting class.

        else if *highest-probability class ∉ NP-list*
            Final prediction = highest-probability class.

        else
            Final prediction = next highest-probability class if it was ∉ NP-list, and so on.
        end
    end
  if *Final prediction == None:*

    Final prediction = majority voting class.
  end
end

To further clarify the proposed ensemble model, it can be summarized in full in the following steps:

- **Step 1:** Loading the dataset, applying pre-processing on images, and then balancing the dataset, as shown in Figure 3.
- **Step 2:** Training the three sub-models from scratch, and then all images for each class in the training subset are isolated, and each class is predicted separately to specialize each class to the sub-model that classifies it better than the rest of the sub-models, as shown in Figure 3.

- **Step 3:** The trained sub-models are used to predict the test subset, as shown in Figure 4.
- **Step 4:** Identifying class (es) to which the sub-models are somewhat certain a specific image does not belong, and each sub-model adds that particular class (es) to the NP-list, as shown in Figure 4.
- **Step 5:** The final classification is carried out using the proposed ensemble model based on the predictions and specializations made in the previous steps in addition to the NP-list, as shown in Figure 5 and Figure 6.

4. RESULTS AND DISCUSSION

The experimental results obtained are detailed in this section.

A. Datasets Used

images of this dataset were acquired using a "Spectralis OCT" device from "Heidelberg Engineering", Germany, and cohorts of adult patients associated with several prestigious institutions were imaged between July 1, 2013, and March 1, 2017.It consists of four categories: Normal, CNV, DME, and DRUSEN, with a total number of (84,484) OCT images (83,484 train, 968 test, 32 validation), which exhibited an imbalance, and the verification data has very few of images, which impacting the training of CNNs. To address this issue, we employed an oversampling technique to equalize the distribution. Then the training set was split into 80% training and 20% validation, resulting in (29,771) in training and (7,442) in validation for each category, as illustrated in Table II. In addition to the original split that occurred, the dataset was re-partitioned into two splits (80:20 and 70:30), and the model was applied to both because the original split had little testing set.

To ensure the success of our ensemble model, we applied it to other datasets, Duke and OCTID.

The Duke dataset includes scans collected from 45 patients, 15 for each class. All scans were collected according to Institutional Review Board-approved protocols using the Spectralis SD-OCT (Heidelberg et al.) imaging device at prestigious institutions such as Duke University and others in 2014. It includes 3,231 OCT images distributed across three distinct categories: 1,407 for normal, 723 (or 686) AMD, and 1,101 for DME.

The OCTID dataset includes 572 spectral-domain OCT scans, classified into different categories: 206-Normal, 102-Macular Hole (MH), 55-Age-related Macular Degeneration (AMD), 102-Central serous retinopathy (CSR), and 107-Diabetic retinopathy (DR). These images were captured with a Cirrus HD-OCT device at Sankara Nethralaya (SN) in India at the end of 2018.

According to the division detailed in Table III for Duke and Table IV for OCTID, where 80% for training, 20% for testing, and from the training set split to 10% for validation and 90% for training.

B. Implementation

We used Python to implement the software using Keras to develop the CNN models, with a batch size of 32, opting for Adam as the optimizer with a learning rate of 0.001. The software was implemented by using PyCharm on an ASUS TUF Dash F15 equipped with a 12th Gen Intel(R) Core (TM) i7-12650H, a ten-core CPU operating at 2.30 GHz, 40 GB RAM, and 8GB NVIDIA GeForce RTX 3070 Laptop GPU. NVIDIA's CUDA Toolkit 11.8 and cuDNN 8.6.0 are used for their ability to improve training speed.

C. Model Evaluation

The performance of the three models across (8, 15, and 17 epochs), respectively, in the splits (Original split, 80:20, and 70:30) on the UCSD-v2 dataset is detailed in Table VI where our ensemble model reached 100% accuracy in the original split, 96.03% in 80:20 split, and 95.85% in 70:30 split. The 100% accuracy was also attained on the Duke dataset, as indicated in the previously mentioned table, albeit after (17) epochs. And 95.73% accuracy was achieved in the OCTID dataset after (50) epochs. To the authors' knowledge, the results achieved demonstrate state-of-the-art accuracy and outperform any other model trained and tested on the UCSD-v2 dataset. The accuracy of each sub-model is detailed in Table VI.

D. Proposed Ensemble Learning

The accuracy of each sub-model is depicted in Table VI. On the testing set of the UCSD-v2 dataset, the Custom model, Xception, and MobileNet achieved 99.79%, 99.59%, and 99.59%, respectively, in the original split. And achieved 94.11%, 95.09%, and 95.63%, respectively, in the 80:20 split. In the 70:30 split achieved 94.01%, 94.67%, and 95.31%, respectively. The specialization of these models, detailed in Table VIII, reveals that in the original split, the Custom model has superior performance in identifying Class 2 (Drusen) and Class 3 (Normal), Xception in Classes 1 (DME), and MobileNet in Class 0 (CNV), that illustrating the unique strengths of each model within their respective domains. On the testing set of the Duke dataset, the sub-models registered accuracies of 99.69%, 95.37%, and 95.22%, respectively. Here, the Custom model specializes in class 0 (AMD), the Xception in class 1 (DME), and the MobileNet in class 2 (Normal). For the testing set of the OCTID dataset, the sub-models achieved accuracies of 89.74%, 70.94%, and 91.45%, respectively. Here, the Custom model specializes in class 3 (MH) and class 4 (Normal), the Xception in class 1 (CSR) and class 2 (DR), and the MobileNet in class 0 (AMD). Although the MobileNet sub-model failed in its speciality, misclassification results from such mistakes can be avoided through the NP-list. All sub-models play a crucial role in correcting misclassifications by identifying classes that a given image is somewhat certain not to belong to, and these classes form the NP-list. The custom model works exceptionally well in most cases. The confusion matrixes are illustrated in Figure 7-a, Figure. 7-b, Figure. 7-c, Figure 7-d, and Figure. 7-e.

TABLE II. UCSD-V2 Dataset Before and After Balancing

State	Data	CNV	DME	Drusen	Normal	Total
Before	Train	37,205	11,348	8,616	26,315	83,484
	Test	242	242	242	242	968
	Val	8	8	8	8	32
After	Train	29,771	29,771	29,771	29,771	119,084
	Test	242	242	242	242	968
	Val	7,442	7,442	7,442	7,442	29,768

TABLE III. DUKE Dataset Before and After Balancing

State	Data	AMD	DME	Normal	Total
Before	All data	723	1,101	1,407	3,231
After	Train	1,013	1,013	1,013	3,039
	Test	145	221	282	648
	Val	112	112	112	336

TABLE IV. OCTID Dataset Before and After Balancing

State	Data	MH	AMD	CSR	No	DR	Total
Before	All data	102	55	102	206	107	572
After	Train	148	148	148	148	148	740
	Test	21	11	21	42	22	117
	Val	16	16	16	16	16	80

TABLE V. Performance Metrics of the UCSD-V2 Dataset

Splitting	Accuracy	Sensitivity	Specificity	Precision
Original split	100%	100%	100%	100%
80:20	96.03%	94.22%	98.59%	94.75%
70:30	95.85%	94.04%	98.53%	94.47%

Some images have one specialized model, others have more than one, or there is no specialist. All these cases are mentioned in Algorithm 1 and illustrated in Table VII, which gives an example for each of these cases based on the Duke dataset. Note that in the first example, there is no specialist, but the correct class is the one with the highest probability. In the second example, there is one specialist who is the one with the correct class. In the third example, there are two specialists, and class 2 has the majority of votes but belongs to the NP-list, which helps correct the misclassification and select the correct class.

E. Comparison

Compared to models developed by other researchers using the UCSD-v2, Duke, and OCTID datasets, our ensemble model achieves an impressive accuracy of 100% on the first dataset, while the highest accuracy in previous studies was 99.7%. For the second dataset, our model also achieved 100% accuracy, equal to one of the research papers, but we used all the images in the dataset and that paper used a subset of them, so our model was considered the outperformer, while on the entire dataset, the highest accuracy in previous studies is 99.69%. On the third dataset, the proposed model outperformed by a clear margin, achieving an accuracy of 95.73%, while the highest accuracy achieved in previous studies on the entire dataset is 88.5% and on a subset of the dataset, they reached 93.7%.

Table IX presents the comparative analysis for the UCSD-v2 dataset, highlighting the performance of the proposed model concerning its counterparts. Moreover, the ensemble learning approach introduced at the table's conclusion exhibits enhanced performance, 100% (accuracy, sensitivity, specificity, and precision). Likewise, Table X delineates the comparative outcomes for the Duke dataset, the ensemble learning method achieved a complete accuracy, sensitivity, specificity, and precision of 100%. Moreover, Table XI illustrates the comparative outcomes for the OCTID dataset, the proposed model achieved an accuracy, sensitivity, specificity, and precision of 95.73%, 90.91%, 98.90%, and 96.17%, respectively.

The comparison in the indicated tables includes the

TABLE VI. Accuracy of Proposed Model and Sub-Models

Datasets	Custom-model accuracy	Xception accuracy	MobileNet accuracy	Ensemble model accuracy
UCSD-v2 (Original)	99.79%	99.59%	99.59%	100%
UCSD-v2 (80:20)	94.11%	95.09%	95.63%	96.03%
UCSD-v2 (70:30)	94.01%	94.67%	95.31%	95.85%
Duke (80:20)	99.69%	95.37%	95.22%	100%
OCTID (80:20)	89.74%	70.94%	91.45%	95.73%

TABLE VII. Examples of Applying the Proposed Model

Specialist		**Predictions**			**NP list**	**Final prediction**	**True label**
		Custom	**Xception**	**MobileNet**			
None	Class Prob.	1 0.469	2 0.977	0 0.699	[1]	2	2
One	Class Prob.	1 0.999	1 0.988	0 0.782	[0, 2]	1	1
Two	Class Prob.	0 0.999	2 0.989	2 0.996	[1, 2]	0	0

TABLE VIII. Specialization of Sub-models

UCSD-v2 dataset (original split)					
Classes	**Class 0 (CNV)**	**Class 1 (DME)**	**Class 2 (Drusen)**	**Class 3 (Normal)**	
Specialized sub-model **Acc. on testing set**	MobileNet 0.996	Xception 0.996	Custom model 1.0	Custom model 0.996	
Duke dataset					
	Classes	**Class 0 (AMD)**	**Class 1 (DME)**	**Class 2 (Normal)**	
	Specialized sub-model **Acc. on testing set**	Custom model 0.993	Xception 0.991	MobileNet 0.997	
OCTID dataset					
Classes	**Class 0 (AMD)**	**Class 1 (CSR)**	**Class 2 (DR)**	**Class 3 (MH)**	**Class 4 (No)**
Specialized sub-model **Acc. on testing set**	MobileNet 0.091	Xception 0.857	Xception 1.0	Custom model 0.857	Custom model 1.0

number of epochs and the number of sub-models, if any, in addition to the various metrics.

5. Conclusions and Future Work

This research demonstrates the efficacy of the proposed novel ensemble model in the classification of retinal diseases, specifically CNV, DME, and Drusen, in addition to the categories from other datasets. This study attempts to benefit as much as possible from the capabilities of all the models used (Custom, Xception, and MobileNet) to improve classification accuracy. This is achieved through a strategic exclusion list (NP-list) that mitigates misclassifications by identifying non-relevant classes for each image.

In this approach, each sub-model specializes in the category in which it achieves higher accuracy than others. One of the most prominent benefits of this method is that if a certain sub-model achieves low accuracy and the rest of the models are higher than it, then it will not be specialized in a specific category, which reduces the risk of misclassification.

Pre-processing had an important role in improving the image, reducing noise, and identifying the region of interest.

The proposed ensemble model achieved state-of-the-art accuracies of (100%, 96.03%, and 95.85%, respectively) on three splits of the UCSD-v2 dataset and similarly high performance on the Duke dataset (100%) and OCTID dataset (95.73%).

In comparison to the separate models and to that in previous studies, the suggested one was better, firstly, as it was a simulation of the human way of decision-making when the outputs of three sub-models are combined, and secondly, utilized from the advantages of each sub-model and avoided its disadvantages, raised the accuracy and reliability of the results.

The proposed model is characterized as less susceptible

TABLE IX. Comparison with Previous Studies (UCSD-V2)

Method	Year	Accuracy	Sensitivity	Specificity	Precision	Epochs	No. of models
[3]	2020	98.53%	97.5%	-	97.02%	250	4
[13]	2020	99.27%	-	-	-	15	5
[14]	2021	99.64%	99.28%	-	99.29%	-	8
[17]	2022	99.48%	99%	-	99%	20	1
[18]	2023	99.7%	99.7%	99.9%	99.7%	50	2
[19]	2023	97%	-	-	-	25	5
[20]	2023	99.6%	99.6%	99.87%	99.6%	5	1
[22]	2023	99.6%	-	-	-	5	3
Proposed custom model	2024	99.79%	99.69%	100%	99.79%	8	1
Proposed ensemble model	2024	100%	100%	100%	100%	8	3

TABLE X. Comparison with Previous Studies (DUKE)

Method	Year	Accuracy	Sensitivity	Specificity	Precision	Epochs	No. of models
[15]	2021	100%	100%	100%	100%	-	18
[20]	2023	97.5%	97.64%	98.91%	96.61%	5	1
[21]	2023	99.69%	99.71%	99.87%	-	50	1
[23]	2024	97.89%	97.89%	-	97.9	100	3
Proposed custom model	2024	99.69%	99.69%	100%	99.69%	17	1
Proposed ensemble model	2024	100%	100%	100%	100%	17	3

TABLE XI. Comparison with Previous Studies (OCTID)

Method	Year	Accuracy	Sensitivity	Specificity	Precision	Epochs	No. of models
[16]	2021	88.5%	84.6%	-	86.2%	45	1
		93.7%	91.2%	-	91.4%	45	1
Proposed Custom model	2024	89.74%	87.18%	97.43%	92.73%	50	1
Proposed ensemble model	2024	95.73%	90.91%	98.90%	96.17%	50	3

to the known problems of ensemble models such as overfitting, variance, and bias, because it uses a specialization strategy and an NP list. One of the drawbacks of the proposed model is that it requires a variety of sub-models and that it may not succeed in specializing sub-models correctly if the dataset is small.

These results emphasize the importance of ensemble learning techniques in medical image analysis, especially in retina OCT images. This study underscores the necessity for cooperation between different specialities and technological progress, especially in health care, to shorten the time and reduce diagnostic errors.

In future work, researchers could aim to advance this study, especially in expanding it to include more diverse and different datasets and many imaging modalities. Some examples are computed tomography (CT), magnetic resonance imaging (MRI), and vascular ultrasound.

Applying and analyzing the performance of other DL models might yield a further increase in the accuracy and effectiveness of the model.

Exploring the opportunities of using common ML algorithms such as SVM and Random Forest as sub-models in the proposed ensemble model to assess their effectivity compared to DL models.

Developing the NP-list and investigating its influence on correcting misclassification when used in the other models and approaches such as bagging, boosting, and stacking.

Refining the details of the proposed model, addressing its weaknesses, and assessing the use of more than three sub-models to enhance accuracy can be considered for future work.

Implementing more sophisticated methods to improve image quality and highlight ROI can improve the model result.

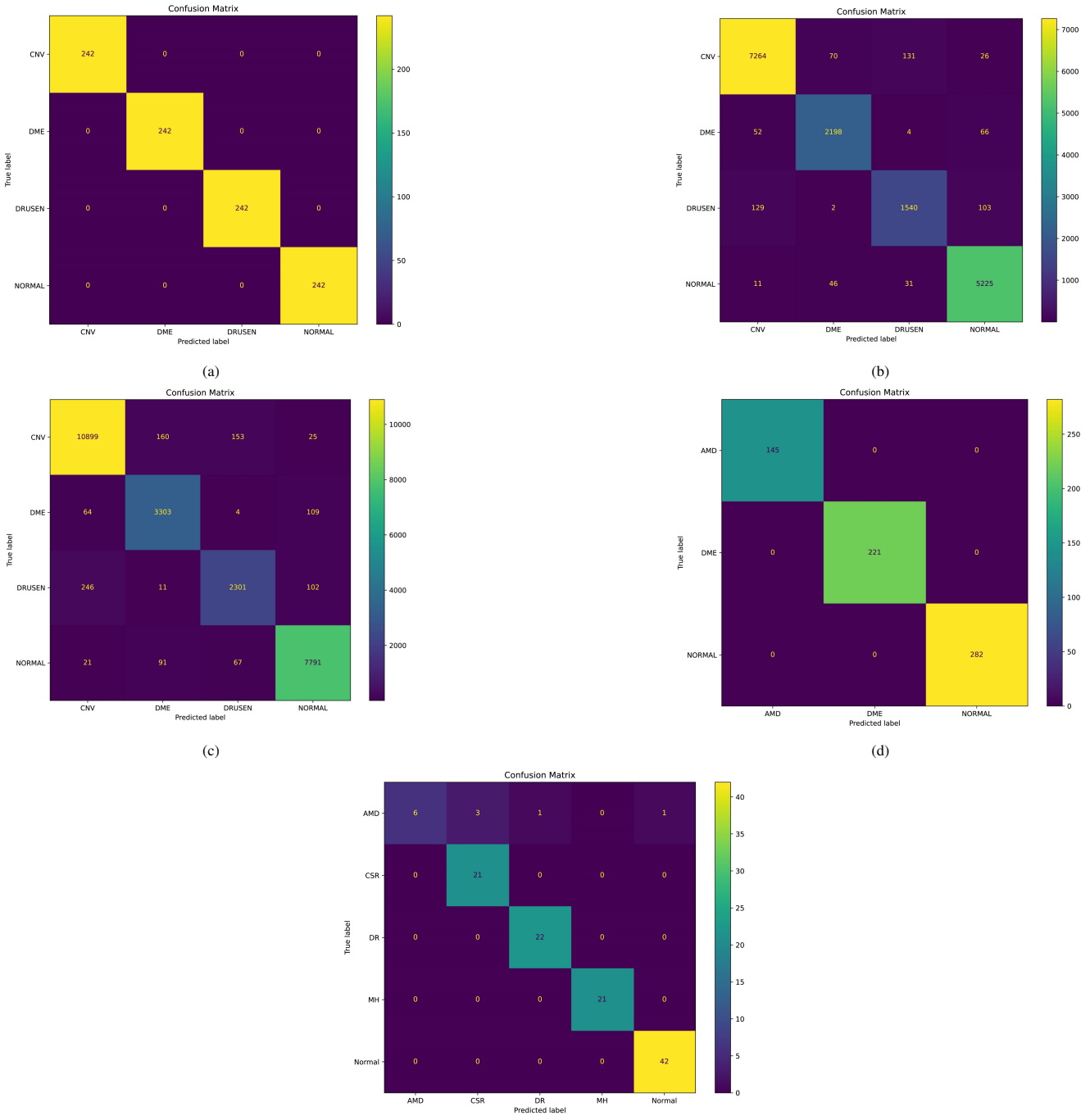

Figure 7. Confusion Matrixes. ((a) for the UCSD-v2 dataset (original split); (b) for the UCSD-v2 dataset (80:20); (c) for the UCSD-v2 dataset (70:30); (d) for the Duke dataset (80:20); (e) for the OCTID dataset (80:20))

By pursuing these ideas, the proposed system can be further optimized, and its applications broadened to include other medical fields, providing more accurate and effective diagnostic tool for doctors and specialists in retinal diseases.

References

[1] M. R. Ibrahim, K. M. Fathalla, and S. M. Youssef, "Hycad-oct: A hybrid computer-aided diagnosis of retinopathy by optical coherence tomography integrating machine learning and feature maps localization," *Applied Sciences*, vol. 10, no. 14, p. 4716, 2020.

[2] L. Fang, C. Wang, S. Li, H. Rabbani, X. Chen, and Z. Liu, "Attention to lesion: Lesion-aware convolutional neural network for retinal optical coherence tomography image classification," *IEEE transactions on medical imaging*, vol. 38, no. 8, pp. 1959–1970, 2019.

[3] D. Paul, A. Tewari, S. Ghosh, and K. Santosh, "Octx: Ensembled deep learning model to detect retinal disorders," in *2020 IEEE 33rd International Symposium on Computer-Based Medical Systems (CBMS)*. IEEE, 2020, pp. 526–531.

[4] J. Ong, A. Zarnegar, G. Corradetti, S. R. Singh, and J. Chhablani, "Advances in optical coherence tomography imaging technology and techniques for choroidal and retinal disorders," *Journal of Clinical Medicine*, vol. 11, no. 17, p. 5139, 2022.

[5] R. K. Ara, A. Matiolański, A. Dziech, R. Baran, P. Domin, and A. Wieczorkiewicz, "Fast and efficient method for optical coherence tomography images classification using deep learning approach," *Sensors*, vol. 22, no. 13, p. 4675, 2022.

[6] M. Subramanian, M. S. Kumar, V. Sathishkumar, J. Prabhu, A. Karthick, S. S. Ganesh, and M. A. Meem, "Diagnosis of retinal diseases based on bayesian optimization deep learning network using optical coherence tomography images," *Computational Intelligence and Neuroscience*, vol. 2022, no. 1, p. 8014979, 2022.

[7] A. G. Roy, S. Conjeti, S. P. K. Karri, D. Sheet, A. Katouzian, C. Wachinger, and N. Navab, "Relaynet: retinal layer and fluid segmentation of macular optical coherence tomography using fully convolutional networks," *Biomedical optics express*, vol. 8, no. 8, pp. 3627–3642, 2017.

[8] J. De Fauw, J. R. Ledsam, B. Romera-Paredes, S. Nikolov, N. Tomasev, S. Blackwell, H. Askham, X. Glorot, B. O'Donoghue, D. Visentin *et al.*, "Clinically applicable deep learning for diagnosis and referral in retinal disease," *Nature medicine*, vol. 24, no. 9, pp. 1342–1350, 2018.

[9] A. Adel, M. M. Soliman, N. E. M. Khalifa, and K. Mostafa, "Automatic classification of retinal eye diseases from optical coherence tomography using transfer learning," in *2020 16th International computer engineering conference (ICENCO)*. IEEE, 2020, pp. 37–42.

[10] P. Dutta, K. A. Sathi, M. A. Hossain, and M. A. A. Dewan, "Conv-vit: a convolution and vision transformer-based hybrid feature extraction method for retinal disease detection," *Journal of Imaging*, vol. 9, no. 7, p. 140, 2023.

[11] N. Ferrara, "Vascular endothelial growth factor and age-related macular degeneration: from basic science to therapy," *Nature medicine*, vol. 16, no. 10, pp. 1107–1111, 2010.

[12] R. Varma, N. M. Bressler, Q. V. Doan, M. Gleeson, M. Danese, J. K. Bower, E. Selvin, C. Dolan, J. Fine, S. Colman *et al.*, "Prevalence of and risk factors for diabetic macular edema in the united states," *JAMA ophthalmology*, vol. 132, no. 11, pp. 1334–1340, 2014.

[13] M. Berrimi and A. Moussaoui, "Deep learning for identifying and classifying retinal diseases," in *2020 2nd International Conference on computer and information sciences (ICCIS)*. IEEE, 2020, pp. 1–6.

[14] H. A. Nugroho and R. Nurfauzi, "Convolutional neural network for classifying retinal diseases from oct2017 dataset," in *2021 4th International Conference on Information and Communications Technology (ICOIACT)*. IEEE, 2021, pp. 295–298.

[15] P. D. Barua, W. Y. Chan, S. Dogan, M. Baygin, T. Tuncer, E. J. Ciaccio, N. Islam, K. H. Cheong, Z. S. Shahid, and U. R. Acharya, "Multilevel deep feature generation framework for automated detection of retinal abnormalities using oct images," *Entropy*, vol. 23, no. 12, p. 1651, 2021.

[16] A. Singh, S. Sengupta, M. A. Rasheed, V. Jayakumar, and V. Lakshminarayanan, "Uncertainty aware and explainable diagnosis of retinal disease," in *Medical Imaging 2021: Imaging Informatics for Healthcare, Research, and Applications*, vol. 11601. SPIE, 2021, pp. 116–125.

[17] S. Asif and K. Amjad, "Deep residual network for diagnosis of retinal diseases using optical coherence tomography images," *Interdisciplinary Sciences: Computational Life Sciences*, vol. 14, no. 4, pp. 906–916, 2022.

[18] V. Latha and K. Sreeni, "Oct image-based macular disease classification using multilayer deep feature fusion," in *2023 International Conference on Control, Communication and Computing (ICCC)*. IEEE, 2023, pp. 1–6.

[19] P. Elena-Anca, "Applications of deep learning algorithms for retinal diseases diagnosis based on optical coherence tomography imaging," in *2023 24th International Conference on Control Systems and Computer Science (CSCS)*. IEEE, 2023, pp. 594–597.

[20] İ. Kayadibi and G. E. Güraksın, "An explainable fully dense fusion neural network with deep support vector machine for retinal disease determination," *International Journal of Computational Intelligence Systems*, vol. 16, no. 1, p. 28, 2023.

[21] O. Akinniyi, M. M. Rahman, H. S. Sandhu, A. El-Baz, and F. Khalifa, "Multi-stage classification of retinal oct using multi-scale ensemble deep architecture," *Bioengineering*, vol. 10, no. 7, p. 823, 2023.

[22] P. Jayanthi, N. Krishnamoorthy, S. Sridharan, R. Tamilkumar, and P. Yokesh, "An enhanced technique to classify oct images using deep learning," in *2023 Second International Conference on Electrical, Electronics, Information and Communication Technologies (ICEEICT)*. IEEE, 2023, pp. 1–5.

[23] 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, Mar 2024.

[24] P. Mooney, "Retinal oct images (optical coherence tomography)," *Kaggle dataset*, 2018.

[25] P. P. Srinivasan, L. A. Kim, P. S. Mettu, S. W. Cousins, G. M. Comer, J. A. Izatt, and S. Farsiu, "Fully automated detection of diabetic macular edema and dry age-related macular degeneration from optical coherence tomography images," *Biomed. Opt. Express*, 2014.

[26] M. M. Mijwil, R. Doshi, K. K. Hiran, O. J. Unogwu, and I. Bala, "Mobilenetv1-based deep learning model for accurate brain tumor classification," *Mesopotamian Journal of Computer Science*, vol. 2023, pp. 29–38, 2023.

[27] N. Altman and M. Krzywinski, "Points of significance: Ensemble methods: Bagging and random forests," *Nature methods*, vol. 14, no. 10, pp. 933–934, Oct. 2017.