2021 International Conference on Electronics, Communications and Information Technology (ICECIT), 14–16 September 2021, Khulna, Bangladesh

Interpretable Retinal Disease Classification from OCT Images Using Deep Neural Network and Explainable AI

Md Tanzim Reza[1], Farzad Ahmed[2], Shihab Sharar[3] and Annajiat Alim Rasel[4]
[1,3,4]Department of Computer Science and Engineering, BRAC University, 66 Mohakhali, Dhaka 1212, Bangladesh
[2]Department of CSE, Ahsanullah University of Science and Technology, 141 & 142, Love Rd, Dhaka 1208, Bangladesh
Email: [1]rezatanzim@gmail.com, [2]farzadahmed6@gmail.com, [3]shihab.sharar@g.bracu.ac.bd, [4]annajiat@gmail.com

***Abstract*—Deep Learning Models (DNN) are being used extensively for medical image classification such as MRI, OCT, x-ray in recent years. The proposed model revolves around the analysis of macular Optical Coherence Tomography (OCT) images to distinguish three eye-related anomalies: Choroidal Neovascularization (CNV), Diabetic Macular Edema (DME), and accumulation of Drusen from OCT images of patients. At first, the dataset was acquired and various pre-processing steps were performed on it. Then we performed a split on the dataset into train-test-validation sets with different numbers of images in each of them. Afterward, we applied pre-trained Resnet, Inception V3, and EfficientNet models in order to classify the images. From our experiment, we achieved the best accuracy of 96.9% from ResNet. Finally, we applied Explainable AI (XAI) framework though the LIME framework in an attempt to explain the reasons for misclassifications. Alongside achieving slightly better accuracy compared to the base model, the purpose of our research is to explain the reasons behind classification errors which can be utilized in the future to develop better models.**

***Index Terms*—Optical Coherence Tomography, Retina, Explainable AI, Deep Learning**

I. Introduction

Eye diseases are a very common occurrence among the general population, especially the aged ones. Choroidal Neovascularization (CNV), Diabetic Macular Edema (DME), and accumulation of Drusen in the macular region are three of the most common anomalies in the eyes. Most of the projection-based reports estimated an increase of affected patients in the coming years which indicates a very ominous future. For example, the number of age-related Macular Degeneration (AMD) patients is expected to rise towards 288 million from the current 150+ million in the year 2040 [1]. Additionally, the number of Americans older than 40 years of age affected by Diabetic Retinopathy (DR) is expected to grow triple by the year 2050 [2]. The risk is slightly lower in the Asian region but the incremental rate is still prevalent. Hence, due to the widespread nature of the disease, an automated approach for quick disease detection is extremely necessary.

Automated classification of medical images such as MRI, OCT, x-ray is being done using deep learning in recent years. In the proposed method, we analyzed macular Optical Coherence Tomography (OCT) image data to determine three eye-related anomalies: CNV, DME, and accumulation of Drusen. The dataset proposed by Kermany et al. is one of the largest OCT-based datasets available on Retinal diseases and we used the dataset for our experiment [3]. The contributions of our research are: (1) it achieved better classification scores compared to the baseline results provided in [3], and more importantly, (2) our research attempted to explain the misclassifications through the XAI framework. The aim of the research is to help the development of a better model for OCT image classification through explaining the currently existing lackings.

The section II of this paper is the background study of which consists of the literature review and model details. Section III is the proposed model section which describes the overall workflow of the research. Afterward, section IV describes the dataset details while section V discusses the results and analysis. Finally, the paper ends with a conclusion in section VI.

II. Background Study

A. Literature Review

There has been a lot of research work to classify different types of retinal diseases from OCT images. Shah et al. [4] tried to detect Stargardt disease from OCT scan images using VGG19. The authors of this paper used a trinary classification system (normal, mild, and severe STGD) to capture the spectrum of retinal deterioration of the scans. Their proposed model gave an accuracy of 99.6% but due to the small amount of dataset used in their paper it had limited capability to distinguish OCT scans with a milder disease phenotype with severe STGD.

Yoo et al. [5] incorporated few-shot learning (FSL) using Generative Adversarial Network (GAN) with Inception v3 to diagnose five different rare retinal diseases. Although their model gave an accuracy of 93.9%, it had few limitations such as producing low-resolution images while using GAN and not producing volumetric analysis of OCT images.

Li et al. [6] used the development retinal OCT image dataset to classify CNV, DME, and Drusen from OCT images. The authors used an ensemble of four classification models based on improved residual neural network (ResNet50) to achieve an accuracy of 96%.

Awais et al. [7] applied VGG model to extract features from different layerswhich was then utilized to classify DME

978-1-6654-2363-2/21/$31.00

from OCT images which were collected from Singapore Eye Research Institute (SERI). The authors performed operations such as noise removal and cropping the images to remove irrelevant features and smoothed the images. This particular model achieved 93% accuracy.

Gulshan et al. [8] applied CNN models to diagnose diabetic retinopathy and DME from retinal OCT images obtained from the EyePACS dataset and Messidor dataset. The model gave a sensitivity of 87% and specificity of 98.5%. Although the model gave good results, it was computationally expensive as it was trained with raw data and required a large amount of data to train.

Retinal disease classification is not only limited to deep learning frameworks, various researchers adopted traditional Machine Learning approaches for disease classification. Hussain et al. [9] used Random Forest to classify Spectral Domain OCT images of DME, AMD, and normal cases. They extracted the thickness of the retinal layers, the volume of pathologies such as Drusen and hyper-reflective intra-retinal spots as features through segmentation. The model achieved an accuracy of 95%.

Lemaitre et al. [10] extracted Local Binary Patterns (LBP) features and different mapping features from OCT images and used them in different linear and nonlinear classifiers to detect DME patients. The model achieved 81.2% Sensitivity and 93.7% Specificity.

Most of the mentioned researches were done with the aim to achieve more accurate OCT image classification. However, these researches are done with black box neural network models without much explanation of classification issues. Our research tries to address these gaps with proper classification explanation on one of the largest OCT datasets.

B. Models

Inception V3: It is a convolutional neural network (CNN) which was developed by Googlenet [11]. Inception V3 is an inheritor of Inception V1 with a total of 24M parameters. This network is capable of avoiding representation bottlenecks and has more systematic computation by using factorization methods.

ResNet50: This version of CNN solves the problem of saturation of accuracy by using shortcut or skip connections while building deeper models [12]. It is one of the early adopters of batch normalization having 26M parameters.

EfficientNetB0: This CNN provides improved efficiency by reducing the parameters to 5.3M and Floating Point Operations Per Second (FLOPS) [13].

III. Dataset Details

The dataset consists of 109309 images in total and it comes in the train set and the test set labelled by ophthalmologists. There are 108309 images in the train set while the test set consisted of 1000 images. The class distribution of the dataset is given in table 1.

In order to have a proper comparison with the original research, we kept the test set as it is. Meanwhile, we created a separate validation by taking 1000 images from the training set.

TABLE I
Label Distribution

Set Name	Classes			
	CNV	***DME***	***Drusen***	***normal***
Train	36955	11098	8366	50890
Validation	250	250	250	250
Test	250	250	250	250

The OCT images in the dataset are grayscale images. They have a deep black background with the macular region vertically centered. Some sample images are given in figure 1.

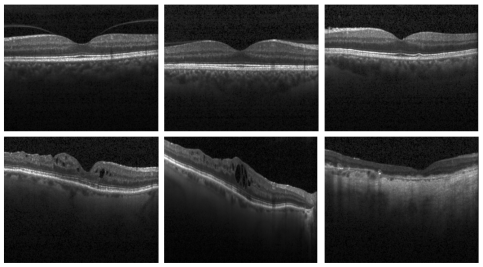

Fig. 1. Sample images from the dataset

IV. Proposed Model

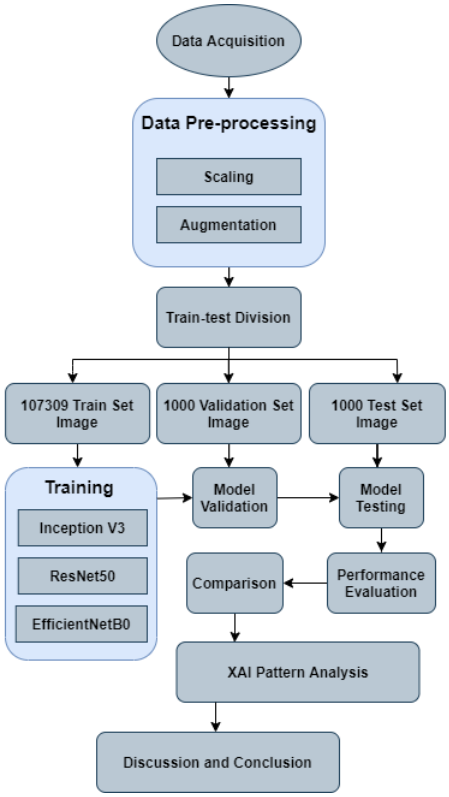

Fig. 2. Proposed workflow of the system

The proposed model is centered around the acquisition of the dataset, pre-processing the data, training deep learning models, evaluation of the performance of the deep learning models and the implementation of XAI based on the trained models. The entire proposed model is provided in figure 2.

First of all, data were collected and augmented. For augmentation, we mirrored the images and applied a slight height-width shift. We also had to scale the pixel values within 0-1 by using min-max scaling. For model training, we used Inception V3, ResNet50, and EfficientNetB0. At the top of each CNN model, we attached two fully connected layers consisting of 1024 and 512 neurons respectively. In order to avoid overfitting, we used 50% dropout within the two fully connected layers. Finally, the output layer had 4 neurons where each neuron corresponded to a single label.

In order to implement XAI, we used the LIME framework. The whole process was performed on a machine with Ryzen 3700X processor, 16 GB RAM, and RTX2070 super GPU.

V. Results and Analysis

After running 10 epochs, we collected data on training, test, and validation accuracy for the three models. In table number II, the training, validation, and test accuracy are provided.

TABLE II
Accuracy Scores

Set Name	Models		
	Inception V3	***EfficientNetB0***	***ResNet50***
Train	98.16%	96.50%	98.16%
Validation	96.00%	95.50%	96.20%
Test	95.10%	96.00%	96.90%

From table II, we can see that the 50 layers deep ResNet model achieved the highest accuracy of 96.9% and this accuracy is better than the maximum accuracy proposed in the base paper, which was 96.6%. While it can be argued that the difference falls within the margin of error, the accuracy was also achieved with just only 10 epochs while the base paper required much more prolonged simulation to achieve their score. The training and validation history graphs for the best performing model ResNet50 is provided in figure 3 and 4 respectively.

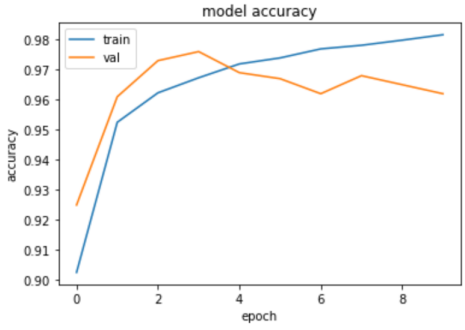

Fig. 3. Training accuracy for ResNet

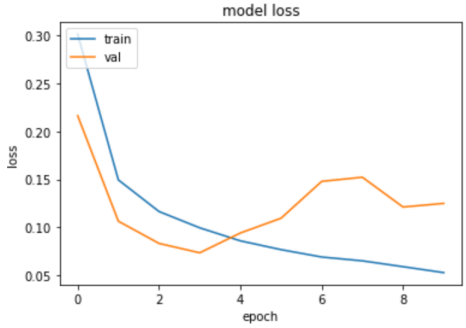

Fig. 4. Training loss for ResNet

In figure 3 and 4, we can see that at the beginning of the training, the training accuracy achieves a lower value compared to the validation accuracy and the training loss value is higher than the validation loss. This unusual behavior can easily be explained by data augmentation and dropout layer usage. The augmented data and the dropout layer during training made it much harder for the models to learn at the training phase. However, at one point the training accuracy catches up with the validation accuracy and eventually starts to overfit later down the line.

In order to get a better insight into the results, let's look at the confusion matrix given in figure 5.

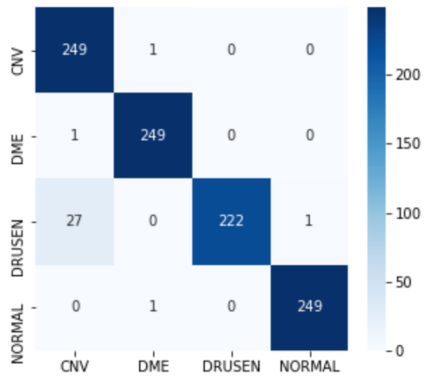

Fig. 5. Confusion matrix for ResNet

From the confusion matrix (Figure 5), it is pretty evident that the ResNet model excels at classifying CNV, DME, Drusen, and normal labeled images. However, it really struggles to classify Drusen, and most of the misclassifications came from the Drusen labeled images. Analyzing the dataset, this somewhat makes sense since Drusen has the lowest distribution among all the four labels and most of the misclassified images were labeled as CNV, which had a very high distribution in the dataset.

Finally, in order to get a better insight into the misclassifications, we implemented XAI in order to visualize the regions of the images that the CNNs used for classification. Some of the XAI analyses on the correctly classified images are given in figure 6.

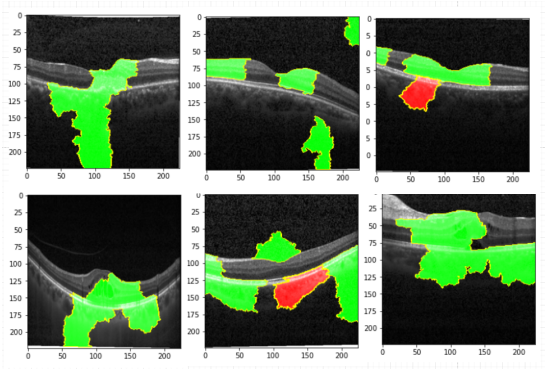

Fig. 6. XAI pattern on correctly classified images

In figure 6, six XAI marked images have been provided. The green marked regions are the ones that the neural networks correctly used for classification. Meanwhile, the red marked regions are the wrongly labeled portion by the neural network models. One interesting observation is the fact that the CNN models highlighted the background pixels on the top-left, top-middle, bottom-left, and bottom-middle images. This pattern is also be seen in the other XAI marked images. This is a clear sign of some overfitting because the background pixels do not carry any useful information. In order to check whether this pattern has any effect on the falsely classified images, we visualized XAI patterns on some misclassified Drusen images. Some sample representative results are provided in figure 7.

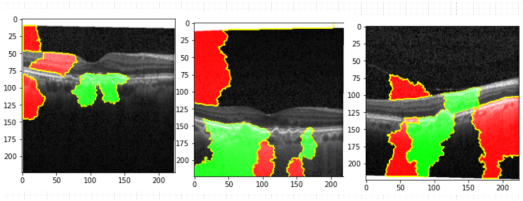

Fig. 7. XAI pattern on falsely classified images

From figure 7, it is clear that the background pixels have a significant impact on misclassification as most of the wrongly classified regions (Red marked area) came from the background pixels. Since this pattern is representative of other wrong classifications too, this is an important aspect to consider for the improvement of OCT classification accuracy.

Looking at the results and interpretations, some of our proposed ideas for future improvement are:

- Since Drusen has the lowest class distribution in the dataset and it has significantly more false negatives compared to other classes, weighting the classes according to their distribution during training may help.
- Overfitting on background pixels has a significant impact on classification results. If some form of background removal or background noise removal can be applied, then the results may improve.
- According to medical studies, choroid regions of an OCT gets most affected by retinal diseases. Therefore, some form of attention model that focuses on the Choroid region during classification may also improve the results.

VI. Conclusion and Future Works

In this paper, we have applied three neural network models namely Inception V3, ResNet50, and EfficientNetB0 in order to classify CNV, DME, Drusen, and normal OCT images. In our experiment, we have managed to achieve slightly more accuracy than the baseline results. However, most importantly, we have proposed some crucial analyses that can be utilized further in order to improve the results for OCT classification. In the future, our goal will be to incorporate the suggestions provided in this paper and improve the results as much as possible.

References

[1] W. L. Wong, X. Su, X. Li, C. M. G. Cheung, R. Klein, C.-Y. Cheng, and T. Y. Wong, "Global prevalence of age-related macular degeneration and disease burden projection for 2020 and 2040: a systematic review and meta-analysis," *The Lancet Global Health*, vol. 2, no. 2, pp. e106–e116, 2014.

[2] J. B. Saaddine, A. A. Honeycutt, K. V. Narayan, X. Zhang, R. Klein, and J. P. Boyle, "Projection of diabetic retinopathy and other major eye diseases among people with diabetes mellitus: United states, 2005-2050," *Archives of ophthalmology*, vol. 126, no. 12, pp. 1740–1747, 2008.

[3] D. S. Kermany, M. Goldbaum, W. Cai, C. C. Valentim, H. Liang, S. L. Baxter, A. McKeown, G. Yang, X. Wu, F. Yan *et al.*, "Identifying medical diagnoses and treatable diseases by image-based deep learning," *Cell*, vol. 172, no. 5, pp. 1122–1131, 2018.

[4] M. Shah, A. Roomans Ledo, and J. Rittscher, "Automated classification of normal and stargardt disease optical coherence tomography images using deep learning," *Acta ophthalmologica*, vol. 98, no. 6, pp. e715–e721, 2020.

[5] T. K. Yoo, J. Y. Choi, and H. K. Kim, "Feasibility study to improve deep learning in oct diagnosis of rare retinal diseases with few-shot classification," *Medical & Biological Engineering & Computing*, vol. 59, no. 2, pp. 401–415, 2021.

[6] F. Li, H. Chen, Z. Liu, X.-d. Zhang, M.-s. Jiang, Z.-z. Wu, and K.-q. Zhou, "Deep learning-based automated detection of retinal diseases using optical coherence tomography images," *Biomedical optics express*, vol. 10, no. 12, pp. 6204–6226, 2019.

[7] M. Awais, H. Müller, T. B. Tang, and F. Meriaudeau, "Classification of sd-oct images using a deep learning approach," in *2017 IEEE International Conference on Signal and Image Processing Applications (ICSIPA)*. IEEE, 2017, pp. 489–492.

[8] V. Gulshan, L. Peng, M. Coram, M. C. Stumpe, D. Wu, A. Narayanaswamy, S. Venugopalan, K. Widner, T. Madams, J. Cuadros *et al.*, "Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs," *Jama*, vol. 316, no. 22, pp. 2402–2410, 2016.

[9] M. A. Hussain, A. Bhuiyan, C. D. Luu, R. Theodore Smith, R. H. Guymer, H. Ishikawa, J. S. Schuman, and K. Ramamohanarao, "Classification of healthy and diseased retina using sd-oct imaging and random forest algorithm," *PloS one*, vol. 13, no. 6, p. e0198281, 2018.

[10] G. Lemaître, M. Rastgoo, J. Massich, C. Y. Cheung, T. Y. Wong, E. Lamoureux, D. Milea, F. Mériaudeau, and D. Sidibé, "Classification of sd-oct volumes using local binary patterns: experimental validation for dme detection," *Journal of ophthalmology*, vol. 2016, 2016.

[11] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 2818–2826.

[12] 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*, 2016, pp. 770–778.

[13] M. Tan and Q. Le, "Efficientnet: Rethinking model scaling for convolutional neural networks," in *International Conference on Machine Learning*. PMLR, 2019, pp. 6105–6114.