Disease Detection from Retinal OCT Images

Arda Uyaroğlu[1]

[1] Artificial Intelligence and Data Engineering, Ankara University, Ankara, Türkiye, ORCID: 0009-0000-4771-4682
* Corresponding author: Arda Uyaroğlu (auyaroglu@ankara.edu.tr)
Received: 30 November 2025, Accepted: 18 December 2025, Published: 19 December 2025

Abstract: In this study, three common Convolutional Neural Network (CNN) architectures—AlexNet, VGG16, and ResNet50—are compared to classify retinal OCT (Optical Coherence Tomography) images into four classes: CNV, DME, DRUSEN, and NORMAL using the OCT2017 dataset. Each model is analyzed using evaluation metrics including accuracy, precision, F1 score, and ROC AUC. The main objective is to reveal the performance differences of various CNN architectures in medical image classification tasks and to determine their potential contribution to the computer-aided diagnosis of retinal diseases. Experimental findings show that all models achieve high success rates; specifically, the VGG16 model achieved the highest performance with 99.70% test accuracy and a 99.70% F1 score. Architectures with deep and residual connections can learn complex textural structures more effectively. These results suggest that CNN-based models provide a reliable basis for advanced diagnostic systems to be developed for automated OCT analysis.

Keywords: Deep Learning; Retinal OCT; CNN; Image Classification

1. Introduction

Retinal imaging techniques are of great importance for the early diagnosis of eye diseases. One of the most used methods in the diagnosis of retinal diseases is Optical Coherence Tomography (OCT). OCT technology is a non-invasive method used to image the detailed structure of the retinal layer.
The success of deep learning in the field of image processing has also begun to be applied in medical image analysis. Many studies have been conducted on OCT images in literature. Kermany et al. (2018) achieved high-accuracy classification with the OCT2017 dataset. De Fauw et al. (2018) presented a significant application by integrating deep learning into clinical-level decision support systems.
Since existing studies generally evaluate only a single architecture, analyses where different architectures are compared together are limited. Furthermore, many models have high hardware requirements, which may restrict their field of application. For this reason, there is a need for a detailed evaluation of different CNN architectures in terms of both accuracy and resource usage.
With this study:

(1) AlexNet, VGG16, and ResNet50 architectures were compared in four classes (CNV, DME, DRUSEN, NORMAL) using the OCT2017 dataset. (This dataset contains 242 test images per class, 8 validation images, and a total of 37,205 CNV, 11,348 DME, 8,616 Drusen, and 26,315 Normal training images).

(2) Analyses were performed using metrics such as accuracy, F1 score, MCC, and training time.

(3) Tests were conducted in both powerful (A100 Google Colab GPU) and limited (GTX 960M) hardware environments.

(4) Additionally, an 8-class classification was performed using a second dataset named Retinal OCT C8 (Obulisainaren, 2023a), and experimental comparisons were presented with a different model group (VGG11, ResNet34, AlexNet).

The remainder of this paper is organized as follows: Section 2 details the materials and methods, including dataset characteristics and network architecture. Section 3 presents the experimental results

and discusses the performance comparison of the models with literature findings. Finally, Section 4 concludes the study and summarizes the key findings.

2. Materials and Methods

In this section, the methodology utilized for the classification of retinal OCT images is detailed. Firstly, the characteristics of the data sets used in the study are described. Secondly, the structural features of the selected Convolutional Neural Network (CNN) architectures—AlexNet, VGG16, and ResNet50—are explained. Finally, the experimental setup, including hardware specifications and training parameters, is mentioned to provide a comprehensive understanding of the analytical framework

2.1 Dataset

The OCT2017 (Mooney, 2018) dataset, obtained from Kaggle, consists of four classes: CNV, DME, DRUSEN, and NORMAL. The dataset contains a total of 84,495 images. The training set is distributed as follows: 37,205 CNV, 11,348 DME, 8,616 DRUSEN, and 26,315 NORMAL images. The test set includes 242 images per class, and the validation set consists of 8 images per class. The images are divided into training, validation, and test subsets to ensure a solid evaluation.

2.2 CNN Architectures

AlexNet, VGG16, and ResNet50 architectures were compared. AlexNet, due to its relatively shallower structure and fewer parameters, provides fast learning and can operate efficiently even in limited hardware environments (Krizhevsky et al., 2012).
In contrast, VGG16 allows for the detailed extraction of low- and mid-level visual features with its deeper and regular layer structure; therefore, it can learn subtle textural differences in medical images more stably (Simonyan & Zisserman, 2014).
ResNet50, on the other hand, largely eliminates the vanishing gradient problem that occurs in deep networks thanks to residual connections and allows much deeper architectures to be trained with high accuracy (He et al., 2016).
Evaluating these three architectures together presents an analysis of how different depth and parameter structures affect performance on complex medical images like OCT and provides clues as to which model might be more suitable in which scenario.

2.3 Training Parameters and Environment

All models were trained on Google Colab (A100 GPU) and local (GTX 960M) devices. Images were resized to 224x224, and the Adam optimization algorithm was used for training. The learning rate was set to 0.0001, and the batch size was selected as 256 or 64. The early stopping method was applied to prevent overfitting. Training was initially scheduled for 15 epochs but was completed in 5 epochs due to the early stopping method.

3. Results and Discussion

3.1 Training Times and Performance

Results were obtained in the following durations using the Google Colab A100 GPU environment:VGG16 ~747 seconds, ResNet50 ~607 seconds, AlexNet ~596 seconds. As model size and parameter count increased, accuracy improved, but training time and hardware requirements also increased. The classification performance and class-wise accuracy distributions for each model are illustrated in the confusion matrices shown in Figure 1, Figure 2, and Figure 3 for VGG16, ResNet50, and AlexNet, respectively.

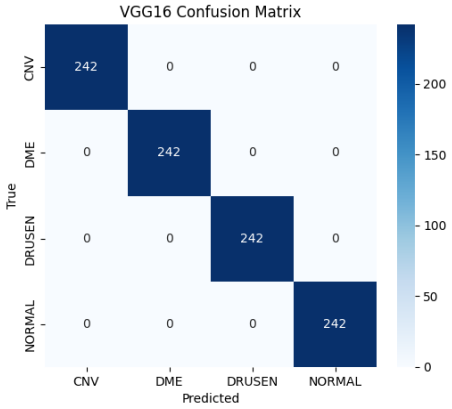

Figure 1. Confusion Matrix for VGG16

This matrix demonstrates that the VGG16 model achieved perfect classification performance, correctly predicting all 242 test images across every category (CNV, DME, DRUSEN, NORMAL) without any misclassification.

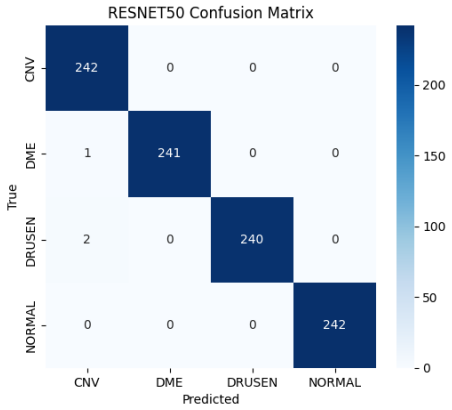

Figure 2. Confusion Matrix for ResNet50

The confusion matrix indicates high accuracy for the ResNet50 architecture. While the model is highly stable, minor misclassifications were observed, particularly distinguishing DME and DRUSEN classes from CNV.

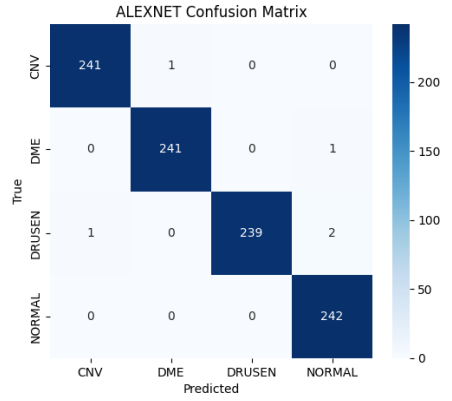

Figure 3. Confusion Matrix for AlexNet

This figure illustrates the classification results for AlexNet. Although the model reached approximately 99% accuracy, it exhibits slightly more deviation and false positives compared to the deeper VGG16 and ResNet50 architectures.

Table 1. Performance Comparison of Models

Model	Test Accuracy	Avg Val Acc	Worst Val Loss	Training Time (s)	Parameters	Size (MB)
VGG16	1.0	0.98	0.097	747.29	134276932	512.24
ResNet50	0.9969	1.0	0.022	607.13	23516228	90.01
AlexNet	0.9948	0.99	0.104	596.	57020228	217.52

VGG16 and ResNet50 offer excellent accuracy and stability, while AlexNet achieves results very close to this success with low resource consumption. Although accuracy may increase as model size increases, hardware needs also rise at the same rate. We can see from Table 1 that AlexNet establishes this balance well. Although the average accuracy and test accuracy are lower than other models, training time is directly related to the model's depth and parameter count; AlexNet has shown that it can be a good model within the possibilities given its parameter count and time.

3.2 Literature Comparison

In this study, deep learning-based classification was performed using two different retinal imaging datasets: OCT2017 (Mooney, 2018) and Retinal OCT C8 (Obulisainaren, 2023a). The C8 dataset was used to test the validity of the study with a different set, given that the results of the first dataset yielded high accuracy values. The results of these studies, where different architectures were used, were compared in detail with similar approaches in the literature.

3.2.1 Comparison of OCT2017 Based Studies

OCT2017 is a common dataset consisting of four classes: CNV, DME, DRUSEN, and NORMAL. In this study, training was performed with VGG16, AlexNet, and ResNet50 architectures, and the highest success was achieved with the VGG16 model.

The model, trained for 5 epochs, reached 99.70% accuracy, 99.70% F1 score, and 0.996 MCC value on the test set. In the study by Barbano (2022), the VGG16 model was trained for only 10 epochs, and 99.28% test accuracy was obtained with a transfer learning approach. However, detailed metrics, training time, or hardware information were not presented in that study. Furthermore, when compared with this study, it was concluded that there was no data leakage or error in the data, as the dataset also gave high results in other generally used studies. Khan (2023) trained the VGG16 model for 25 epochs on the same dataset and reported a test accuracy of 95.17%. Metrics and hardware details remained limited. A detailed comparison of these results with the findings of the current study is presented in Table 2.

Table 2. Result Comparison

Model	Test Acc	Val Acc	Epoch	F1 Score	Source
VGG16	1.0	0.9875	5	0.997	This work
VGG16	0.9928	-	10	-	A. Khan
VGG16	0.9517	-	25	-	Obulisainaren

In this study, high success was achieved with a short training time, and analysis was presented with metrics such as ROC curves (Figure 4), hardware comparisons, MCC, and F1.

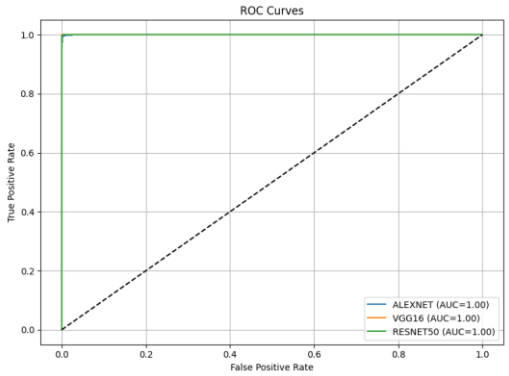

Figure 4. ROC Curve Comparison for Three Models

All metrics achieved a maximum score (1.00). Models exhibited high success in metrics such as MCC, F1, and Accuracy. AlexNet reached levels of 99% with very little deviation.

3.2.2 Comparison of Retinal OCT C8 Based Studies

The Retinal OCT C8 dataset is more complex for classification as it contains eight classes. In this study, VGG11, ResNet34, and AlexNet architectures were tested; 92.61% accuracy, 92.64% F1 score, and 0.924 MCC were obtained on the test set with the VGG11 model. In the study conducted by Obulisainaren (2023b), training was performed with the VGG19 architecture, and test accuracy was reported as 89.79%. However, F1 score, MCC, training time, or hardware information was not shared in the study. In this study, not only was higher success achieved, but a more comprehensive evaluation was also presented with training times, ROC analyses, and metric diversity. Additionally, considering the accuracy results of other models, it was revealed that the VGG model worked very well, but AlexNet and ResNet could not provide sufficient performance, as shown in Table 3.

Table 3. Performance Comparison of Models

Model	Test Acc	Avg Val Acc	Precision	F1 Score	ROC AUC	Training Time (s)	Parameters	Size (MB)
VGG11	0.92	0.91	0.93	0.92	0.95	101.17	128799112	491.34
ResNet34	0.75	0.75	0.88	0.75	0.86	102.16	21288776	81.35
AlexNet	0.57	0.58	0.79	0.59	0.75	101.76	57036616	217.38

4. Conclusion

As a result of this study, it was observed that AlexNet, VGG16, and ResNet50 models classify OCT images with high accuracy. When the first dataset (Mooney, 2018) is considered, VGG16 and ResNet50 models operate with nearly 100% accuracy on the test set, while the AlexNet model reached an accuracy level of approximately 98–99%. This indicates that CNN-based approaches are effective in medical image analysis.

In terms of hardware, while VGG16 presents higher resource requirements, AlexNet offers a suitable solution for limited resource environments by providing reasonable accuracy with low resource

consumption. ResNet50 stands out as the structure that best provides the balance between performance and efficiency.

In this study, additionally, the performance of VGG11, ResNet34, and AlexNet models on 8-class OCT images were evaluated using the 'Retinal OCT C8' dataset (Obulisainaren, 2023a) found on Kaggle. This dataset includes Retina Pigment Epithelium Detachment (RPED), Macular Hole (MH), Epiretinal Membrane (ERM), and Central Serous Retinopathy (CSR) classes in addition to CNV, DME, DRUSEN, and NORMAL classes.
Based on this, when the second dataset (Obulisainaren, 2023a) was used, it was revealed that the VGG11 model gave a good result, whereas ResNet34 and AlexNet could not actually provide sufficient performance. It was understood that the main reason for this situation was the normalization rate in the parameters used.
Additionally, considering the datasets, the very small amount of validation data (32 images) in the Mooney (2018) set compared to train and test sets, and the excess of images in the train part of the first set were observed to affect the accuracy rate. When the Obulisainaren (2023a) dataset was used, it was realized that the parameters were not suitable for AlexNet and ResNet, and that distributing the train, test, and validation parts of the 2nd dataset more proportionally (18,400 train images, 2,800 test and 2,800 validation images) influenced the change in model accuracies.
Finally, the main reason why accuracy rates were high in the Mooney (2018) set compared to values in other studies is that the images taken from different hospitals and countries are directly high-quality tomography images that show healthy and disease visuals clearly and cleanly, without any effects or manipulation on the image.

Funding: This research received no external funding.
Conflicts of Interest: The authors declare no conflict of interest.

References

Barbano, C. A. (2022). VGG16 transfer learning with PyTorch. Kaggle. https://www.kaggle.com/code/carloalbertobarbano/vgg16-transfer-learning-pytorch

De Fauw, T., Ledsam, J. R., Romera-Paredes, B., Nikolov, S., Tomasev, N., Blackwell, S., … Ronneberger, O. (2018). Clinically applicable deep learning for diagnosis and referral in retinal disease. *Nature Medicine*, 24(9), 1342–1350. https://doi.org/10.1038/s41591-018-0107-6

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (pp. 770–778). https://doi.org/10.1109/CVPR.2016.90

Kermany, D. S., Goldbaum, M., Cai, W., Valentim, C. C., Liang, H., Baxter, S. L., … Zhang, K. (2018). Identifying medical diagnoses and treatable diseases by image-based deep learning. *Cell*, 172(5), 1122–1131.e9. https://doi.org/10.1016/j.cell.2018.02.010

Khan, A. (2023). Retina damage classification with 95% accuracy. Kaggle. https://www.kaggle.com/code/arbazkhan971/retina-damage-classification-95-accuracy

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). ImageNet classification with deep convolutional neural networks. *Advances in Neural Information Processing Systems*, 25, 1097–1105.

Mooney, P. T. (2018). Kermany2018 retinal OCT dataset. Kaggle. https://www.kaggle.com/datasets/paultimothymooney/kermany2018

Obulisainaren. (2023a). Retinal OCT C8 – 8 class retinal OCT image dataset. Kaggle. https://www.kaggle.com/datasets/obulisainaren/retinal-oct-c8

Obulisainaren. (2023b). Retinal OCT VGG19 training notebook. Kaggle. https://www.kaggle.com/code/obulisainaren/retinal-oct-vgg19

Rajalakshmi, R., Subashini, R., Anjana, H., & Mohan, V. (2020). Application of artificial intelligence in retinal disease screening. *Journal of Diabetology*, 11(2), 55–59. https://doi.org/10.4103/jod.jod_34_20

Simonyan, K., & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. arXiv. https://arxiv.org/abs/1409.1556