

OCT Retinal Scan Classification Using Machine Learning

Amna Hassan

Abstract

Optical Coherence Tomography (OCT) scans are non-invasive images that use light waves to take detailed pictures of the retina. These images can detect a variety of optical diseases that can affect the vision and function of the eye. Our project aims to compare different machine learning models in their ability to classify retinal images. We want to provide support to ophthalmologists by finding an algorithm that can produce consistent accurate diagnosis. In this project, we classified the OCT scans into four classes of diseases: Normal, Drusen, Choroidal Neovascularization (CNV), and Diabetic-Related Macular Edema (DME). We used a Logistic Regression, MLP Classifier, convolutional neural network (CNN) with a VGG16 architecture, and compared each on their testing accuracy score. We found the Logistic Regression gave us an accuracy of 52%, the MLP Classifier gave us 73%, and the highest, CNN, with 98% accuracy. These results show that convolutional neural networks can efficiently classify OCT retinal scans accurately. The successful results from using a pre-trained architecture like VGG16 suggests that transfer learning is sufficient in separating the four classes in this dataset. However, before implementing in clinical settings, the model would require validation using scans from different patient populations and imaging devices.

1 Introduction

Retinal diseases are a main cause of vision impairment making early classification and diagnosis important for necessary treatment. Optical Coherence Tomography (OCT) images are non-invasive images used by many optometrists and ophthalmologists to take pictures of a patient's retina. It measures the thickness of the retina which is used to diagnose different retinopathic diseases [1]. Interpreting and understanding these images requires clinical knowledge and experience. As the use of OCT scans in healthcare settings increases, machine learning models have the potential to offer additional diagnostic support for eye-care specialists.

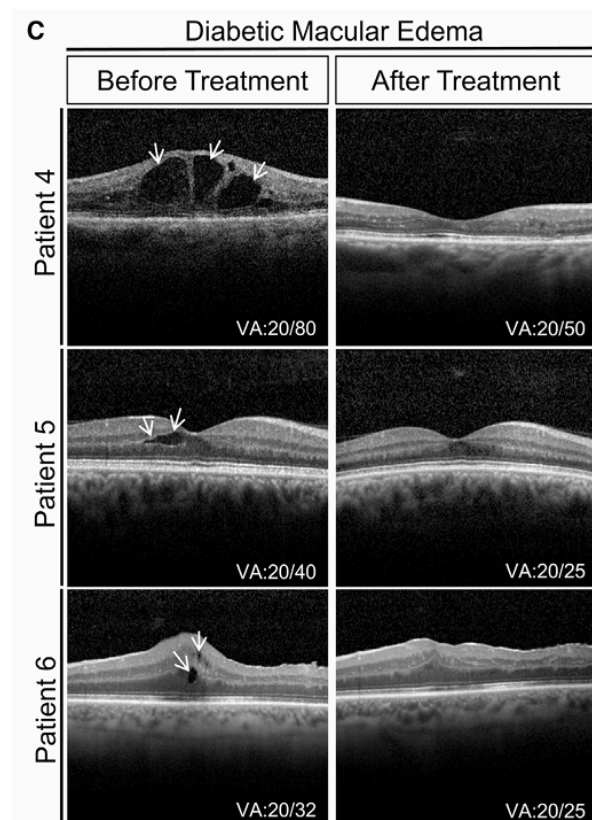
The application of artificial intelligence and machine learning in healthcare is a growing concept due to its potential to aid healthcare professionals by improving diagnosis accuracy. The emerging integration is already set to improve detection and diagnostic accuracy in daily medical settings [8]. Comparing models with different architectures can highlight how these algorithms can perform on classification tasks and which ones are better suited for possible use in a medical setting. For example, a linear model like logistic regression model cannot identify relationships between neighboring pixels, while a convolutional neural network (CNN) is known for learning spatial patterns in images [12] [13]. A recent review of the use of artificial intelligence in retinal screening using OCT images found that pre-trained networks like deep-learning models were practical and useful when working with the classification of complex images

like OCT scans [8]. Additionally, OCT technology has applications beyond ophthalmology. This included cardiovascular imaging, where it is used to examine the structure of coronary arteries [2].

In this project, we compare how accurately different machine learning models can classify OCT images and which model can provide the highest classification accuracy. Since we were working with a dataset of labeled retina images, this project would fall under the category of supervised machine learning involving a multi-class image classification. The input data is the OCT retinal images and the output is a predicted class label corresponding to each image which falls into one of the four diagnostics. By comparing the performance and accuracy scores of classification models, we aim to identify the model which provides the most accurate classification of the retinal diseases using the scans.

2 Background

The research paper [9] uses a transfer learning based CNN to classify retina diseases from a dataset of OCT scans. They used a pretrained Inception V3 architecture and retrained it by initializing convolution and softmax layers. Similar to our project, it was trained to classify images from four classes: Normal, Drusen, CNV, and DME. Using a dataset including over 108,000 scans, they accomplished an accuracy of 96.6% making its performance comparable to expert ophthalmologists. A unique feature used in the research paper were occlusion maps. These maps highlight the locations on the scans that influenced the models prediction. The added feature proved the model was focusing on relevant features and made the models diagnosis interpretable showing how a CNN-based model can identify important visual-based features on an OCT scan and accurately classify the disease presented.



3 Dataset

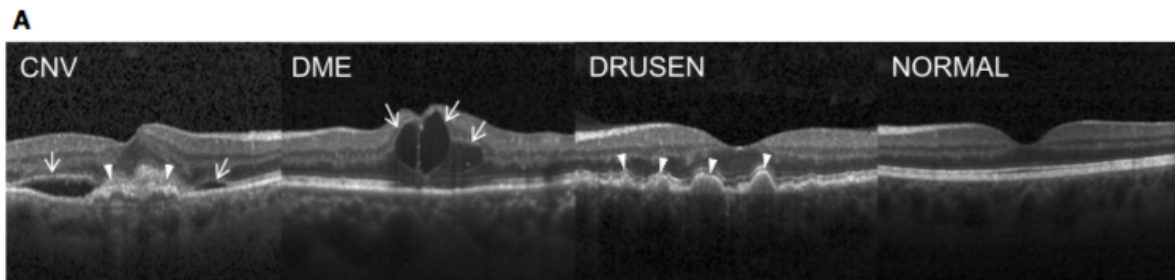
The dataset we used for this project consisted of OCT retina images, making it a vision-based dataset [6] [7]. It was published in January 2018 on Mendeley Data by the same team who used transfer learning in OCT classification which was discussed earlier [9]. The dataset is organized into training and testing folders. In total, the dataset contains 84,484 images, consisting of 83,484 training images and 1,000 testing images. For the VGG16 model, 20% of the training data was used as a validation set during training through the `validation_split` parameter.

Our dataset included 4 types of images from the OCT scans: Normal, Drusen, Choroidal Neovascularization (CNV), and Diabetic-Related Macular Edema (DME) images. Normal images represented scans of normal healthy retinas. Drusen are lipid and protein filled deposits found under the retina which can cause blurred vision and difficulty seeing in some circumstances [3]. CNV is a retinal condition which is due to an abnormal growth of blood vessels into the retina [4]. Finally, DME is a diabetes-related condition which causes swelling in the center of the retina. It can cause difficulty in vision and if untreated can cause vision loss [5].

Before training the model, a few preprocessing steps were performed. First, each image was converted into grayscale to reduce complexity. Since OCT images encode retinal tissue structure rather than color based information, converting the scans to greyscale helped simplify the input without losing important features [10]. Next, all the images were resized to 100x100 to ensure an identical input size. The images were then encoded into numerical pixel values and flattened into vectors. The pixel values were normalized to a range of 0 to 1 by dividing each pixel value by 255 to improve training performance. The labels of each class were also encoded into numeric values for classification. The features in this dataset are the visual properties shown in each OCT image. They are very important as these patterns help the model to differentiate between the four classes.

The large amount of labeled images in the dataset and preprocessing make the images well suited to be used in training an accurate classification model.

For better understanding of each class in the dataset, example images of the four classes are attached below.



4 Models

The first model applied to the dataset was a Logistic Regression model using a Stochastic Gradient Descent (SGD) algorithm. This algorithm updates the model's weights using portions of training data and is mostly used for large datasets [11]. With an iteration of 500, the pre-processed training data was used to train the model with early stopping and regularization to reduce overfitting. After training completed, we measured the results by producing an accuracy score, a confusion matrix, and a classification report.

The second model used was a Multi-Layer Perceptron (MLP) Classifier. The neural network used a hidden layer with 100 neurons and used ReLU activation function to accurately learn the patterns in the OCT retina images. The Adam Optimizer adjusted the model's weights while training. Similar to the logistic regression model, the MLP was graded on the accuracy, confusion matrix, and classification report.

The third model created was a Convolution Neural Network (CNN) based on a VGG16 architecture. The model used parameters and pre-trained ImageNet weights from TensorFlow which allowed it to utilize features learned from the large dataset [13]. We resized the input images 224 x 224 and converted them from greyscale to a 3-channel format to meet the model requirements. The pre-trained convolution layers were frozen to retain the learned features and additional layers were added for classification. These layers included Average Pooling, Batch Normalization, Dropout, and a Dense layer with a softmax activation function to classify images into the four class categories. The model was trained on 20 epochs using a batch size of 32. We additionally used the Adam optimizer with sparse categorical cross-entropy as the loss function.

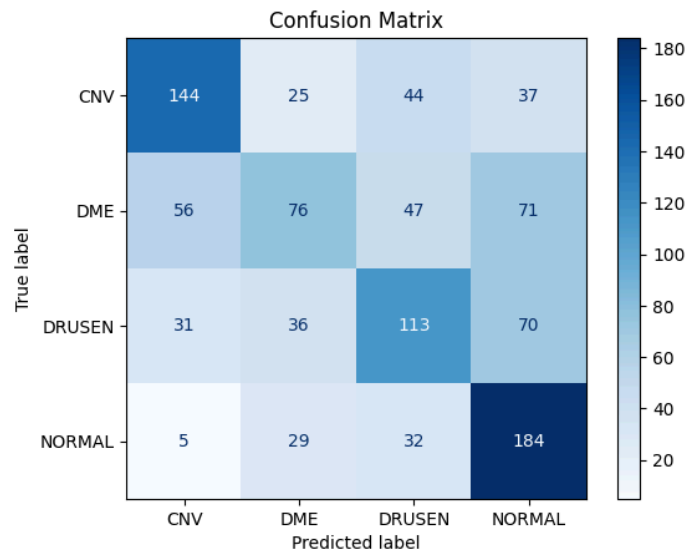
5 Results

After training and testing the three models, we based their performance by comparing each of their testing accuracy, confusion matrix, and classification report. The classification report included precision, recall, and F1 score values. Precision measured how often the model's predictions were correct while recall measured how well the model detected the images that belonged to each class. The F1 score is a score that joins precision and recall using the formula: $2 \times (\text{Precision} \times \text{Recall}) / (\text{Precision} + \text{Recall})$ [14].

1) The Logistic Regression model achieved a test accuracy score of 52%. As seen on the confusion matrix below, the classes which were most frequently predicted correctly were CNV and Normal. Our classification report included these values for each class:

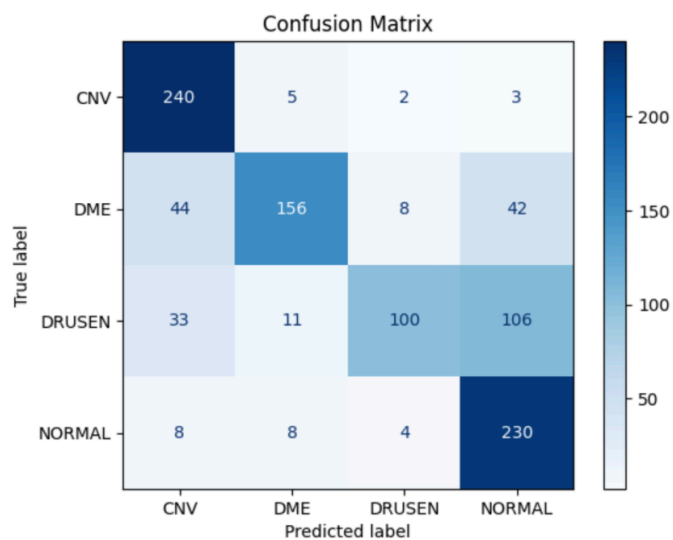
class	precision	recall	f1-score
CNV	0.61	0.58	0.59

DME	0.46	0.30	0.37
DRUSEN	0.48	0.45	0.47
NORMAL	0.51	0.74	0.60



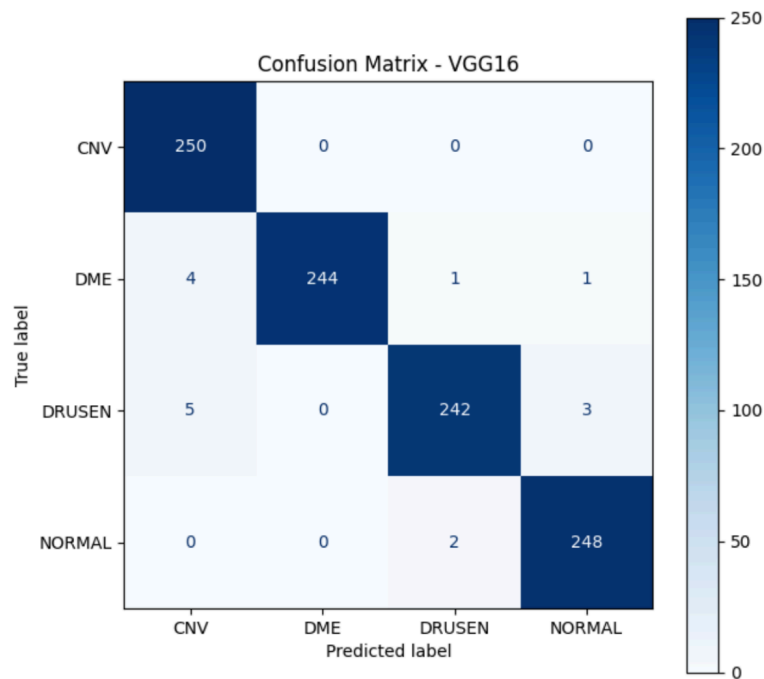
2) The MLP classifier achieved a test accuracy of 73% produced a 21 percentage improvement over the Logistic Regression model. The confusion matrix shows that CNV had the highest number of correctly classified predictions at 240/250 images, followed by Normal with 230/250 correct predictions.

class	precision	recall	f1-score
CNV	0.74	0.96	0.83
DME	0.87	0.62	0.73
DRUSEN	0.88	0.40	0.55
NORMAL	0.60	0.92	0.73



3) The VGG16 CNN achieved an accuracy of 98% making it the highest accuracy out of the three models evaluated. The confusion matrix below shows that out of 250 images from each class, CNV was correctly predicted in 250/250 images, DME in 244/250, Drusen in 242/250, and Normal in 248/250 images.

class	precision	recall	f1-score
CNV	0.97	1.00	0.98
DME	1.00	0.98	0.99
DRUSEN	0.99	0.97	0.98
NORMAL	0.98	0.99	0.99



6 Discussion

Since the logistic regression model is a linear classifier, it assumes relationships between pixel values and the classes can be separated by a straight line, making it unable to capture complex visual patterns in the OCT retinal scans. The Logistic Regression model input requirements are 1D feature vectors which require us to flatten the 2D retinal images. This preprocessing causes the model to lose awareness of how the pixels relate to each other and their overall structure.

The improvement from the MLP Classifier shows that the model was able to learn more complex non-linear relationships in the retinal images than the Logistic Regression. Its multi-layer structure

allowed the model to identify patterns that the simpler linear-model missed. However, since the model used flattened input images, it was unable to properly capture the spatial relationships between neighboring pixels. This limitation may have contributed to the 73% accuracy the model achieved.

The MLP classifier struggled to identify the Drusen class, achieving a precision of 0.88 and a recall of 0.40. This means the model only identified 40% of the actual Drusen images correctly. One possible explanation for this imbalance can be the dataset. Drusen had approximately 8,600 images compared to 37,000 CNV images. Resizing the images to 100x100 pixels could have also caused some visual details in the Drusen images to be lost, making its classification more difficult.

The 98% accuracy achieved by the VGG16 CNN model shows its ability to identify complex visual patterns from the OCT scans. CNN's are specifically designed for visual data classification and can preserve the spatial arrangement of neighboring pixels [14]. This model used pretrained ImageNet weights to extract visual patterns when applied to the classification of OCT retinal images. This likely contributed to the model's ability to accurately distinguish between the 4 classes.

Comparing the three models reflected differences in how image information is processed. The logistic regression and MLP classifier both used flattened pixel values while the VGG16 extracted features using its pretrained frozen convolutional layers. Since the convolutional layers were frozen, this experiment thus observed how accurately pretrained ImageNet features could be applied to the classification of OCT images without retraining the layers for specifically retinal images.

7 Conclusion

In this project, we investigated how machine learning could be used to accurately classify retinal scans to assist health care professionals in diagnosing retinal diseases. The Logistic Regression, MLP Classifier, and CNN models were used to classify the OCT retinal scans into four diagnosis classes: CNV, DME, Drusen, and Normal. These models were trained using 83,484 images and evaluated on 1,000 testing images. The data was pre-processed and used to train the classification models. From the models tested, the CNN achieved the highest accuracy in classification with a score of 98.4% with the help of a pre-trained VGG16 architecture and image pre-processing. In the future we can use a larger data set of OCT scans to test our model and observe its performance. We can also explore how this architecture can be used to classify scans from different medical areas.

8 Acknowledgments

I would like to thank Erick Siavichay for his help throughout this project as well as Inspirit AI for providing this opportunity.

9 References

- [1] Aumann, S., Donner, S., Fischer, J., & Müller, F. (2019, August 14). *Optical Coherence Tomography (OCT): Principle and Technical Realization* (J. F. Bille, Ed.). PubMed; Springer.
<https://www.ncbi.nlm.nih.gov/books/NBK554044/>
- [2] Castaldi, G., Zormpas, G., Frederiks, P., Adriaenssens, T., & Bennett, J. (2025). The Rise of Optical Coherent Tomography in Intracoronary Imaging: An Overview of Current Technology, Limitations, and Future Perspectives. *Reviews in Cardiovascular Medicine*, 26(8). <https://doi.org/10.31083/rcm38123>
- [3] California Retina ConsultantsTM | Retina Specialists in Central CA. (2026, March 25). What Are Drusen? Understanding a Common Early Sign of Macular Degeneration. California Retina; California Retina ConsultantsTM | Retina Specialists in Central CA.
<https://www.californiaretinacare.com/blog/what-are-drusen-understanding-a-common-early-sign-of-macular-degeneration>
- [4] Choroidal Neovascularization: OCT Angiography Findings - EyeWiki. (2026, August 28). Eyewiki.Org. https://eyewiki.org/Choroidal_Neovascularization:_OCT_Angiography_Findings#Disease
- [5] What is Diabetes-Related Edema (DME)? (2023). Cleveland Clinic.
<https://my.clevelandclinic.org/health/diseases/24733-diabetes-related-macular-edema>
- [6] Retinal OCT Images (optical coherence tomography). (n.d.). WwW.Kaggle.Com. Retrieved September 2, 2026, from <https://www.kaggle.com/datasets/paultimothymooney/kermany2018>
- [7] Kermany, D., Zhang, K., & Goldbaum, M. (2018). Labeled Optical Coherence Tomography (OCT) and Chest X-Ray Images for Classification. *Data.Mendeley.Com*, 2. <https://doi.org/10.17632/rscbjbr9sj.2>
- [8] Muhammed Halil Akpinar, Abdulkadir Sengur, Faust, O., Tong, L., Molinari, F., & U. Rajendra Acharya. (2024). Artificial Intelligence in Retinal Screening Using OCT Images: A Review of the Last Decade (2013-2023). *Computer Methods and Programs in Biomedicine*, 254, 108253–108253.
<https://doi.org/10.1016/j.cmpb.2024.108253>
- [9] Kermany, D. S., Goldbaum, M., Cai, W., Valentim, C. C. S., Liang, H., Baxter, S. L., McKeown, A., Yang, G., Wu, X., Yan, F., Dong, J., Prasadha, M. K., Pei, J., Ting, M. Y. L., Zhu, J., Li, C., Hewett, S., Dong, J., Ziyar, I., ... 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>
- [10] Brar, M., Bartsch, D.-U. G., Nigam, N., Mojana, F., Gomez, L., Cheng, L., Hedaya, J., & Freeman, W. R. (2009). Colour versus grey-scale display of images on high-resolution spectral OCT. *British Journal of Ophthalmology*, 93(5), 597–602. <https://doi.org/10.1136/bjo.2008.146233>

[11] 1.5. Stochastic Gradient Descent — scikit-learn 0.23.2 documentation. (n.d.). Scikit-Learn.Org. Retrieved September 2, 2026, from <https://scikit-learn.org/stable/modules/sgd.html>

[12] scikit-learn. (n.d.). 1.1. Linear Models — scikit-learn 0.22.2 documentation. Scikit-Learn.Org. Retrieved September 2, 2026, from https://scikit-learn.org/stable/modules/linear_model.html#logistic-regression

[13] TensorFlow. (n.d.). *VGG16*. TensorFlow API documentation. https://www.tensorflow.org/api_docs/python/tf/keras/applications/VGG16

[14] GeeksforGeeks. (2023, December 27). F1 Score in Machine Learning. GeeksforGeeks. <https://www.geeksforgeeks.org/machine-learning/f1-score-in-machine-learning/>