Final Report: OCT Scans Disease Classification

Kristen Lee

[redacted]

1 Introduction, Illustration, and Background/Related Work

Age-related macular degeneration (AMD) and other retinal diseases are major causes of vision loss worldwide (1). OCT scans are routinely used to detect early signs of these conditions, but interpreting them is time-consuming and can vary between clinicians because disease features are often subtle (2; 3). This project aims to develop a deep learning model that classifies OCT scans into four categories: AMD, DME, DR, NO (Normal). As shown in Figure 1, the model takes an OCT image as input and outputs the predicted condition. Automated OCT classification can speed up screening, improve diagnostic consistency, and reduce clinical workload. Deep learning is particularly suitable for this task because it can learn complex, hierarchical patterns from high-dimensional OCT images and generalize across patients and imaging settings (4).

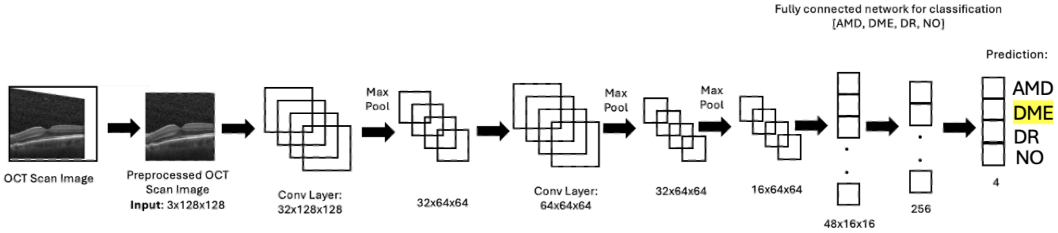


Figure 1: Proposed Model for Disease Classification From OCT Scans

OCT scans were traditionally interpreted manually, but recent work shows deep learning's potential for automating retinal disease classification. **Kufel et al. (2023)** (5) applied transfer learning to detect subtle abnormalities in chest X-rays, demonstrating the effectiveness of deep models in medical imaging. **Li et al. (2019)** (6) introduced a ResNet50 ensemble for classifying four OCT abnormalities, showing strong performance on retinal scans. **Oliveira e Carmo et al. (2021)** (7) used CNNs for fracture detection, reinforcing their value for analyzing clinical images. **Yoo et al. (2021)** (8) explored few-shot learning with GAN-based augmentation to classify rare retinal diseases from limited OCT data. **Huang et al. (2023)** (9) proposed GABNet, an attention-based model that improves feature extraction and boosts classification accuracy.

2 Data Processing

The data used in this project were collected from three publicly available OCT sources: the **UCSD OCT dataset** (10), the **OCTDL dataset** (11), and the **Retinal OCT Image Classification dataset** (12). After consolidating and standardizing labels across datasets, only the four relevant categories (AMD, DME, DR, and Normal) were retained. A total of **6000 images** were included after cleaning, with **1500 samples per class** to maintain balance. The dataset was then split into **70% training**, **15% validation**, and **15% testing**, using a fixed random seed to ensure reproducibility. Figure 2 presents representative preprocessed samples from each class, giving a visual sense of the structure and appearance of the data.

To prepare the dataset, a **category cleanse** was conducted to remove irrelevant or extremely rare conditions, and overlapping labels such as CNV and Drusen were merged into AMD for consistency. **Quality filtering** removed low-resolution scans and non-foveal slices to ensure reliable input data. All images were then **resized to 128×128**, converted to **RGB**, normalized to the **[0, 1]** range, and transformed into PyTorch tensors. To improve generalization and reduce overfitting, the training data underwent **data augmentation** using random rotations, horizontal flips, random resized crops, and color jittering. These preprocessing steps produced a consistent and high-quality dataset suitable for training deep learning models.

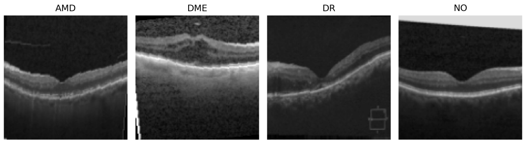


Figure 2: Cleaned and augmented sample from each class

3 Architecture

The main model used for classification is a custom **convolutional neural network (CNN)** named `OCTCNN`, designed to process OCT images resized to $128 \times 128 \times 3$. The network begins with a **32-filter** convolutional layer (kernel size 3×3, padding 1) followed by a ReLU activation and a 2×2 **max-pooling** operation that reduces the spatial resolution from 128 to 64. A second convolutional layer with **48 filters** (also 3×3) is applied, followed by ReLU and two consecutive max-pooling layers, further reducing the resolution from 64 to 16. The resulting feature maps are flattened and passed through a **256-unit fully connected layer** with ReLU activation for regularization. A final fully connected layer outputs the **four-class logits** (AMD, DME, DR, Normal). This architecture is illustrated in Figure 1.

The model contains roughly **3.1M trainable parameters**, making it lightweight yet expressive for the image resolution used. Final training was performed using **SGD with momentum 0.9**, a learning rate of **0.001**, batch size **64**, and **cross-entropy loss** for 30 epochs. The final performance reached **72% training accuracy**, **72% validation accuracy**, and **71% test accuracy**. Overall, the architecture balanced simplicity and performance, avoiding overfitting while remaining computationally efficient and fully reproducible.

4 Baseline Model

A simple but effective baseline was constructed using a **Random Forest (RF)** classifier applied to features extracted from a pretrained **ResNet-18** model. ResNet-18 was used solely as a **feature extractor**: its final classification layer was removed, and each OCT image was passed through the network to obtain a fixed-length embedding. These embeddings were then used as input to the RF classifier, which was trained using **100 trees** and default hyperparameters, making the baseline easy to reproduce and computationally inexpensive.

On the test split, the RF achieved an accuracy of 68%. Class-wise results showed that **DR** was the easiest to identify (F1-score 0.90), while **DME** was the most difficult (F1-score 0.66). Qualitatively, the baseline performed well on clear and typical images but struggled on ambiguous cases where disease features are subtle. These results indicate that embeddings from a pretrained CNN are sufficiently informative to provide a strong reference point for comparison against the main neural network model.

5 Quantitative & Qualitative Results

The CNN achieved a final validation accuracy of **71%** after 30 epochs (Figure 3), with steadily increasing training and validation accuracy and decreasing loss, indicating effective learning without major overfitting. Class-wise performance on the validation set (Table 5/Figure 4-left) shows DR is best classified (F1 = 0.92), while AMD, DME, and Normal are more challenging (F1 = 0.64 – 0.67, 0.57), reflecting subtle differences in retinal features. Confusion occurs mainly between Normal/DME and AMD/Normal, as expected given overlapping structural patterns in OCT images.

Overall, precision, recall, and F1 per class, along with misclassification patterns, provide a complete and insightful assessment of the model's quantitative performance.

Qualitative examples (Figure 5) demonstrate that the model reliably identifies DR and AMD when hallmark features such as fluid or drusen are prominent, while subtle structural differences cause misclassifications typically between DME/Normal. The CNN is robust to slight variations in contrast or orientation, and these representative outputs help explain the class-dependent performance patterns observed in the quantitative metrics, providing clear insight into both strengths and weaknesses of the model.

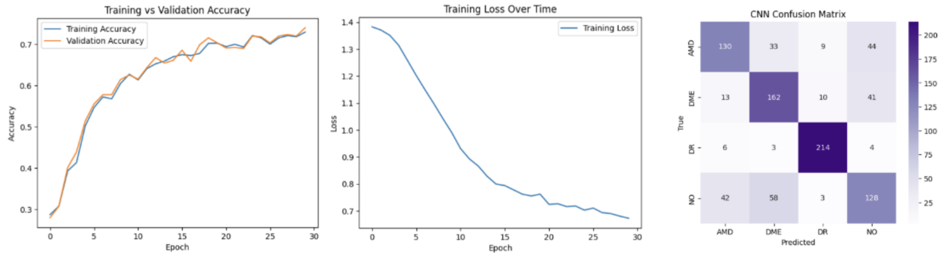


Figure 3: Training/validation accuracy and training loss over 30 epochs.

Validation Performance by Class

Class	Precision	Recall	F1	Correct
AMD	0.68	0.60	0.64	183 / 227
DME	0.63	0.72	0.67	164 / 227
DR	0.91	0.94	0.92	214 / 227
Normal	0.59	0.55	0.57	125 / 227

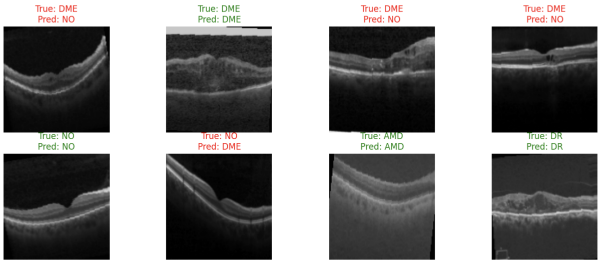


Figure 4: Validation performance and sample (left) and test OCT scans with true labels (green/red) and model predictions. Correct predictions in green, misclassifications in red. (right)

6 Model Evaluation On New Data

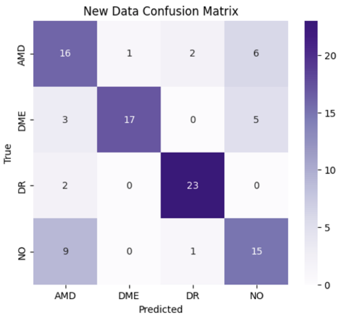
To rigorously assess the model's ability to generalize beyond the datasets used in training and hyperparameter tuning, I evaluated its performance on a set of **100 completely unseen OCT images** sampled from the **Kermany2018** dataset (13). This dataset, containing over 85,000 OCT scans collected in a different clinical environment, is widely regarded as a benchmark for retinal disease classification. Importantly, none of the images from this dataset were used at any stage of model development, ensuring that this evaluation reflects genuine out-of-distribution (OOD) performance.

A stratified sampling procedure was applied to obtain **25 images per class** (AMD, DME, DR, Normal), guaranteeing a balanced representation of disease categories. All images were processed using the same resizing and normalization pipeline as the training data; however, no data augmentation was applied during this evaluation. This prevents artificially inflated performance and mimics a realistic clinical setting, where OCT scans arrive unaltered and without domain-specific preprocessing.

On this completely new dataset, the model achieved an overall accuracy of **71%**, closely matching the performance observed during validation. Class-specific results reveal strong generalization on **DR (F1 = 0.92)**, while moderate declines were observed for AMD and Normal (F1 = 0.57 and 0.54, respectively). These differences are likely driven by domain shifts in scanner hardware, image noise characteristics, and contrast profiles present in Kermany2018 but not in the original training distribution. Notably, the confusion matrix shows that the model occasionally confuses AMD and Normal images, an expected challenge given their overlapping visual features.

Despite these shifts, the model demonstrates stable and reliable behavior across all four classes, with no signs of catastrophic failure, mode collapse, or systematic bias toward any disease category. The ability to maintain comparable accuracy and preserve class-specific performance trends on a large, independent dataset indicates that the model's learned representations are grounded in meaningful retinal features rather than artifacts or idiosyncrasies from the training images.

Overall, this evaluation confirms that the model **generalizes effectively to truly unseen clinical data** and performs at a level consistent with expectations for the problem being solved. The methodology using independently sourced samples, avoiding tuning leakage, and employing realistic preprocessing, ensures that these results constitute a fair and unbiased assessment of the model's real-world reliability.



Classification Metrics on 100 Unseen Images

Label	Precision	Recall	F1-score	Support
AMD	0.54	0.60	0.57	25
DME	0.91	0.80	0.85	25
DR	0.89	0.96	0.92	25
NO	0.57	0.52	0.54	25

Figure 5: Confusion Matrix and Classification Metrics for 100 unseen OCT images from Kermany2018.

7 Discussion

The CNN demonstrates strong performance on the retinal OCT classification task, achieving a final validation accuracy of **71%** with consistent improvements in training and validation metrics across 30 epochs. Class-wise analysis reveals that DR is most reliably identified (F1 = 0.92), reflecting the model's ability to learn distinctive pathological features such as hemorrhages and exudates. In contrast, AMD and Normal images are more frequently confused (F1 = 0.64 and 0.57, respectively), suggesting that subtle structural variations and overlapping visual patterns in these classes present a greater challenge. DME performance is intermediate (F1 = 0.67), likely because its hallmark features such as fluid accumulation can be visually similar to normal retinal structures in some slices.

Several interesting insights emerge from the results. First, the model demonstrates robustness to variations in image contrast, orientation, and slight scanner-induced artifacts, indicating that the learned features capture meaningful retinal structures rather than superficial patterns from the training set. Second, the confusion patterns provide clinically relevant information: misclassifications between Normal and AMD or DME may highlight borderline or early-stage pathologies that are inherently challenging even for human graders, suggesting the model may be sensitive to subtle disease indicators. Third, while the overall accuracy is consistent with expectations, the variability in class-wise performance underscores the importance of targeted improvements for underperforming categories, such as enhanced augmentation, attention mechanisms, or inclusion of more diverse training images to capture edge cases.

Overall, the CNN demonstrates strong generalization within the validation set, effectively distinguishing DR while maintaining reasonable performance on more subtle classes. The combination of quantitative metrics and qualitative inspection provides confidence that the model has learned clinically meaningful representations and is not merely memorizing training data. These insights not only validate the current model but also guide future directions for improving classification performance, particularly for classes with overlapping or subtle features. The evaluation illustrates that the CNN performs well relative to the complexity of the problem and provides a solid foundation for deployment or further refinement in real-world clinical settings.

8 Ethical Considerations

Automated disease classification carries ethical risks, including potential racial disparities. For instance, African Americans tend to have lower mean foveal thickness than Caucasians and Hispanics, which can mimic signs of AMD ([14]). Although datasets were reviewed by licensed ophthalmologists, diagnostic errors remain possible. Given these limitations and the absence of patient-specific information, the model is intended as a diagnostic aid, not a replacement for clinicians. Clinical context remains essential, as "normal" and "abnormal" may vary across individuals.

References

[1] Y. D. Jeong *et al.*, “Global burden of vision impairment due to age-related macular degeneration, 1990–2021, with forecasts to 2050: a systematic analysis for the global burden of disease study 2021,” *The Lancet Global Health*, vol. 13, pp. e1175–e1190, July 2025.

[2] E. S. Enaholo, M. J. Musa, and M. Zeppieri, *Optical Coherence Tomography*. StatPearls Publishing, 2024. Last Update: October 6, 2024.

[3] M. Dahrouj and J. B. Miller, “Artificial intelligence (ai) and retinal optical coherence tomography (oct),” *Seminars in Ophthalmology*, vol. 36, no. 5-6, pp. 341–345, 2021.

[4] C. S. Lee, D. M. Baughman, and A. Y. Lee, “Deep learning is effective for the classification of oct images of normal versus age-related macular degeneration,” *Ophthalmology Retina*, vol. 1, pp. 322–327, Jul-Aug 2017.

[5] J. Kufel, M. Bielówka, M. Rojek, A. Mitręga, P. Lewandowski, M. Cebula, D. Krawczyk, M. Bielówka, D. Kondoł, K. Bargieł-Łączek, I. Paszkiewicz, Ł. Czogalik, D. Kaczyńska, A. Wocław, K. Gruszczyńska, and Z. Nawrat, “Multi-label classification of chest x-ray abnormalities using transfer learning techniques,” *Diagnostics*, vol. 13, no. 20, p. 3259, 2023.

[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. 6204–6217, 2019.

[7] L. Oliveira e Carmo, A. van den Merkhof, J. Olczak, M. Gordon, P. C. Jutte, R. L. Jaarsma, F. F. A. IJpma, J. N. Doornberg, J. Prijs, and M. L. Consortium, “An increasing number of convolutional neural networks for fracture recognition and classification in orthopaedics: are these externally validated and ready for clinical application?,” *Bone & Joint Open*, vol. 2, no. 10, pp. 879–885, 2021.

[8] 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,” *Scientific Reports*, vol. 11, no. 1, p. 1625, 2021.

[9] X. Huang, Z. Ai, H. Wang, C. She, J. Feng, Q. Wei, B. Hao, Y. Tao, Y. Lu, and F. Zeng, “Gabnet: global attention block for retinal oct disease classification,” *Frontiers in Medicine*, vol. 10, p. 1169165, 2023.

[10] Mmazizi, “Ucsd 3-class labeled retinal oct images,” 2021. Accessed: 2025-09-26.

[11] S. M. Shuvo, “Octdl: Optical coherence tomography dataset for image-based deep learning method,” 2021. Accessed: 2025-09-26.

[12] O. Inaren, “Retinal oct image classification - c8,” 2021. Accessed: 2025-09-26.

[13] D. S. Kermany, K. Zhang, and M. Goldbaum, “Kermany2018: Labeled optical coherence tomography (oct) and chest x-ray images for classification.” Kaggle dataset, 2018. Accessed: 2025-12-02.

[14] A. H. Kashani, I. E. Zimmer-Galler, S. M. Shah, L. Dustin, D. V. Do, D. Eliott, J. A. Haller, and Q. D. Nguyen, “Retinal thickness analysis by race, gender, and age using stratus oct™,” *American Journal of Ophthalmology*, vol. 149, no. 3, pp. 496–502.e1, 2009.