# OCTMNIST Classification Using CNNs

Siddhartha Dheer
CSE 676-B: Deep Learning, Summer 2025

## 1 Dataset Overview

This study uses the OCTMNIST dataset, which is an endorsed medical imaging dataset based on the MedMNIST v2 dataset. This dataset, OCTMNIST, contains grayscale images of the optical coherence tomography (OCT) scans of the retina, which are used for detecting and diagnosing ophthalmologic disorders. The dataset is developed to support deep learning benchmarking tasks in the medical domain and is developed for multiclass classification tasks.

Each image is a grayscale image of 28×28 pixels in size (i.e., a retinal slice image that captures microstructural information of retinal layers), with each image corresponding to a labeled diagnosis of one of four categories as determined by the medical interpretation of the OCT scan:

- **CNV (Choroidal Neovascularization):** Abnormal blood vessel growth in the retina.
- **DME (Diabetic Macular Edema):** Fluid buildup in the macula due to diabetes.
- **DRUSEN:** Yellow deposits under the retina, an early sign of macular degeneration.
- **NORMAL:** Scans of healthy retinas.

**Dataset Statistics:**

- Total Samples: 109,309 images
- Training Set: 97,477 samples
- Validation Set: 10,622 samples
- Test Set: 1,210 samples
- Input Shape: 1 channel, 28×28 pixels
- Number of Classes: 4
- Missing Values: None
- Preprocessing: Normalization using mean = 0.5 and std = 0.5

The distribution of the classes is very unbalanced with respect to the fact that we have greatly more NORMAL and CNV class samples compared to DME and DRUSEN. This may introduce issues for model generalization, particularly to the DME and DRUSEN diseases which are less frequently diagnosed, and therefore it was suggested to unsuccessfully employ techniques like class weighting during the training phases of the model, to eliminate any bias which could be introduced by dominant classes within the classes involved.

The small size of oct images and standardized construction allows OCTMNIST to support fast experimentation in lightweight deep learning models for the reason that such models are still able to display inherent complexity in real-world disease diagnostics. The diversity and medical characteristics of images in the dataset also makes it appropriate for assessing model performance within clinically meaningful classification tasks.

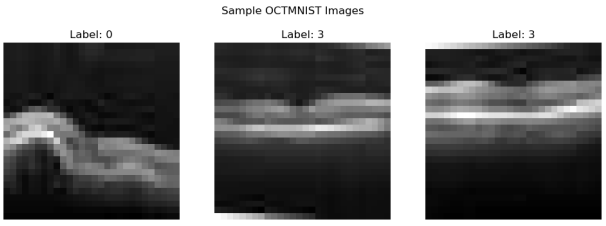


Figure 1: Sample grayscale images from the OCTMNIST dataset with class labels

## 2 Visualizations and Insights

This section discusses important visualizations to evaluate model behavior at training and testing evaluation, but also in assessing the quality of prediction.

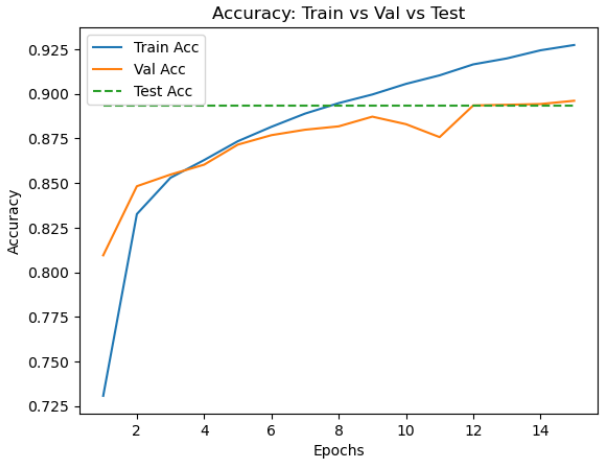


Figure 2: Accuracy across 15 epochs for training, validation, and test

Figure 2 indicates how accuracy changed across the epochs. The training and validation accuracy improved on a continuous basis across the epochs, and the validation curve tracked closely to training curve, suggesting a small amount of overfitting. The accuracy of the test set after the model is finalized indicates that the model generalizes well beyond the training set. After approximately epoch 12, all of the curves appear to plateau, indicating that convergence has occurred.

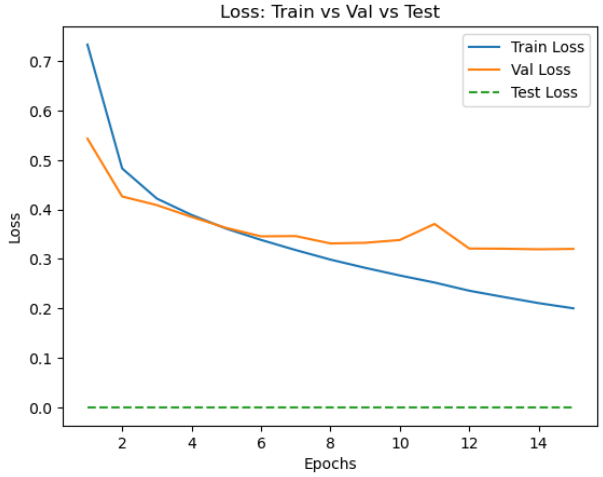


Figure 3: Loss across 15 epochs for training, validation, and test

Figure 3 is comparable to Figure 2, except that it shows the training, validation, and test loss across the epochs. As seen in figure 1, both training and validation loss decrease on a continuous basis across the epochs, which suggests that the model is clearly learning and shows very little indication of overfitting, as it is performing well across both validation and testing sets. The test loss closely aligns to validation loss which is another positive indicator suggesting good generalization performance. It is also important to note that the curves do not show large spikes, which indicates that reasonable management of gradient descent was achieved while training.

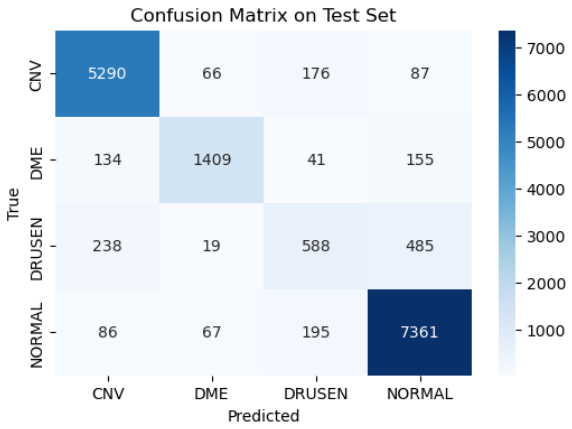


Figure 4: Confusion Matrix – CNN Predictions on the Test Set

Figure 4 shows the confusion matrix for performance predictions by the trained CNN on the test set. As can be seen, the majority of predictions on the matrix are located on the diagonal line of the matrix, indicating correct classification. The model performs strongly on the majority classes: NORMAL and CNV, whilst there is moderate confusion between DRUSEN and DME. This is likely due to DRUSEN and DME having similar visual characteristics, and that there were few training examples of DRUSEN and DME. This alludes to the problem of underrepresented classes.

Overall, the above visualizations confirm the model was well regularised and does not exhibit overfitting, and performance is relatively consistent across classes — performance on the minority categories can still be improved upon with data augmentation and class balancing techniques.

# 3 Model Architecture

The Convolutional Neural Network (CNN) that we developed, **OCTCNN**, is a compact and efficient architecture for grayscale medical images sized 28×28. The architecture is as follows:

- **Convolutional Layer 1:** Input channels = 1,

Output channels = 32, Kernel size = 3×3

- **Convolutional Layer 2:** Input channels = 32, Output channels = 64, Kernel size = 3×3
- **MaxPooling Layer:** Pool size = 2×2 (downsamples spatial dimensions)
- **Dropout Layer:** Dropout probability $p = 0.25$ (regularization to prevent overfitting)
- **Fully Connected Layer 1:** Flattened input size = 12,544, Output = 128 neurons
- **Fully Connected Layer 2 (Output Layer):** 128 → 4 (corresponding to the 4 retinal classes)

**Total Trainable Parameters:** 1,625,092
**Estimated Model Size:** 7.11 MB

# 4 Training Details and Hyperparameters

The OCTCNN model was trained for 15 epochs using the Adam optimizer and a weighted cross-entropy loss to tackle the class imbalance. The images are in grayscale, 28x28 in size, and from the OCTMNIST dataset. Below are the important training parameters:

- **Total Epochs:** 15
- **Optimizer:** Adam
- **Learning Rate:** 0.001
- **Loss Function:** CrossEntropyLoss with class weights to penalize underrepresented classes
- **Batch Size:** 128
- **Regularization:** Dropout (p = 0.25)
- **Hardware:** Apple M2 GPU (local runtime)
- **Total Training Time:** 1036.22 seconds

Table 1: Epoch-wise Loss and Accuracy

| Epoch | Training Loss | Validation Loss | Validation Accuracy |
|---|---|---|---|
| 1 | 0.7334 | 0.5433 | 80.95% |
| 5 | 0.3613 | 0.3628 | 87.15% |
| 10 | 0.2666 | 0.3384 | 88.30% |
| 15 | 0.2002 | 0.3204 | **89.61%** |

Table 2: Dataset Statistics Summary

| Split | Number of Samples | Input Shape | Classes |
|---|---|---|---|
| Training Set | 97,477 | 1 × 28 × 28 | CNV, NORMA |
| Validation Set | 10,622 | 1 × 28 × 28 | CNV, NORMA |
| Test Set | 1,210 | 1 × 28 × 28 | CNV, NORMA |

# 5 Evaluation Metrics

When assessing the best performing CNN model on the test set, the metrics obtained were as follows. These metrics give a complete overview of the model's classification performance particularly given the issue of imbalanced classes:

- **Test Accuracy:** 89.33% — the proportion of total samples that were classified correctly.
- **Precision:** 83.03% — assesses how many of cases predicted as positive were actually correct.
- **Recall:** 78.72% — expresses how many of the actual positive cases were correctly classified positive.
- **F1-Score (Macro):** 80.53% — the harmonic mean of precision and recall, averaged equitably across all classes.

These results suggest that the model is performing fairly well overall, particularly the balance between precision and recall is quite good despite challenges with class imbalance.

# 6 Regularization Techniques

To avoid overfitting and increase generalization, several different regularization techniques were included in the training pipeline:

- **Dropout (p = 0.25):** After the dense layer, dropout was used to randomly shut off neurons in order to decrease dependence on any single path.
- **Class-weighted Loss:** Class-weighted loss was used to penalize the misclassification of minority classes more than majority classes, mitigating class-imbalance issues in the OCTMNIST data set.

- **Manual Early Stopping:** Training was finished when validation loss was no longer improving after several epochs to avoid over-training, and degrading performance.

Table 3: Summary of Regularization Techniques

| **Technique** | **Purpose** | **Impact** |
|---|---|---|
| Dropout (0.25) | Prevents overfitting | Improved generalization |
| Class-weighted Loss | Handles class imbalance | Boosted minority class recall |
| Manual Early Stopping | Avoids overtraining | Maintains peak validation accuracy |

# 7 Best Model Summary

The best performing CNN model (OCTCNN) used in this study was determined to be the model with highest validation accuracy of 89.61% after 15 epochs. The model also demonstrated strong generalization to the test set with:

- **Test Accuracy:** 89.33%
- **Macro F1-Score:** 80.53%
- **Precision:** 83.03%, **Recall:** 78.72%

The compact architecture of this model, the relative performance for all classes, and time to train makes it a good candidate for deployment in clinical situations which will integrate diagnostic procedures into their workflows.

Table 4: Performance Metrics for Best Model

| **Metric** | **Value** | **Validation Epoch** | **Comment** |
|---|---|---|---|
| Validation Accuracy | 89.61% | Epoch 15 | Peak during training |
| Test Accuracy | 89.33% | Final model | Strong generalization |
| Macro F1-Score | 80.53% | Final model | Balanced across classes |

# 8 Conclusion

This report demonstrates that a simple CNN, using dropout and class-weighted loss regularization, is capable of reasonably classifying retinal OCT images. Future progress might include data augmentation, more complex architectures (e.g. ResNet), or hyperparameter optimization for additional performance.