Project Summary

1. Our approach is primarily based on adding one neuron to the CNN to represent the "unknown" class, along with using data augmentation to generate additional, more challenging examples.

   We also set a confidence threshold for each image's predicted label. Specifically, if the model's confidence for a label is below 97.5% (after training for 20 epochs), the image is considered most likely "unknown."

   The rationale behind this is that by training the model for 20 epochs—with each epoch processing 50,000 distinct images—the model becomes highly confident in its predictions on the test set. Therefore, if the model is not sufficiently confident about an image, it is likely that the image comes from CIFAR10 or FashionMNIST, leading the model to label it as "unknown."

   In our first attempt, we built the model without any dropout layers, resulting in an OOD accuracy of 87.10%. We then added two dropout layers to reduce overfitting, improve feature space separation, and enhance the model's robustness to biases in the training data. As a result, we achieved a significantly improved OOD accuracy of 97.40%.

2. The primary hyperparameters of the model were the number of epochs, training batch size, the number of layers in the CNN, the dropout probability, and the threshold value for the model's confidence regarding an image's label.

   I selected these parameters based on validation set experiments, with the final values chosen after various validation sets consistently showed high performance for these hyperparameter configurations.

3. We aim to support the claim that our approach achieves robust performance on the OSR problem for datasets with images that closely resemble those in MNIST. Specifically, as the number of training epochs increases and the model's confidence threshold for assigning a label is raised accordingly, the performance improves significantly (We see that the claim holds true based on an analysis of the EMNIST dataset).

   We used a training set of 80,000 images (60,000 from MNIST and 20,000 from data augmentation) and ran the training algorithm for 50 epochs. Afterwards, we set the model's confidence threshold to 0.99995.

   According to image 1, we can see that the model's error decreases as the number of epochs increases, which bodes well for its performance on the EMNIST dataset. This trend is confirmed in image 2, where the model achieves an accuracy of 87.20% (OOD Accuracy) on a dataset composed of EMNIST and the MNIST test set. Image 3 illustrates the model's

accuracy by showing a relatively clear separation between the different data groups. Finally, image 4 demonstrates the model's high separation capability, with an AUC of 0.9 between the EMNIST and MNIST datasets.

We note that the performance of this model on the dataset composed of CIFAR10 and FashionMNIST achieves an accuracy of 99.05% (OOD Accuracy). In other words, the model's confidence in assigning a label to a given image has increased significantly.

Figure 1

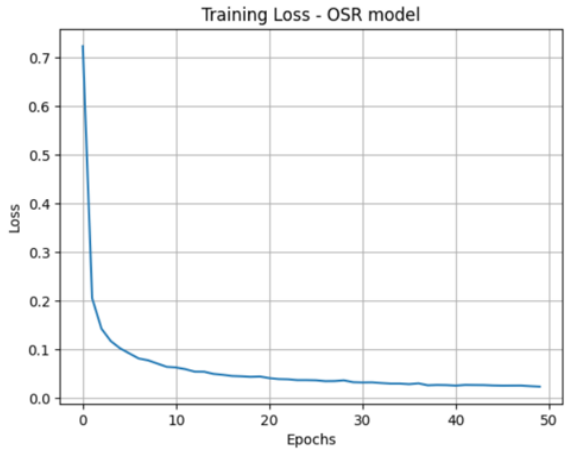


Figure 2

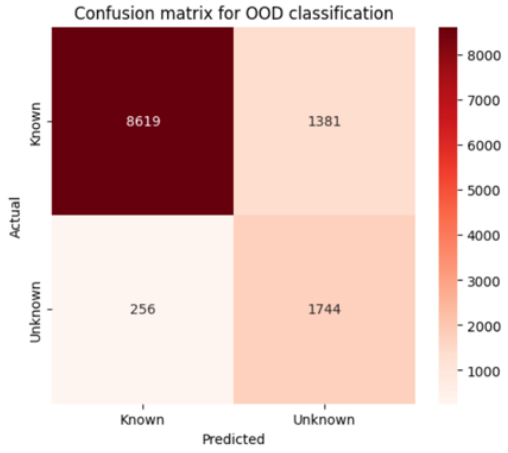


Figure 3

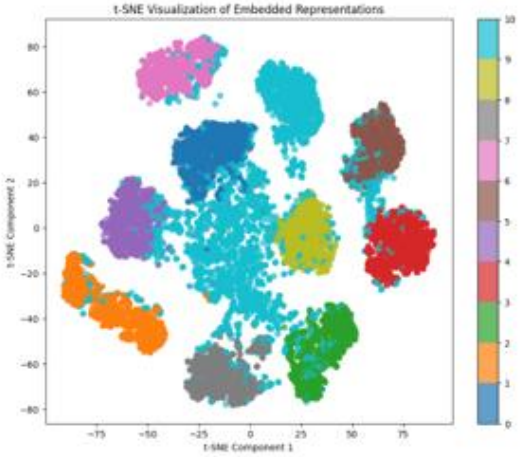
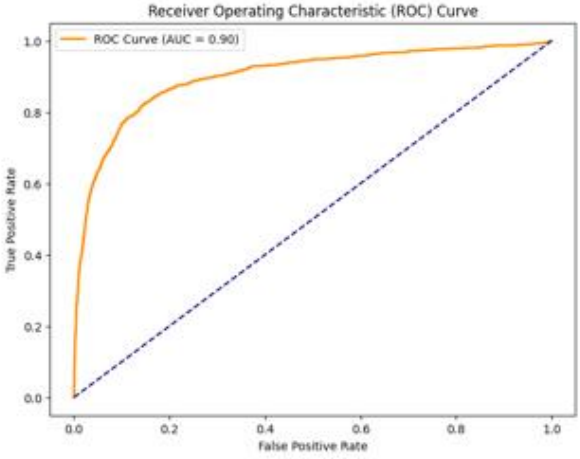


Figure 4



4. One of the main limitations of our approach is the training time, which is capped at 10 minutes. As a result, we selected the appropriate number of epochs and training set size to fit within this timeframe. We observed that as we increase the number of epochs and expand the training set, the model's confidence in its predictions for a given image improves, leading to fewer errors.

For the EMNIST dataset (handwritten characters), we obtained an accuracy of 66.05%.

For the SVHN dataset (Street view numbers), the accuracy reached 99.80%.

The Omniglot dataset (Diverse handwritten alphabets) yielded 99.55% accuracy,

while the USPS dataset (Postal handwritten digits) achieved 57.60%.

Lastly, the STL10 dataset (Larger natural images) produced an accuracy of 98.20%.

(All the percentages shown are based on OOD accuracy)

We observe that our approach performs well on datasets where there is a significant difference between the images of the selected dataset and those of MNIST. However, it struggles with datasets whose images are very similar to MNIST, which necessitates increasing the number of epochs and enlarging the training set in such cases.