Applied Machine Learning Final Project

By Dean Zruya (207153628) and Emanuel Goldman (315690750)

Project Overview

Title: Out-of-Distribution (OOD) Detection Using Extreme Value Theory (EVT) and Sparse Reconstruction Errors

Objective:

The primary goal of this project is to detect out-of-distribution (OOD) samples in an image classification task by leveraging sparse reconstruction errors and Extreme Value Theory (EVT). The approach involves training an autoencoder, analyzing reconstruction errors, and using EVT to set an adaptive threshold for OOD detection.

Methodology:

1. **The Basic Autoencoder-Decoder:**

   The autoencoder consists of an encoder that compresses input images into a low-dimensional latent space and a decoder that reconstructs the images from this representation. It is trained using Mean Squared Error (MSE) loss to minimize reconstruction error.

2. **Baseline Classification Model:**

   The autoencoder is extended with a classification head to predict digit classes (0-9) using encoded representations. The classifier is trained with Negative Log Likelihood (NLL) loss to optimize prediction accuracy.

3. **OOD Detection with EVT:**

   To detect out-of-distribution (OOD) samples, reconstruction errors are analyzed. Instead of using a fixed threshold, Extreme Value Theory (EVT) is applied to model the tail distribution of reconstruction errors using the Generalized Pareto Distribution (GPD). The EVT-derived threshold provides a more adaptive and statistically sound approach for identifying OOD samples.

   This methodology ensures robust feature extraction, accurate classification, and effective OOD detection.

   Important note:

   Since the baseline and final models share the same parameters, we trained a single model for both. In the baseline model, OOD detection was effectively disabled by setting the threshold to a very high value (10^8), ensuring that all inputs were classified as in-distribution.

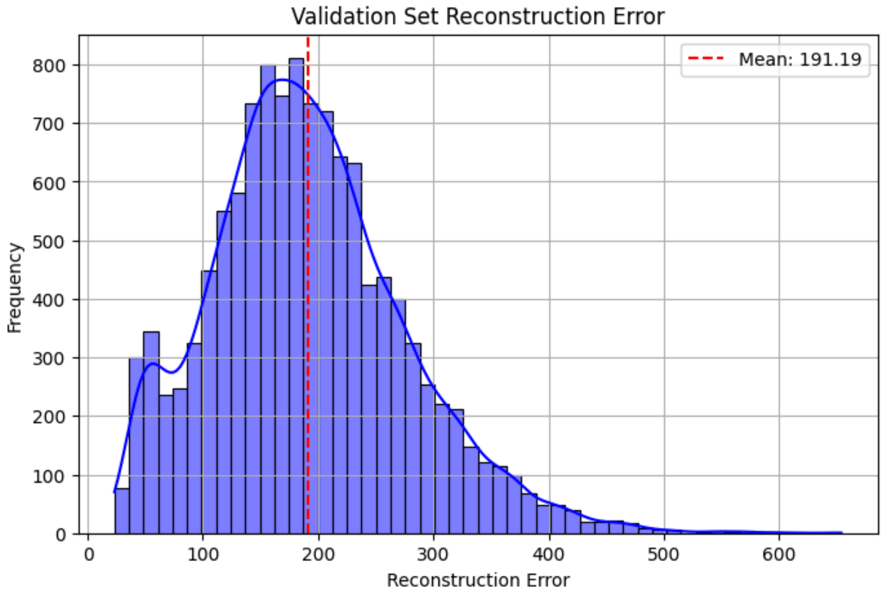
Model Training:

Our initial model exhibited overfitting, as indicated by a decreasing training loss while the validation loss plateaued. To improve generalization, we introduced dropout layers and weight decay, which mitigated overfitting and enhanced performance on unseen data. After these modifications, the loss curves demonstrated that classification stabilized around epoch 20, while reconstruction continued to improve. To optimize training, we limited classification training to the final 20 epochs while maintaining reconstruction training for the full 60 epochs.

Hyper-Parameters:

1. **Latent Dimension** *`(code_size = 32)`*
   - Controls the size of the compressed representation.
   - Chosen as a balance between compression and preserving useful features for classification.
2. **Learning Rate** *`(lr = 1e-3)`*
   - Determines step size for weight updates.
   - Selected through experimentation, ensuring convergence without overshooting optima.
3. **Weight Decay** *`(1e-5)`*
   - Helps prevent overfitting by penalizing large weights.
   - A small value is chosen to maintain generalization.
4. **Batch Size** *`(batch_size = 1024)`*
   - Impacts training stability and memory usage.
   - A batch size of 1024 provides a good trade-off between efficiency and gradient estimation.
5. **Loss Functions:**
   - **Reconstruction Loss** *`(MSELoss)`* for autoencoder training.
   - **Classification Loss** *`(NLLLoss)`* for training the classifier.
   - These choices align with standard practices for autoencoder and classification training.
6. **Optimizer** *`(Adam)`*
   - Adaptive learning rate optimization for faster convergence.
   - Used due to its effectiveness in deep learning tasks.

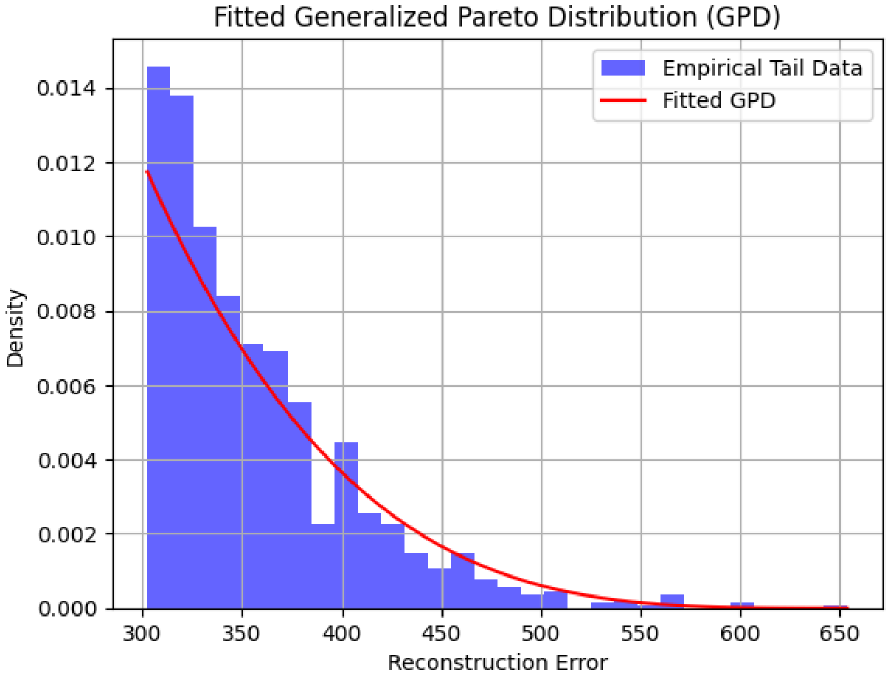
Choosing the Right Threshold for OOD Detection



Initial Thoughts: At first, we considered using a **fixed threshold** based on the mean reconstruction error, assuming that errors significantly higher than the mean could indicate out-of-distribution (OOD) samples.

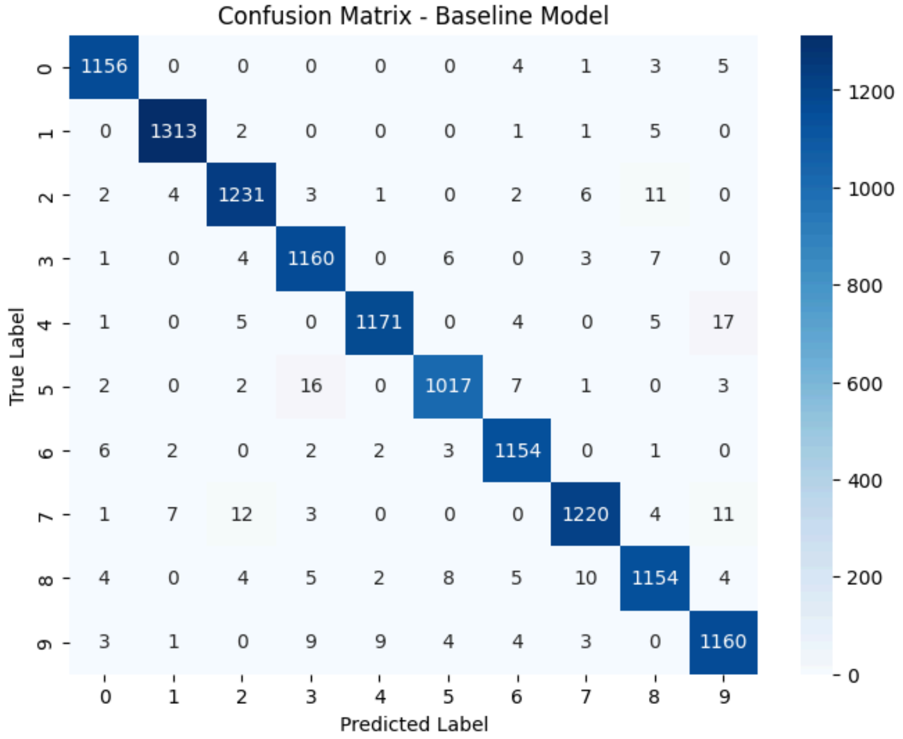
However, we noticed that the distribution has a long tail, meaning that a simple threshold based on the mean might not effectively capture extreme reconstruction errors.

This led us to explore **Extreme Value Theory (EVT)** to model the tail of the distribution. EVT provides a more **adaptive** and **statistically robust** way to set the threshold, ensuring better generalization across datasets rather than relying on an arbitrary cutoff.



We defined the tail of the reconstruction error using the **90th percentile** threshold. To classify a sample as OOD, we set a confidence threshold of **50%**, meaning that if the model is at least **70% certain** that a sample is out-of-distribution, it will be classified as such.

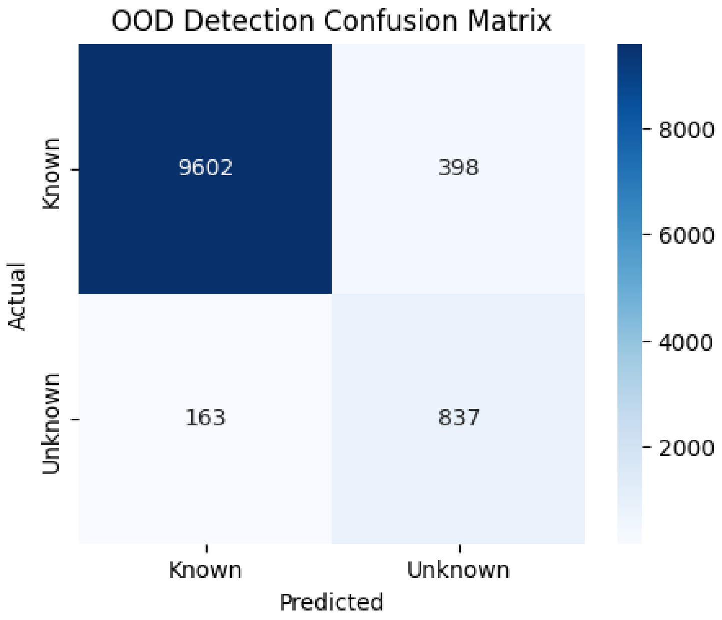
Evaluation Metrics:



Confusion Matrix - Baseline Model

True Label \ Predicted Label	0	1	2	3	4	5	6	7	8	9
0	1156	0	0	0	0	0	4	1	3	5
1	0	1313	2	0	0	0	1	1	5	0
2	2	4	1231	3	1	0	2	6	11	0
3	1	0	4	1160	0	6	0	3	7	0
4	1	0	5	0	1171	0	4	0	5	17
5	2	0	2	16	0	1017	7	1	0	3
6	6	2	0	2	2	3	1154	0	1	0
7	1	7	12	3	0	0	0	1220	4	11
8	4	0	4	5	2	8	5	10	1154	4
9	3	1	0	9	9	4	4	3	0	1160

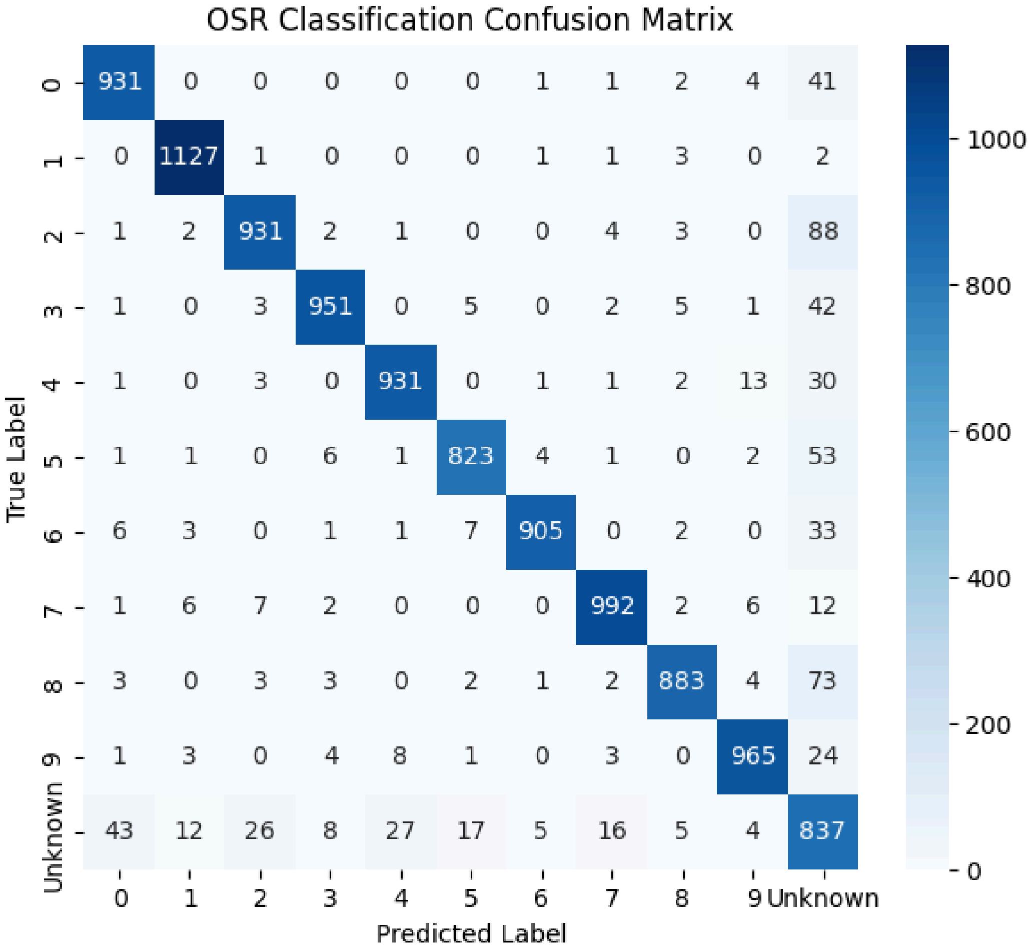
The confusion matrix shows the baseline model's strong classification performance on MNIST, with minimal misclassifications. The **baseline model accuracy** reached **97.80%**.



OOD Detection Confusion Matrix

Actual \ Predicted	Known	Unknown
Known	9602	398
Unknown	163	837

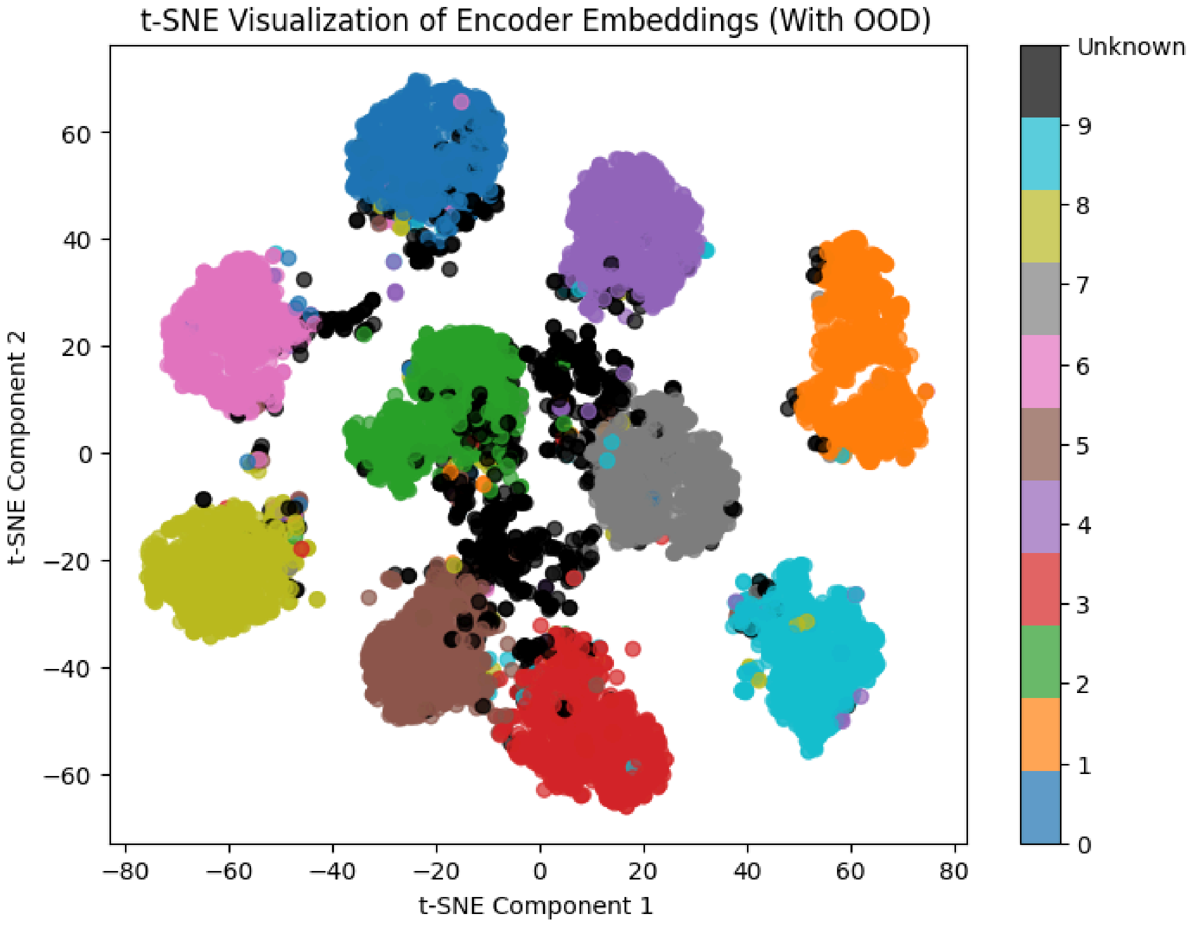
The model achieved an **OOD binary classification accuracy** of **94.90%**, an **MNIST classification accuracy** of **94.39%**, and an **OOD classification accuracy** of **83.70%**, resulting in an overall **total accuracy** of **93.42%**.



True Label \ Predicted Label	0	1	2	3	4	5	6	7	8	9	Unknown
0	931	0	0	0	0	0	1	1	2	4	41
1	0	1127	1	0	0	0	1	1	3	0	2
2	1	2	931	2	1	0	0	4	3	0	88
3	1	0	3	951	0	5	0	2	5	1	42
4	1	0	3	0	931	0	1	1	2	13	30
5	1	1	0	6	1	823	4	1	0	2	53
6	6	3	0	1	1	7	905	0	2	0	33
7	1	6	7	2	0	0	0	992	2	6	12
8	3	0	3	3	0	2	1	2	883	4	73
9	1	3	0	4	8	1	0	3	0	965	24
Unknown	43	12	26	8	27	17	5	16	5	4	837

The OSR confusion matrix shows strong classification for MNIST digits, with most "Unknown" samples correctly classified. However, class 2, 8 and 5 are frequently misclassified as "Unknown," indicating a challenge in distinguishing them from OOD samples. **Total OSR Classification Accuracy is: 93.42%.**

t-SNE Visualization of Encoder Embeddings: Separating MNIST Digits and OOD Data



In this visualization, we applied t-SNE to reduce the high-dimensional encoder embeddings to 2D, allowing us to observe their structure. Each color represents a digit class (0-9) from MNIST, while the black points correspond to OOD (Out-of-Distribution) samples. We ensured proper label mapping so that the OOD class (10) is distinctly represented in black. The clusters indicate that the encoder effectively separates digit classes, while OOD samples are spread across different regions, showing where the model struggles to distinguish them.

Conclusion: The encoder successfully learns compact and well-separated embeddings for MNIST digits, demonstrating strong intra-class consistency. However, the OOD samples do not form a distinct cluster, indicating that they are not easily separable from the known classes. This suggests that while the model can recognize in-distribution digits with high confidence, its ability to detect and isolate OOD samples could be improved, possibly through enhanced OOD detection mechanisms such as better thresholding, uncertainty estimation.

Limitations & Future Work

The method's performance is dependent on the quality of the fitted GPD.
Additional validation is needed across multiple datasets to ensure generalization.
Future work can explore improvements using more sophisticated deep generative models (e.g., normalizing flows, diffusion models).
This approach ensures a reliable OOD detection system by combining deep learning with statistical extreme value modeling, enabling better robustness in real-world classification tasks.

Limitations:

- **Dependence on Training Data:** The model performs well when trained on diverse and representative data but may struggle with unseen distributions significantly different from the training set.
- **Separation of OOD Samples:** While the model detects OOD samples, they do not always form a distinct cluster, making clear separation difficult.
- **High-Dimensional Data Complexity:** The t-SNE projection to 2D may not fully capture high-dimensional relationships, limiting interpretability in some cases.
- **Threshold Sensitivity:** The OOD detection threshold may need fine-tuning for different datasets, impacting performance across domains.
- **Limited Performance on Subtly Shifted Data:** The method is effective for detecting drastically different OOD samples but may struggle with **near-OOD** cases (e.g., blurry or perturbed versions of in-distribution data).

Expected Performance:

- **Performs Well On:** Clearly structured datasets with distinct class boundaries (e.g., MNIST, CIFAR-10).
- **Performs Poorly On:** Highly complex or overlapping distributions, subtle dataset shifts, or adversarially perturbed images that resemble in-distribution samples.

Future work should focus on refining feature representations, improving OOD separation, and enhancing adaptability across different datasets.