COMP90073 Security Analytics Assignment 2: Blue Team Report

Muhammad MD NASREIN

21 September 2024

1 Executive Summary

This report delves into the development and evaluation of anomaly detection (AD) and out-of-distribution (OOD) detection algorithms using the derived dataset of CIFAR-10. The primary objective is to identify anomalies and OOD samples accurately to prevent the integrity of the company's data from being compromised leading to undesirable damages towards the company.

Key findings indicate that anomaly detection using One-Class Support Vector Machines (OCSVM) and Isolation Forest models exhibit robust performance in detecting anomalies, with AUROC scores exceeding 0.9 on validation sets. Isolation Forest exhibits clearer separation of anomalous and normal sample. For deep learning models such as Variational Autoencoder (VAE) and Convolutional Autoencoder (CAE), VAE outperform CAE substantially with AUROC margin of 0.3. For OOD detection, Vision Transformer (ViT) with Mahalanobis distance calculated receive a AUROC score of 0.92.

2 Introduction

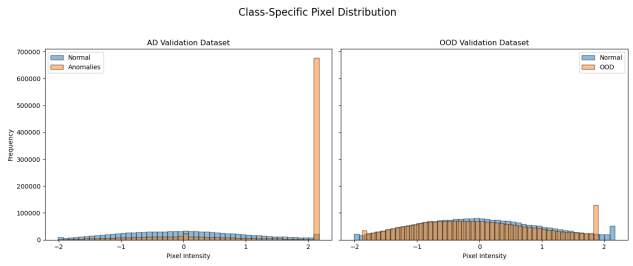
AD and OOD detection are critical tasks in cybersecurity, particularly for companies like Pixels.com, which rely on machine learning to process and validate large volumes of data. This report aims to develop and evaluate multiple algorithms to identify anomalies and OOD samples within the CIFAR-10 dataset, a collection of 32x32 color images across 10 categories normalised to pixel value of [-2,2] range.

For anomaly detection, we employed both shallow models, such as OCSVM and Isolation Forest, and deep learning models such as VAE and CAE. For OOD detection, we employed pre-trained base ViT with Mahalanobis distance as the OOD score. Our analysis includes a comprehensive evaluation of these models using metrics like AUROC, AUPRC, F1-score and FPR@95% TPR for OOD detection, alongside a critical examination of their discriminative performance and the display performance curves for the best interpretation .

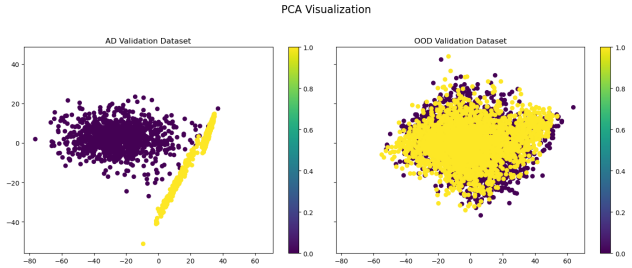
3 Methodology

3.1 Data Exploration

Our methodology includes the data exploration to understand different kinds of anomaly and OOD samples. Pixel value distributions are plotted for both anomalous and OOD samples against the normal samples in their respective validation dataset. We also plotted the heatmap for these value and the PCA plot to understand the nature of the anomalous and OOD sample and how different they are.



(a) AD (left) and OOD (right) pixel distribution

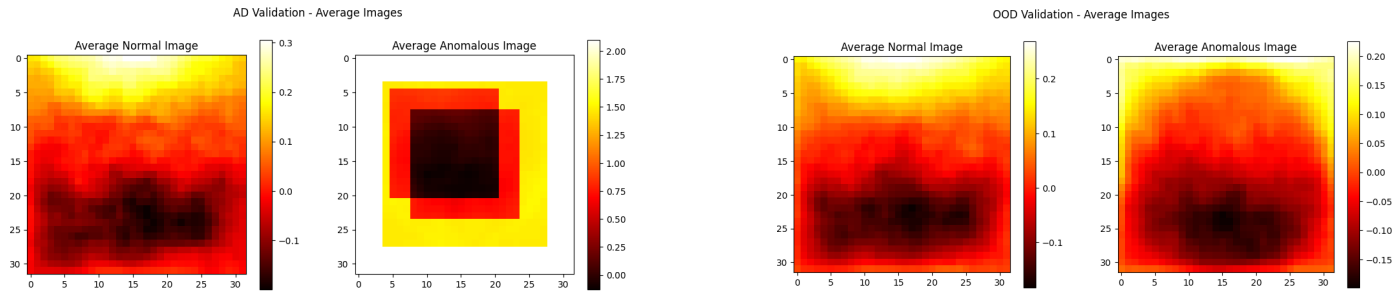


(b) AD (left) and OOD (right) PCA plot

Figure 1: Pixel distribution and PCA plot for the validation dataset

In the Figure 1 PCA visualization comparing the AD (Anomaly Detection) and OOD (Out-Of-Distribution) validation datasets, we observe distinct clusters. For the AD dataset, the anomalous points (yellow) form a clearly separated linear cluster, while the normal points (purple) are tightly grouped together, showing minimal overlap. This suggests a stark difference between the normal and anomalous samples in the AD dataset, making them relatively easy to distinguish. In contrast, the OOD dataset shows a more interspersed distribution between the normal and OOD points, indicating

greater overlap between these classes. This highlights the challenge in distinguishing OOD samples from the normal ones, as they seem to share more similarities in this projection.



(a) AD vs. normal samples heatmap

(b) OOD vs. normal samples heatmap

Figure 2: Pixel distribution and PCA plot for the validation dataset

The pixel intensity distribution histograms and heatmaps in Figure 1 and 2 respectively further illustrate these distinctions. In the AD dataset, the pixel intensity distribution for anomalies is sharply skewed, with a significant peak at one extreme, reflecting the high level of distortion in anomalous images. On the other hand, the OOD dataset presents a more balanced distribution, with normal and OOD samples having comparable intensities but showing variation in frequency across the pixel range. The heatmaps confirm these trends: the average normal image in both AD and OOD datasets exhibits similar intensity patterns, while the anomalous images differ markedly in AD but show subtler variations in OOD, reinforcing the greater challenge in distinguishing OOD samples compared to anomalies.

3.2 Data Preprocessing

For One-Class SVM (OCSVM) and Isolation Forest, preprocessing involves flattening each 2D image (e.g., 32x32 pixels) into a 1D vector, followed by feature scaling using StandardScaler to normalize the data (zero mean, unit variance). This step is crucial to improve the performance of distance-based (OCSVM) and tree-based (Isolation Forest) models. The scaler is fitted only on the training data to avoid data leakage, and the same transformation is applied to validation and test sets to maintain consistency.

For deep learning models like Variational Autoencoders (VAE), Convolutional Autoencoders (CAE), and Vision Transformers (ViT), preprocessing includes normalizing pixel values (e.g., [0, 1] or [-1, 1]) and, for convolutional models, reshaping images to meet input requirements. Data augmentation (e.g., flipping, cropping, rotations) is applied only to the training set to improve model robustness. For ViT, images are resized (commonly to 224x224 pixels) and normalized based on the pre-trained model's mean and standard deviation. During validation and testing, only normalization and reshaping are applied, with no data augmentation, ensuring consistent evaluation. This tailored preprocessing ensures optimal performance for each model type.

3.3 Training and Validation Process

The dataset used for training and validation is AD-val-dataset and OOD-val-dataset. These datasets wertr systematically divided into training and validation sets, typically following a split ratio such as 80% training and 20% validation. For OCSVM and Isolation Forest, the training set consisted exclusively of normal samples to enable the models to learn the boundary of normality. Hyperparameter tuning for these shallow models was conducted using a grid search approach, exploring combinations of key parameters like nu, kernel, and gamma for OCSVM, and n-estimators, max-samples, and contamination for Isolation Forest. To ensure robust parameter selection, a Stratified K-Fold cross-validation with 5 folds was employed, maintaining the proportion of normal and anomalous samples across each fold. The best-performing hyperparameters based on metrics such as AUROC,AUPRC and F1-score were then used to retrain the final models on the entire training dataset before evaluating their performance on the test set.

In contrast, VAE, CAE, and ViT are deep learning-based models that require more intensive training and validation strategies. These models utilized the same train-validation split, with the training set containing only normal samples for VAE and CAE. ViT train with both normal and OOD samples for outlier exposure (OE) that perform 80-20 train-validation split on both normal and OOD sample. Hyperparameter tuning for these architectures involved exploring parameters like latent dimension size, learning rates, batch sizes, and architectural configurations through methods such as random search or Bayesian optimization to efficiently navigate the extensive hyperparameter space. Due to the high computational demands of deep learning models, K-Fold cross-validation was generally impractical; instead, a single validation split was used alongside techniques like early stopping and model checkpointing to prevent overfitting and ensure optimal model performance. Once the optimal hyperparameters were identified based on validation performance metrics; AUROC, AUPRC and F1-score for VAE and CAE, and AUROC, AUPRC and False Positive at 95% True Positive (FP@95TR) for ViT .the final versions of the VAE, CAE, and ViT models were trained on the entire training dataset. These final models were then rigorously evaluated on the test set to assess their generalization capabilities in anomaly detection tasks.

4 Result and Analysis

4.1 Shallow Models Performance

For anomaly detection, two shallow models: OCSVM and Isolation Forest will be compared.

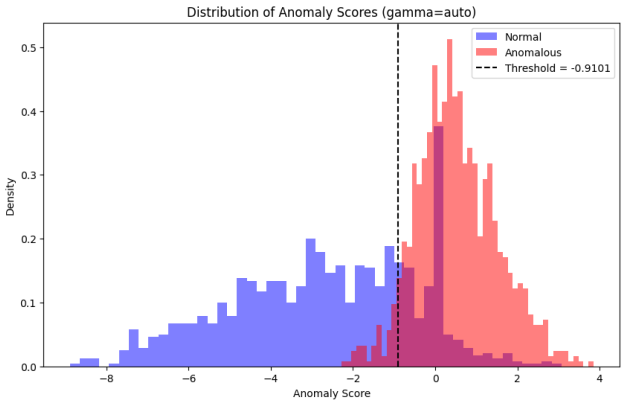
OCSVM is a kernel-based anomaly detection model that defines a decision boundary around normal data points, classifying those far from it as anomalies. It requires careful tuning of parameters like nu, which controls the expected proportion of anomalies, and gamma in the RBF kernel, which adjusts the boundary's flexibility. Proper tuning is crucial to avoid overfitting or underfitting, and the model can become computationally expensive for large datasets.

Isolation Forest is a tree-based model that isolates anomalies by randomly partitioning data points. It needs minimal tuning, with key parameters like n-estimators (number of trees), contamination (anomaly proportion), and max-samples (data subset size). These parameters make IsoForest efficient and scalable, particularly suited for large datasets with rare anomalies.

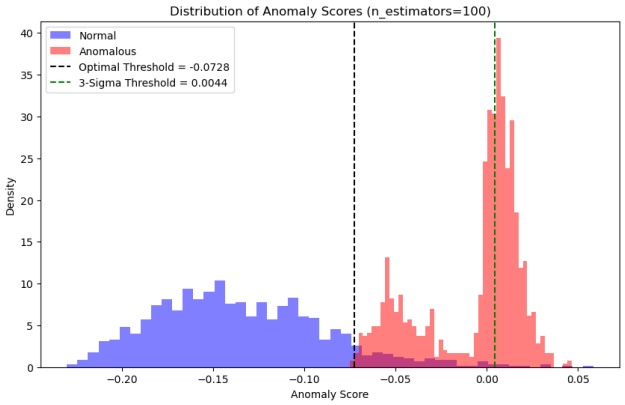
Model	Best Hyperparameters	Accuracy	Precision	Recall	F1-Score	ROC AUC	AUPRC
OCSVM	{'gamma': 'auto', 'kernel': 'rbf', 'nu': 0.1}	0.8495	0.7905	0.9510	0.8634	0.9173	0.8892
Isolation Forest	{'contamination': 0.01, 'max_samples': 100, 'n_estimators': 200}	0.9550	0.9190	0.9980	0.9569	0.9813	0.9607

Table 1: Comparison of OCSVM and Isolation Forest Best Hyperparameters Performance Metrics

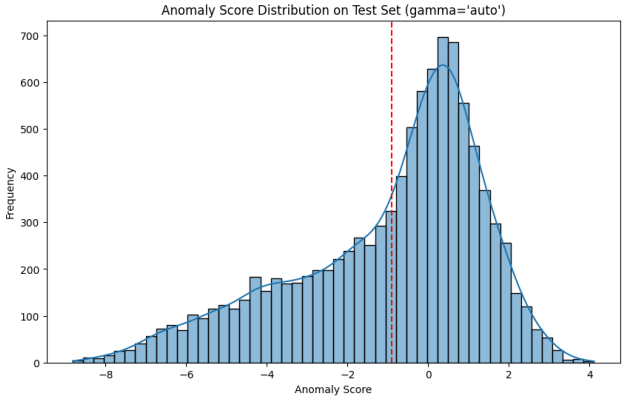
Table 1 shows Isolation Forest significantly outperforms OCSVM in nearly all metrics. It achieves higher accuracy (0.9550 vs. 0.8495), precision (0.9190 vs. 0.7905), and recall (0.9980 vs. 0.9510), leading to a stronger F1-score of 0.9569 compared to OCSVM's 0.8634. Additionally, IsoForest's area under the ROC curve (ROC AUC) is 0.9813, which is notably better than OCSVM's 0.9173, showing superior classification ability. Similarly, the area under the precision-recall curve (AUPRC) is higher for IsoForest (0.9607 vs. 0.8892), further highlighting its stronger precision-recall trade-off.



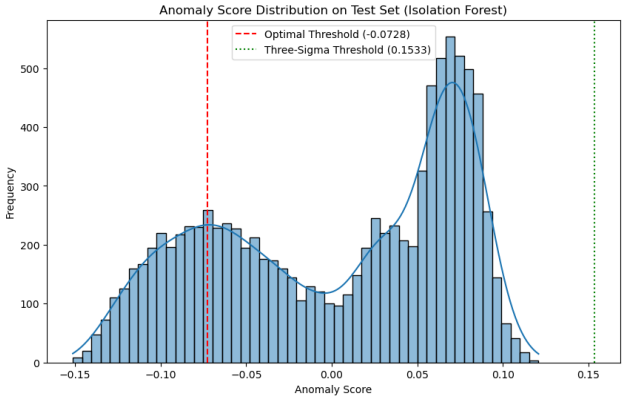
(a) OCSVM validation decision score distribution



(b) Isolation Forest validation decision score distribution



(c) OCSVM test decision score distribution



(d) Isolation Forest test decision score distribution

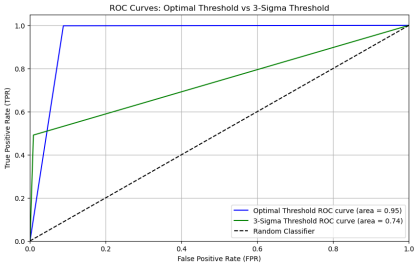
Figure 3: Decision score on validation and test set across both shallow models

The analysis in Figure 3 shows that Isolation Forest outperforms OCSVM in anomaly detection by providing clearer separation between normal and anomalous samples with minimal overlap around the optimal threshold (-0.0728), leading to fewer misclassifications. In contrast, OCSVM shows significant overlap, increasing the risk of misclassification near its decision boundary. Although the test set's anomaly score distribution for Isolation Forest shows a clear separation around 0, which makes it tempting to select 0 as the threshold, it's important not to adjust the threshold based on the

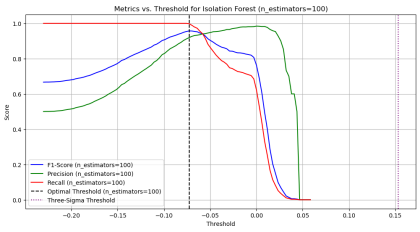
test data. Doing so introduces overfitting, as the test set should only be used for final evaluation, not for tuning model parameters. The threshold should be derived from the validation set to ensure the model generalizes well to unseen data. Choosing a threshold based on test set separation risks inflating performance on this specific dataset but may lead to poorer results on future, real-world data.

In terms of time complexity, OCSVM has a higher cost of $O(n^2)$, due to kernel matrix calculations and quadratic optimization, making it inefficient for large datasets. Isolation Forest, with a time complexity of O(t * n log n), where t is the number of trees and n is the number of data points, is more scalable and efficient, especially for large datasets. This is especially desirable when the cybersecurity department always viewed as cost center. Reducing unnecessary cost will shed a positive light on the department in the company.

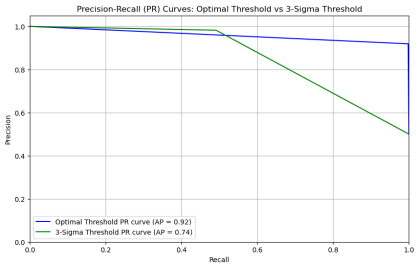
While Isolation Forest is more computationally efficient, threshold selection is a post-training task. To find the best threshold, methods such as the ROC curve, Precision-Recall curve, or statistical rules like the three-sigma rule can be used to balance precision and recall, minimizing false positives and false negatives. Therefore, we will use Isolation Forest to predict and annotate the test set.



(a) Isolation Forest ROC curve



(b) Isolation Forest Metric Thresholding



(c) Isolation Forest Precision-Recall curve

Figure 4: Isolation Forest ROC, Metric Thresholding, and Precision-Recall curves

The provided plots in Figure 4 compare the performance of Isolation Forest using two thresholds: the optimal threshold and the 3-Sigma threshold. The ROC curve shows that the optimal threshold yields an AUC of 0.95, indicating excellent separation between normal and anomalous samples, while the 3-Sigma threshold performs worse with an AUC of 0.74. The Precision-Recall (PR) curve further supports this, with the optimal threshold achieving an average precision (AP) of 0.92, compared to the lower AP of 0.74 for the 3-Sigma threshold, reflecting its poorer performance. The metric thresholding plot reveals that the optimal threshold maximizes the F1-score, providing a balanced trade-off between precision and recall. This shows that the optimal threshold is clearly superior, offering better anomaly detection accuracy with fewer false positives and a stronger overall performance than the 3-Sigma threshold.

4.2 Deep Learning Model Performance

We are motivated to use Variational Autoencoder (VAE) and Convolutional Autoencoder (CAE) as deep learning models for anomaly detection due to their ability to learn compressed, meaningful representations of data through unsupervised training. VAEs leverage a probabilistic framework to capture the underlying distribution of normal data, making them effective for identifying anomalies based on deviations from typical patterns. Similarly, CAEs utilize convolutional layers to capture spatial hierarchies in image data, enabling the detection of subtle anomalies through high-fidelity reconstruction of normal samples. Both models are well-suited for complex, high-dimensional datasets where traditional methods struggle to perform.

Table 2 shows that the VAE outperforms the CAE across all key metrics, with a higher F1-score (0.9911), AUROC (0.9857), and AUPRC (0.9942) compared to the CAE's F1-score (0.9510), AUROC (0.8374), and AUPRC (0.9214). The VAE's lower reconstruction error mean (0.9967) also reflects more accurate modeling of normal data.

We tune parameters like latent dimension to control the model's ability to compress data, hidden layers to manage model complexity, and learning rate to balance convergence speed and performance. The VAE's probabilistic framework benefits from this tuning, as it helps model data variability and detect anomalies more effectively. In contrast, the CAE's deterministic nature makes it less capable of capturing the underlying distribution, leading to lower performance. Based on these metrics and hyperparameter tuning, the VAE is the better choice for anomaly detection.

In Figure 5, the VAE shows clearer separation between normal and anomalous samples compared to the CAE, as seen in its well-defined peaks and optimal threshold of 1.88. On the test set, the VAE maintains this separation, while the CAE exhibits more overlap between the two classes near its threshold of 0.66, leading to potential misclassifications. The VAE's reconstruction errors are more compact, while the CAE's errors for anomalies are spread out, indicating less consistent detection performance.

The VAE's superior performance likely stems from its probabilistic framework, allowing it to model data variability more effectively. By learning a latent space that captures normal data patterns, the VAE distinguishes anomalies based on statistical deviations. In contrast, the CAE, while good at capturing spatial features, lacks this probabilistic modeling, leading to more overlap in reconstruction errors and less precise anomaly detection. Therefore, VAE has been used to annotate the AD test set.

Model	Metric	Value	Best Hyperparameters
VAE	Best Validation Reconstruction Error Mean	0.9967	{'epochs': 30, 'hidden_dims': [512, 256], 'latent_dim': 2, 'learning_rate': 0.001}
	Best Threshold	1.8821	Same as above
	F1-Score	0.9911	Same as above
	AUROC	0.9857	Same as above
	AUPRC	0.9942	Same as above
	Precision	0.9833	Same as above
	Recall	0.9990	Same as above
CAE	Best Validation Reconstruction Error Mean	1.1465	{'epochs': 30, 'hidden_dims': [32, 64, 128], 'latent_dim': 4, 'learning_rate': 0.0001}
	Standard Deviation	0.3399	Same as above
	Optimal Threshold for Anomaly Detection	0.6582	Same as above
	F1-Score	0.9510	Same as above
	Precision	0.9066	Same as above
	Recall	1.0000	Same as above
	AUROC	0.8374	Same as above
	AUPRC	0.9214	Same as above

Table 2: Performance Metrics and Best Hyperparameters for VAE and CAE Models

The ROC and PR curves both show excellent performance, with AUC and AUPRC of 0.99. The model achieves a high F1-score of 0.99, indicating strong precision and recall. This suggests the model effectively detects anomalies with minimal false positives and negatives, making it highly reliable for anomaly detection.

The confusion matrix in Figure 6 comparison shows that VAE outperforms Isolation Forest in terms of misclassification. VAE has no false positives, correctly classifying all anomalies without misclassifying any normal data as anomalous. In contrast, Isolation Forest misclassifies 2472 normal instances as anomalies, indicating a higher false positive rate. Both models correctly identify 5419 true anomalies, but Isolation Forest's sensitivity to normal data leads to a higher error rate in detecting false positives. Overall, VAE provides better performance with fewer misclassifications, making it the more precise model in this case.

4.3 OOD Detection Performance

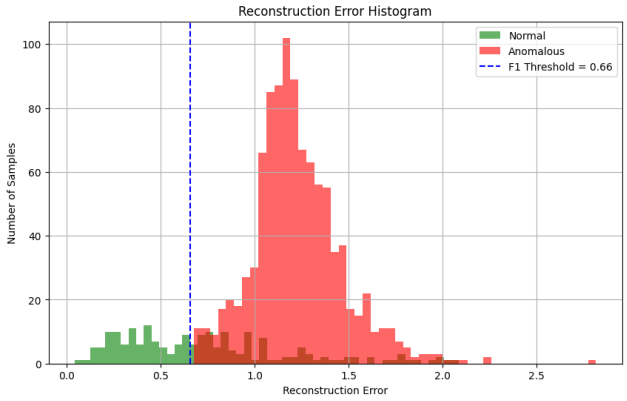
The Vision Transformer (ViT) was chosen for its ability to capture global relationships within images, making it highly effective for distinguishing between normal and out-of-distribution (OOD) samples. ViT's self-attention mechanism allows it to generalize better in OOD scenarios compared to convolutional models. Mahalanobis distance is used as the scoring system for its ability to quantify how far a sample is from the normal data distribution, enhancing OOD detection. This combination leverages ViT's global feature extraction with Mahalanobis distance's precise anomaly scoring for more accurate OOD detection.

Metric	Value
Precision	0.9276
Recall	0.9220
F1-Score	0.9248
AUROC	0.9668
AUPRC	0.9586
Mean Validation Accuracy	0.9240
Best Hyperparameters	{'learning_rate': 0.0001, 'optimizer': 'AdamW', 'weight_decay': 0.0001}

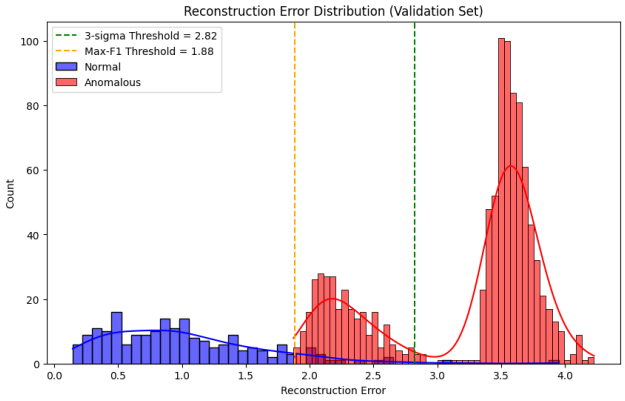
Table 3: ViT Performance Metrics and Best Hyperparameters

The ViT model demonstrates strong performance as shown if Table 3 with a Precision of 0.9276, Recall of 0.9220, and an F1-Score of 0.9248, indicating a balanced ability to correctly classify both normal and OOD samples. The high AUROC (0.9668) and AUPRC (0.9586) further reflect its strong discriminatory power, making it highly reliable for OOD detection. The mean validation accuracy of 0.9240 confirms consistent performance. The model's best performance was achieved with the AdamW optimizer, learning rate of 0.0001, and weight decay of 0.0001, highlighting effective hyperparameter tuning.

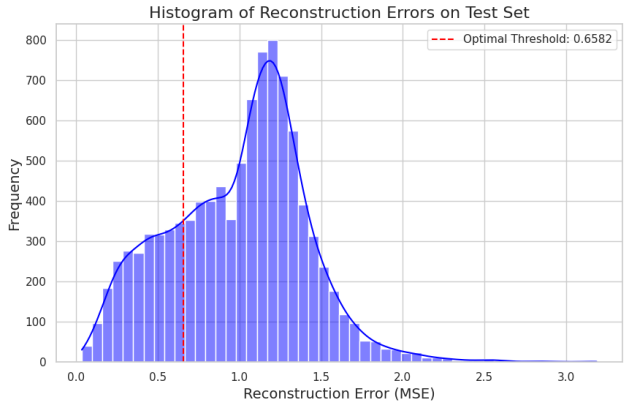
Based on Figure 7, the ViT OOD detection using Mahalanobis distance demonstrates strong performance, as indicated by the clear separation between in-distribution (ID) and out-of-distribution (OOD) samples in both the validation and test sets. The validation score distribution shows distinct peaks for ID and OOD samples, with minimal overlap near the threshold of 77.6099, reflecting effective classification. The test set distribution mirrors the validation set, suggesting



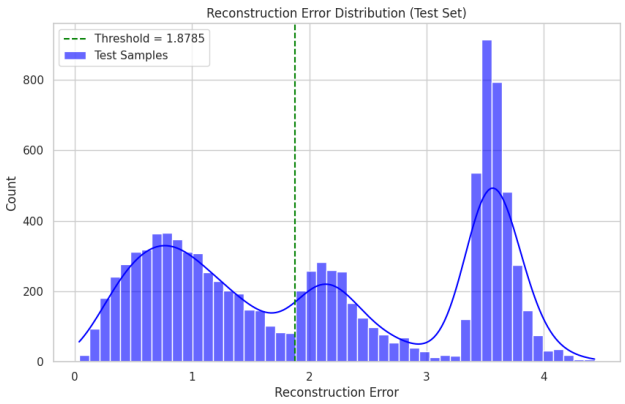
(a) CAE validation reconstruction error score distribution



(b) VAE validation reconstruction error score distribution



(c) CAE test reconstruction error score distribution



(d) VAE test reconstruction error score distribution

Figure 5: Reconstruction error score on validation and test set across both deep learning models

that the model generalizes well to unseen data. However, the slight overlap between the two distributions indicates a small potential for misclassification, though this is mitigated by the model's strong overall performance.

The FPR@95% TPR metric indicates the False Positive Rate (FPR) when the True Positive Rate (TPR) is fixed at 95%. In this case, the model has an FPR of 0.1080, meaning that at 95% TPR, approximately 10.8% of normal samples are incorrectly classified as OOD. The threshold at 95% TPR is 77.6099, which is the Mahalanobis distance score used to maintain a high TPR while controlling for false positives.

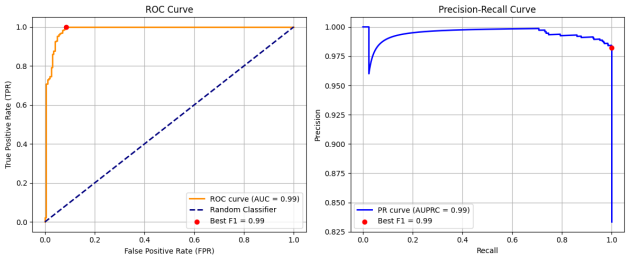
The ROC curve, with an AUC of 0.9668, and the PR curve, with an AUPRC of 0.9586, further confirm the model's high accuracy in distinguishing ID from OOD samples. The ROC curve shows a low false positive rate, while the PR curve demonstrates that precision remains consistently high as recall increases. Despite the small area of overlap, the ViT model, paired with Mahalanobis scoring, provides excellent OOD detection performance.

t-SNE plot in Figure 8 visually supports the quantitative performance of the model seen in the ROC and PR curves. The clear clustering of normal and OOD samples in this t-SNE plot aligns with the model's high AUROC (0.9668) and AUPRC (0.9586), indicating strong discrimination capabilities. The separation shown in the t-SNE plot is consistent with the minimal overlap observed in the Mahalanobis score distributions, further reinforcing the conclusion that the ViT model, with Mahalanobis distance scoring, is highly effective at OOD detection.

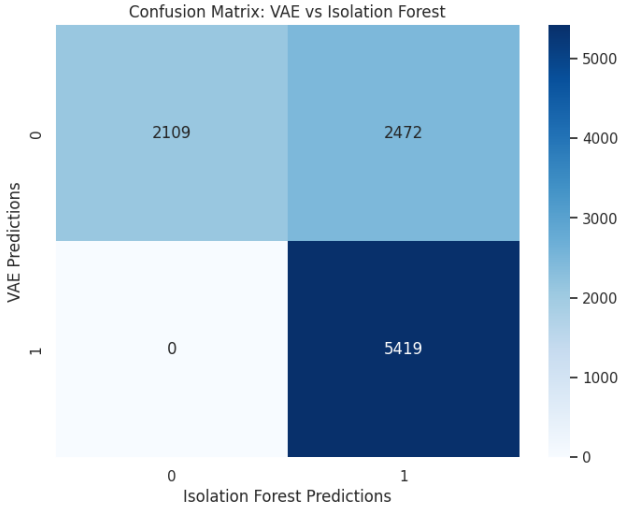
5 Conclusion and Future Work

The analysis of both shallow models (OCSVM and Isolation Forest) and deep learning models (CAE and VAE) for anomaly detection reveals that deep learning models generally outperform the shallow ones in distinguishing between normal and anomalous samples. The VAE, in particular, achieves superior results with a high F1-score (0.9911), AUROC (0.9857), and AUPRC (0.9942), making it the most reliable among all models tested. As for OOD detection, the t-SNE visualization and Mahalanobis distance scoring further validate the effectiveness of the ViT model in OOD detection, showing clear separation between normal and OOD samples. In contrast, while Isolation Forest and OCSVM perform reasonably well in AD detection, they exhibit higher false positive rates compared to the deep learning models. This makes the ViT and VAE models preferable for detecting anomalies and OOD samples, especially in high-dimensional image datasets where complex feature extraction is critical.

Future work can focus on enhancing both shallow and deep learning models for anomaly detection. For shallow models, integrating ensemble techniques or developing hybrid models that combine tree-based methods with deep learning can reduce false positives and improve accuracy. In deep learning models like CAE and VAE, further refining thresholding



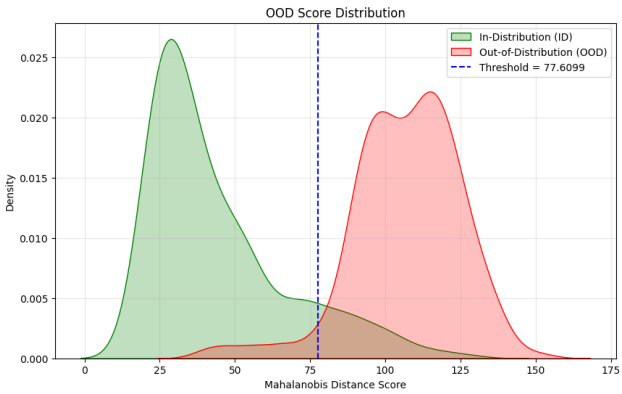
(a) VAE ROC and Precision-Recall curve



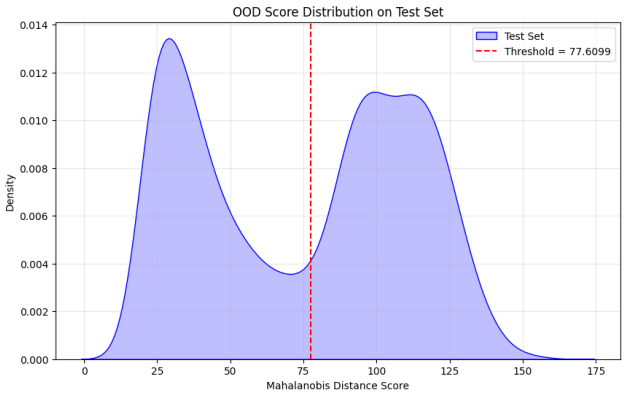
(b) Confusion matrix of Isolation Forest and VAE test prediction

Figure 6: VAE ROC and Precision Recall curve (left), Confusion Matrix for VAE and IF test prediction

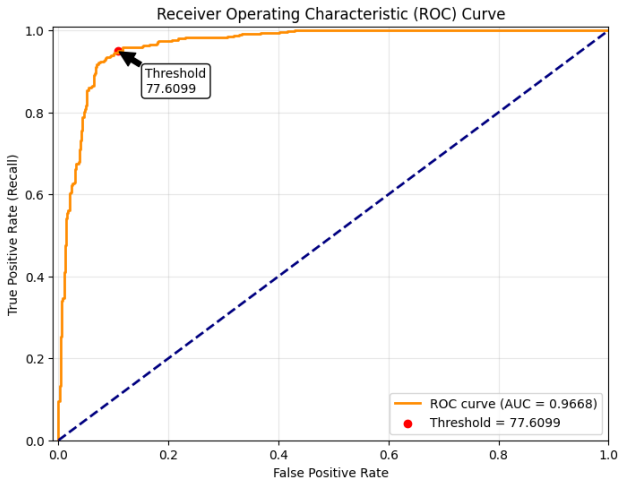
methods (e.g., FPR@95% TPR) can sharpen decision boundaries and lower misclassifications. Exploring unsupervised approaches for anomaly detection without labeled data, as well as improving model interpretability, are key areas to pursue. Moreover, testing across more diverse datasets and anomaly types will enhance model generalization. Additional strategies include adopting self-supervised learning to boost performance with limited data, and using active learning to prioritize labeling of ambiguous cases, enabling more efficient anomaly detection in real-world scenarios.



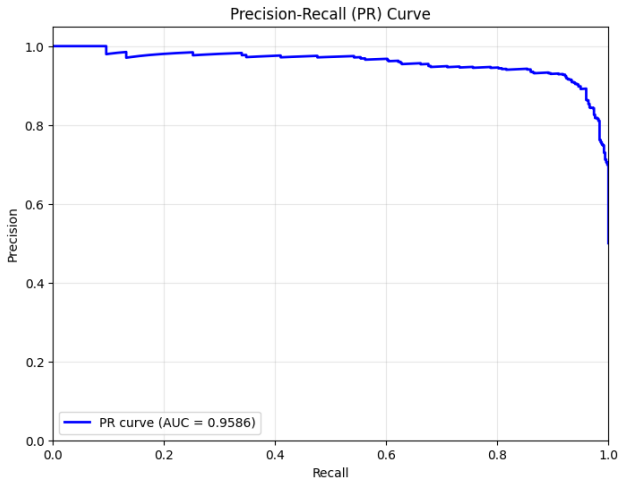
(a) ViT validation Mahalanobis score distribution



(b) ViT test decision Mahalanobis distribution



(c) ViT ROC curve



(d) ViT Precision-Recall curve

Figure 7: ViT Mahalanobis score disdtribution, ROC curve and Precision-Recall curve

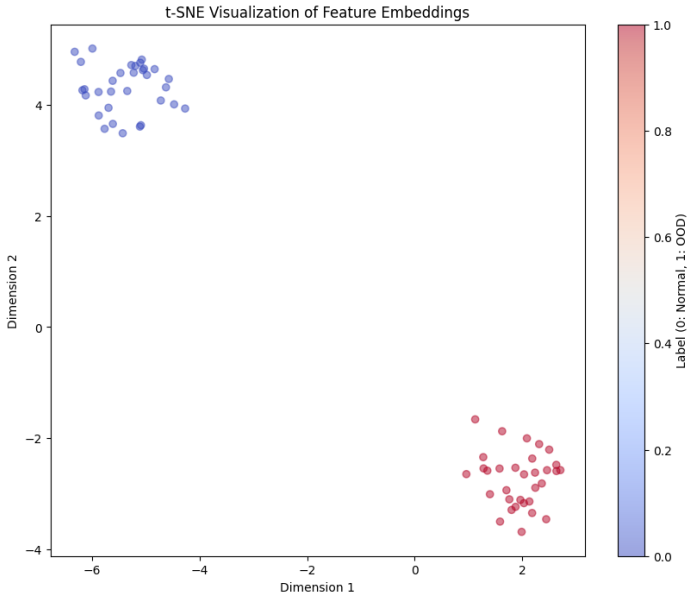


Figure 8: -SNE Visualization of Feature Embeddings for Normal and Out-of-Distribution (OOD) Samples