Safeguarding Training Data Integrity through Anomaly and Out-of-Distribution Detection

Jessica Sihite – 1275588 – COMP90073: Trustworthy Machine Learning

1,647 words

October 18, 2025

Executive Summary

Mosaic.com's competitive advantage relies on automated image classification systems that leverage product categorisation and inventory. These machine learning models continuously improve through incremental retraining on user-uploaded images to adapt to evolving product trends and visual characteristics unavailable in initial training data. However, this strategy introduces a counterintuitive vulnerability. By opening our training pipeline to user-contributed data for accuracy improvements, we inadvertently expose the system to data integrity attacks that can degrade classification performance.

We identified two distinct threat vectors that can compromise our training pipeline. First, our models are susceptible to accepting anomalous samples, which are legitimate product images that have been corrupted. Second, out-of-distribution (OOD) samples represent content completely unrelated to our expected product categories, uploaded either through user error or malicious intent. If undetected, both threats cascade misclassifications across our client base with potential liability and reputational consequences.

To address these risks, we established a Blue Team developing detection systems to filter compromised data before training pipeline ingestion. This report presents three detection algorithms designed to safeguard data quality: two anomaly detection approaches (shallow and deep learning models) and one OOD detector, alongside their performance evaluation and operational recommendations.

Methodology

Data

We evaluated detectors on CIFAR-10 grayscale 32x32 images. Separate modified versions were created to mirror the two distinct threat scenarios. The **anomaly detection validation set** contains 2,000 labelled images: 1,000 normal samples and 1,000 anomalous images representing legitimate product images that have been corrupted through noise injection, compression artifacts, rotation, cropping, and transmission distortions. The **anomaly detection test set** contains 10,000 unlabelled images maintaining similar corruption types and distribution as the validation set. The **OOD detection validation set** contains 5,000 labelled images: 2,500 in-distribution product images that match our 10 expected categories and 2,500 OOD images, including human faces and non-product objects. The **OOD detection test set** contains 10,000 unlabelled images with a similar distribution to the corresponding validation set.

Methods

Data Pre-processing and Feature Extraction

For shallow anomaly detection and OOD detection, we extracted 512-dimensional feature representations from the pre-trained ResNet-8 model to obtain semantically meaningful representations, enabling efficient detection. Deep learning models for anomaly detection operated directly on 1,024-dimensional raw pixel values.

Algorithms

Shallow Anomaly Detector: One-Class Support Vector Machine (OCSVM)

OCSVM learns a decision boundary encapsulating normal data without requiring anomalous training samples. It maps 512-dimensional features into a high-dimensional space and constructs a hyperplane separating normal data from the origin, flagging samples beyond this boundary as anomalies. We trained OCSVM exclusively on 512-dimensional features extracted from the CIFAR-10 training images, optimising two critical hyperparameters through grid search: the **nu** parameter that controls outlier fraction bounds (lower **nu** values reduce false positives at the cost of increased false negatives, while higher **nu** values produce more permissive boundaries, improving recall at the cost of precision) and kernel functions (linear for efficiency and interpretability, while RBF kernels capture non-linear patterns but risk overfitting in high dimensions). Selected configuration was **nu** = 0.1 with linear kernel. We also implemented an alternative algorithm, Local Outlier Factor (LOF), which computes the ratio of a sample's density to its neighbours' densities and identifies samples in sparse regions as anomalies.

Deep Anomaly Detector: Variational Encoder (VAE)

While shallow models use fixed features, deep learning models simultaneously learn hierarchical representations and detection mechanisms. Autoencoders compress normal data into latent representations, then reconstruct accurately. Anomalous samples produce high reconstruction errors as corruption patterns absent from training cannot be recreated. VAE extends this by modelling latent space probability distributions rather than deterministic codes, providing regularisation and uncertainty quantification. We implemented VAE with fully connected architecture on flattened 1,024 dimensional inputs. The encoder projects inputs into mean and log-variance vectors parametrising a 16-dimensional Gaussian latent distribution. The decoder transforms sampled codes through 256 hidden units to 1,024-dimensional reconstructions with sigmoid activation, forcing efficient compression. For comparison, we also implemented Standard Autoencoder and Denoising Autoencoder (DAE) that explicitly corrupts inputs with additive Gaussian noise during training.

Post-hoc OOD Detector: Mahalanobis Distance

Post-hoc OOD detection methods provide computational efficiency by leveraging pre-trained classifiers without retraining. Among post-hoc methods, Mahalanobis distance assumes features from in-distribution samples form class-conditional Gaussian distributions in the learned feature space. We extracted 512 features from the pre-trained model for all CIFAR-10 training images and computed class-specific mean vector and covariance matrix. Mahalanobis distance measures how many standard deviations a sample lies from the nearest class distribution, accounting for feature correlations through the inverse covariance matrix. Samples whose minimum distance across all classes exceeds the threshold are classified as OOD. We compared Mahalanobis against another post-hoc OOD detection method, Maximum Softmax Probability (MSP).

Validation

In addition to accuracy, precision, recall, and F1-score, we assessed performance using Area Under the Receiver Operating Characteristics Curve (AUROC), measuring the algorithm's ability to discriminate across all thresholds, and Area Under the Precision-Recall Curve (AUPRC), which captures performance on imbalanced datasets. For OOD detection tasks, we also calculated False Positive Rate at 95% True Positive Rate (FPR@95%TPR), indicating the proportion of in-distribution samples incorrectly flagged when detecting 95% of OOD samples. Thresholds were optimised via Youden's index, maximising TPR minus FPR. Following validation, the best-performing models were applied to test sets for final predictions.

Results and Discussion

Based on validation performance presented in Table 1, OCSVM was selected over LOF for shallow detection and VAE over standard AE and DAE for deep detection. However, the choice between OCSVM and VAE for production deployment involves business trade-offs beyond raw metrics. OCSVM's higher AUROC and AUPRC indicate superior ranking quality where anomalies consistently score higher than normal samples across the score distribution, providing operational robustness when upload patterns shift due to seasonal trends or new retail partnerships. Conversely, VAE achieved higher accuracy and recall, detecting 37 more corruptions than OCSVM. However, VAE's superior performance in detecting these corruptions concentrates around its optimised threshold. When upload patterns change, recalibration may be needed, requiring additional engineering overhead.

As illustrated in Table 2, post-hoc Mahalanobis detector achieved 8.2 times lower false positive rate than MSP at 95% sensitivity. The performance gap stems from architectural differences. Both approaches utilise the same pre-trained model that exhibits systematic overconfidence on boundary cases. Legitimate uploads with atypical lighting or unusual viewpoints from mobile captures reduce softmax confidence despite being valid inventory items, producing high false positive rates that damage client trust. Contrastingly, Mahalanobis operates in the model's intermediate 512-dimensional feature space, computing statistical distance to class distributions rather than relying on confidence calibration. This design leverages the space where the classifier exhibits 93% accuracy, maximising consistency between OOD detection and classification systems. This provides robustness critical for Mosaic.com's diverse client base.

However, Figures 1 and 2 expose contrasting failure modes between feature-space and pixel-space detection. OCSVM, operating on classification-optimised features, exhibited 53 false negatives compared to VAE's 16, demonstrating that compressed features miss pixel-level corruptions. Although [1] shows that penultimate layer features provide semantically rich representations optimal for downstream classification tasks, the 512-dimensional compression struggles to detect pixel-level corruptions such as blur, compression artifacts, and noise visible in Figure 4 that VAE's raw 1,024-dimensional input preserves. This architectural difference manifests in complementary strengths where OCSVM achieved superior ranking quality and threshold stability as evident in higher AUROC and AUPRC, while VAE detected 37 additional corruptions through pixel-level reconstruction errors despite marginally lower ranking metrics. Both methods exhibit comparable false positive rates (OCSVM = 104; VAE = 100), indicating convergent failure modes on unusual but legitimate semantic patterns, such as extreme lighting conditions, atypical object orientations, and compositionally rare scenes visible in Figures 1 and 2. This suggests the false positive problem stems from insufficient training diversity, while false negative disparity reflects architectural trade-offs between feature granularity and ranking consistency. Both detectors required critical design compromises where OCSVM's nu parameter bounds outlier fraction to 10% to match anticipated anomaly prevalence, while VAE's 16-dimensional latent space balances reconstruction fidelity against overfitting risk.

Figure 3 reveals systematic failure patterns where Mahalanobis conflated intra-class diversity with distributional shifts. False positives predominantly feature legitimate CIFAR-10 objects from atypical viewpoints, such as trucks captured from overhead angles and animals at image edges or in unusual poses. These remain semantically valid class members but fall outside the narrow Gaussian distributions learned from prototypical perspectives, as shown in Figure 5 where in-distribution samples exhibit consistent canonical orientations. Frontal boat viewpoints produced Mahalanobis distances exceeding detection thresholds because the learned distributions model exclusively broadside orientations, leaving frontal angles in unmapped feature space interpreted as OOD. This represents binary distributional failure where underrepresented viewpoints in training data cause legitimate images captured from these angles to be incorrectly flagged as OOD. For Mosaic.com, this manifests when retail clients photograph inventory from diverse contexts, such as bulk storage verification versus damage documentation.

Mahalanobis' false negatives expose a fundamentally different failure mechanism through geometric mimicry. Figure 3 shows multiple human faces and standing figures incorrectly classified as in-distribution despite being clear OOD samples in Figure 5. These human images share critical geometric properties with animal categories, such as vertical bilateral symmetry (centered facial features, upright postures), appendage structures (arms resembling limbs), and edge-based representations (silhouettes activating identical contour detectors). While the 512-dimensional features capture semantic concepts for within-distribution classes, they reduce to geometric primitives when encountering out-of-distribution samples that share structural similarities with learned classes. Consequently, the representation struggles to distinguish human faces from animal faces when both exhibit bilateral symmetry because both activate identical mid-level feature detectors learned for within-distribution classification, creating exploitable blind spots where OOD samples sharing geometric patterns with learned classes produce small Mahalanobis distances indistinguishable from legitimate samples.

Conclusion

Mosaic.com's training pipeline vulnerabilities necessitate multi-layered defence architectures. OCSVM delivered threshold-stable anomaly detection across dynamic upload patterns, while VAE captured pixel-level corruptions, detecting 37 additional anomalies despite lower AUROC. Although Mahalanobis distance reduced false positive rates 8.2-fold compared to MSP, failure analysis exposed fundamental architectural constraints: compressed feature representations sacrifice corruption detection granularity, Gaussian distributional assumptions break under viewpoint diversity, and geometric feature encoding enables mimicry-based evasion. These findings suggest production deployment would benefit from ensemble approaches combining pixel-space and feature-space detection, training data augmentation incorporating diverse viewpoints, and continuous threshold recalibration as upload distributions evolve seasonally.

References

[1] S. Kornblith, J. Shlens, and Q. V. Le, “Do better ImageNet models transfer better?,” in *Proc. IEEE/CVF CVPR*, 2019, pp. 2661-2671. Available: https://openaccess.thecvf.com/content_CVPR_2019/papers/Kornblith_Do_Better_ImageNet_Models_Transfer_Better_CVPR_2019_paper.pdf

Appendix

Table 1

Anomaly Detection Performance Evaluation on Validation Set

Model	Algorithm	Evaluation Metrics					
		Accuracy	Precision	Recall	F1-Score	AUROC	AUPRC
Shallow	OCSVM	0.922	0.901	0.947	0.924	0.969	0.945
	LOF	0.524	0.551	0.259	0.352	0.556	0.535
Deep Learning	VAE	0.9420	0.908	0.984	0.944	0.963	0.929
	Standard AE	0.610	0.677	0.420	0.519	0.677	0.639
	DAE	0.676	0.653	0.751	0.698	0.744	0.738

Table 2

Out-of-Distribution Performance Evaluation on Validation Set

Algorithm	Evaluation Metrics						
	Accuracy	Precision	Recall	F1-Score	AUROC	AUPRC	FPR@95%TPR
Mahalanobis Distance	0.951	0.977	0.924	0.950	0.982	0.987	0.057
MSP	0.821	0.813	0.834	0.823	0.871	0.849	0.468

Figure 1

OCSVM Misclassification Examples on Validation Set: False Positives and False Negatives

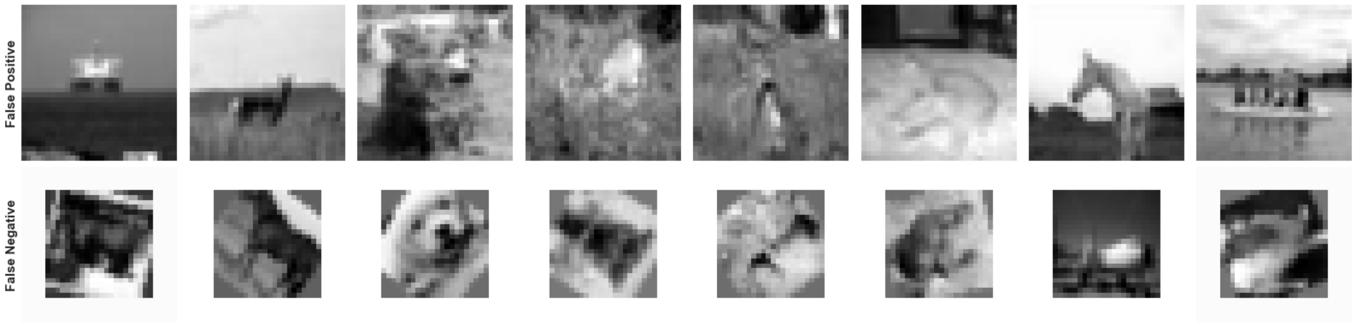

Figure 2

VAE Misclassification Examples on Validation Set: False Positives and False Negatives

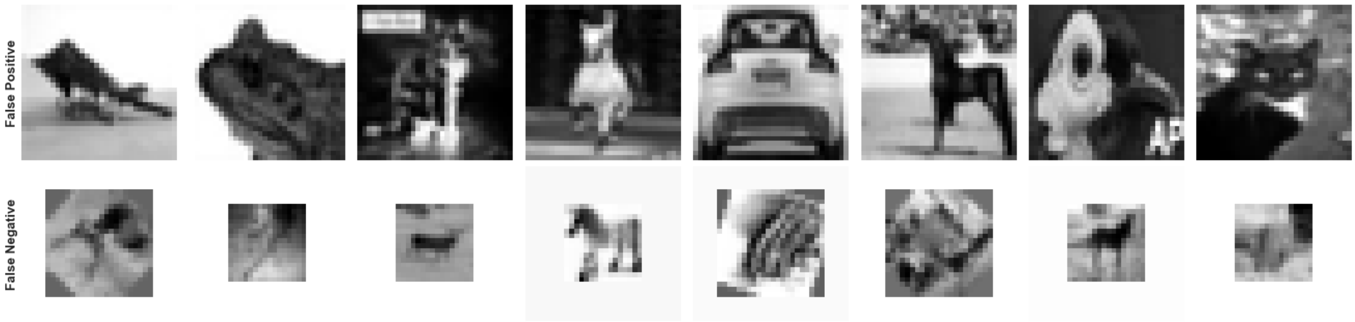

Figure 3

Post-hoc Mahalanobis Misclassification Examples on Validation Set: False Positives and False Negatives

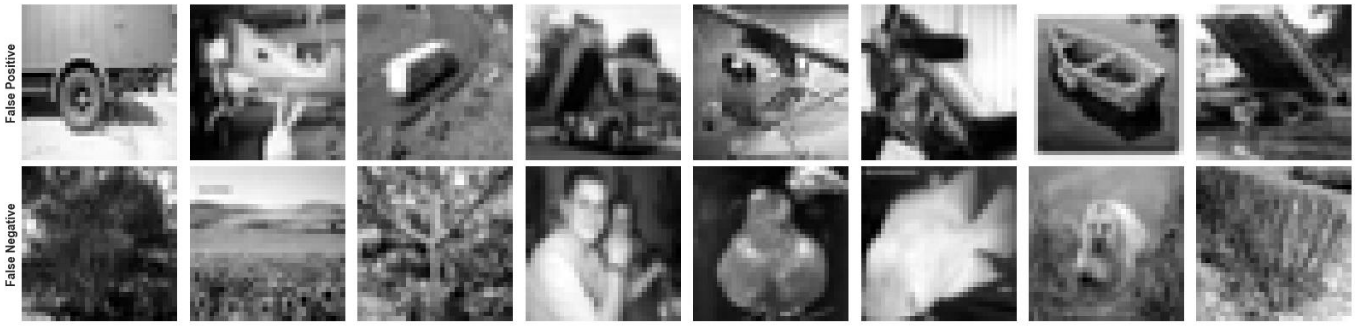

Figure 4

Anomaly Detection Validation Set Visualisation

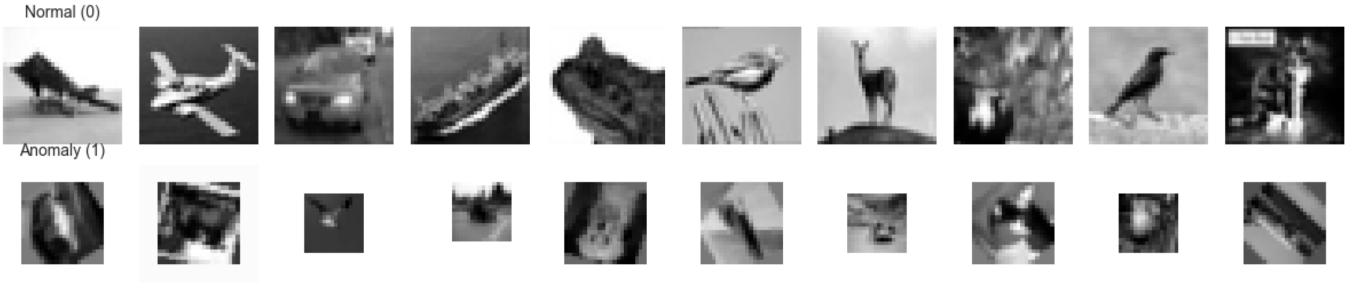

Figure 5

OOD Detection Validation Set Visualisation