



Out of Distribution Detection for Reliable Deployment of Machine Learning Models

Received: July 02, 2026

Revised: July 28, 2026

Accepted: September 16, 2026

Publish: September 28, 2026

William Yeoh*, Fadly, Maria Ulfa

Abstract:

Background: Machine learning models can give confident predictions when they receive inputs that are different from the data used during training. These predictions may not always be reliable when the model encounters unfamiliar inputs. OOD detection can help identify such inputs and reduce the risk of incorrect predictions in real-world applications.

Aims: This study evaluates several OOD detection methods and examines their ability to reduce unreliable automatic predictions.

Methods: Several post hoc detection methods were evaluated using ResNet classifiers trained on CIFAR datasets. The methods included Maximum Softmax Probability, ODIN, Energy-Based Detection, Mahalanobis Distance, Deep k Nearest Neighbors, and ReAct. The evaluation covered near-OOD, far-OOD, and corrupted-in-distribution conditions. Detection performance was also examined together with prediction errors, review needs, processing time, and memory use.

Results: The methods showed different performance across the tested conditions. Deep k-Nearest Neighbors performed well for near OOD inputs, while ReAct showed strong performance for far OOD detection. The deployment simulation showed that ReAct-based gating reduced OOD leakage and improved the accuracy of predictions that were processed automatically. It also reduced incorrect predictions made without additional review. Energy-based detection provided competitive results with relatively low computational requirements.

Conclusion: OOD detection can improve the reliability of machine learning systems. When combined with abstention and fallback mechanisms, unfamiliar inputs can be identified and sent for further review. This approach reduces unreliable predictions while allowing familiar inputs to continue through the automated process.

Keywords: Deployment Reliability; Machine Learning; Model Robustness; Out-Of-Distribution Detection; Uncertainty Estimation.

1. INTRODUCTION

Machine learning models are increasingly deployed in environments that are more dynamic and less predictable than the controlled conditions under which they were developed. During training and validation, a model generally assumes that future inputs will follow a distribution similar to that of the training data. This closed-world assumption is difficult to maintain after deployment because changes may occur in data sources. Consequently, a deployed model may receive inputs that differ substantially from those observed during training.

user behavior, operating environments, sensing devices, population characteristics, or application requirements. Such inputs are commonly referred to as out-of-distribution (OOD) data, whereas samples drawn from the training-related distribution are considered in-distribution (ID) data (Gwon & Yoo, 2023; Kaur et al., 2022; B. Wu et al., 2023; Yang et al., 2024; Zhao & Cao, 2024). Detecting these changes is important because pre-deployment evaluations cannot fully represent real-world dynamics, and post-deployment monitoring is needed to identify unexpected inputs, distribution shifts, and unreliable system behavior (Rao et al., 2026).

The presence of OOD inputs creates a fundamental reliability problem. A machine learning classifier may assign a known class label to an unfamiliar input even when the input lies outside the model's domain of competence. More importantly, modern neural networks can produce highly confident predictions for samples that are statistically or semantically different from the training data. In such situations, predictive confidence does not necessarily represent actual reliability. An incorrect but highly confident prediction can be particularly problematic when the model is used to support consequential decisions. An OOD detector is

Publisher Note:

CV Media Inti Teknologi stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Copyright

©2026 by the author(s).

Licensee CV Media Inti Teknologi, Indonesia. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution-ShareAlike (CCBY-SA) license (<https://creativecommons.org/licenses/by-sa/4.0/>).

therefore expected to identify potentially unfamiliar inputs before the primary prediction is accepted, allowing the system to abstain, request additional verification, or direct the input to an alternative processing mechanism (Han & Li, 2022; Lahoti et al., 2023; Tosaki et al., 2025; H. Wang et al., 2025; Xia & Bouganis, 2024).

Early OOD detection research introduced probability-based approaches that use information obtained directly from the classifier's output. Luo et al. (2023), for example, demonstrated that the maximum softmax probability could serve as a simple baseline for distinguishing ID inputs from misclassified or OOD samples. Berger et al. (2021) advanced the discussion with ODIN, which uses temperature scaling and small perturbations to better differentiate the ID and OOD confidence distribution overlap. Although softmax-based techniques continue to be susceptible to overconfidence in model predictions, as high values can be achieved for well outside the training distribution inputs. To counter this, Liu et al. (2020) proposed an energy-based scoring function that better delineates the observed from the unobserved input in a more formal sense. These works established the family of probabilistic scoring functions for OOD detection and demonstrated that relying solely on a classifier's confidence may offer an inadequate basis for a safe system deployment.

Other research focuses on the model's internal representations and their geometry. (Lee et al., 2018) estimated class-specific feature distributions and computed the Mahalanobis distance to assess the nearness of a given test sample to the known classes. Distance-based detection was subsequently implemented in a non-parametric fashion; for example, Sun et al. (2022) computed nearest-neighbor distances in the feature space that the model has learned, making no class-conditional Gaussian assumption. Representation-based methods look at intermediate model activations, the structure and properties of the deep layers, and the degree of embedding that may differentiate between seen (known) and unseen (unknown) inputs. Outlier Exposure is another training-based method for improving out-of-distribution (OOD) awareness by training a model on auxiliary outlier data, though its success is influenced by the quality and quantity of the auxiliary data (Chen et al., 2024; Hendrycks & Dietterich, 2019; Kim et al., 2024; Yu et al., 2024). From the perspective of these studies, it appears that OOD awareness may be derived from an output probability, a distance measure in a particular feature set, and the organization of the learned representation.

One of the more significant challenges, even with the advancements that have been made, is determining which OOD detector is appropriate for real-world use. Reported OOD detection methodologies differ in the datasets used, model architectures, preprocessing, selection of thresholds, and how different types of distribution shifts are defined. OpenOOD provides a

benchmark with over 30 OOD detection methodologies and related problem settings to facilitate ease of comparison (Yang et al., 2022). However, recent studies have shown that excellent performance in a traditional benchmark is not necessarily applicable to other types of distribution shifts. Guille-Escuret et al. (2023) demonstrated that OOD detectors that encompass an effective detection of novel class scenarios perform poorly in the presence of data corruption, adversarial data, synthetic data, and a host of other novel and challenging scenarios. Consequently, the performance of an OOD detector should not be evaluated based on a single dataset or a single OOD sample.

Research into OOD detection is frequently constrained by a clear focus on OOD detection as a standalone discrimination task (Guérin et al., 2023; Lu et al., 2025a; Nie et al., 2025; Yang et al., 2024). For example, the area under the curve (AUC) and the false positive rate at a fixed true positive rate are helpful for evaluating and ranking different scoring functions, but do not help evaluate and rank different detector systems for an inference task. Detection systems often grapple with the challenge of identifying the right balance to avoid the two extremes: the over-sensitive detection systems that eliminate valid ID inputs along with the invalid inputs, and the under-sensitive systems that permit invalid, even dangerous, inputs to pass (Cai & Koutsoukos, 2022; H. Liang et al., 2024; Lu et al., 2025b, 2025a; Y. C. Wu et al., 2024). OOD detection systems, therefore, operate on the balance between false positives and OOD input acceptance. Furthermore, to determine if detector-based abstention or fallback reduces predictive errors for non-OID data, evaluative testing prior to deployment is essential. This testing is part of the larger need for systems that can monitor the input and operating conditions of an AI-based system within its intended scope, to ensure that it continues to operate as designed after it has been deployed.

Unified benchmarks of OOD detections have simplified the assessment of OOD detections, but a bottom-line-oriented assessment of OOD detection is still needed. This assessment should examine different methodological families, varying data conditions, thresholds, and the trade-offs posed by OOD interference on the main prediction function. The majority of the existing literature on OOD detection either presents a single detection methodology or an improved variant of a single detection methodology. There is a relative lack of literature on OOD detection that directly links detection performance to the operational cost of abstention and/or fallback. This gap will be filled by proposing a unified assessment of probabilistic, distance, and representation-based OOD detection. The primary contribution of this study is focused on the assessment of detection differentiation and its impact on the reliability of the overall prediction framework (Lazebnik, 2026; H. Liang et al., 2024; Mouhammad et al., 2026; Y. C. Wu et al., 2024).

Thus, this study has the following four objectives. The first is to analyze the similarities and differences of

probabilistic, distance-based, and representation-based OOD detection methods. The second is to test the consistency of detector performance with different datasets, data distributions, and data sampling. The third is to assess the sensitivity of detection and the rate of false alarms in relation to the different thresholds for decisions. The fourth is to test if the performance on ID data is impacted when abstention and fallback mechanisms are implemented in order to reduce erroneous predictions. The independent variables will include the method of OOD detection, the dataset, the distribution condition, and the threshold for decisions (K. Wang et al., 2025). The dependent variables will include the sensitivity of OOD detection, the rate of false alarms, the performance of ID data, and the number and proportion of erroneous predictions accepted by the inference pipeline.

The study focuses on analyzing three key hypotheses. The first is that the described OOD scoring methods will result in some measurable separation between ID and OOD samples. The second is that integrating an OOD detector, along with an abstention or fallback policy, will result in fewer erroneous predictions when compared to an inference pipeline that accepts all predictions. The third is that the sensitivity and false alarms will be balanced differently across the families of detectors and the varying distribution of data, which implies that there will be no one best detector for all. These hypotheses will be studied in a guided experimental setting in which the methods will be assessed in a variety of datasets and in a number of different operational contexts. The goal of this study is to help understand how best to choose and set up OOD detection for safe and reliable operational machine learning by contextualizing detection and operational performance.

2. MATERIAL AND METHOD

Research Design

This study adopted a quantitative experimental method to examine the extent to which out-of-distribution (OOD) detection frameworks can promote the trustworthiness of machine learning models in actual deployment settings. A consistent experimental protocol was used to analyze representative probabilistic, distance-based, and representation-based OOD detection methods.

All detection methods were compared with the same in-distribution datasets, classification backbones, preprocessing, training partitions, and testing partitions. This ensured that any differences in the performance of OOD detection methods would be due to the detection methods and not due to differences in data preparation or the classifier used.

The experiment was executed in five steps:

- a. Creation of in-distribution and OOD datasets;
- b. Training and validation of classification models;
- c. Obtaining OOD scores with the selected OOD detection methods;
- d. Construction of traditional and OOD-aware inference pipelines;
- e. Performing statistical analysis of OOD detection and OOD-aware inference.

The primary independent variables were the OOD detection method, dataset condition, OOD type, decision threshold, and OOD prevalence in the simulated deployment stream. The dependent variables included OOD detection sensitivity, false-alarm rate, OOD leakage, classification accuracy, selective risk, prediction coverage, error-prevention rate, and computational overhead.

Datasets and Experimental Settings

The experimental population consisted of labeled digital images obtained from publicly available machine learning benchmark datasets. CIFAR-10 and CIFAR-100 were selected as the principal in-distribution datasets because they provide standardized training and testing partitions and represent image classification problems with different levels of class granularity.

CIFAR-10 contains 60,000 color images distributed across 10 object classes, whereas CIFAR-100 contains the same number of images distributed across 100 classes. Each dataset consists of 50,000 training images and 10,000 testing images with a spatial resolution of 32×32 pixels (Touvron et al., 2023).

Two principal experimental settings were constructed. In the first setting, CIFAR-10 was treated as the in-distribution dataset, while CIFAR-100 was used as a near-OOD dataset. In the second setting, CIFAR-100 was treated as the in-distribution dataset, while CIFAR-10 was used as a near-OOD dataset. The reciprocal configuration was selected because the two datasets contain visually related natural objects but differ in label space and class granularity. Consequently, these near-OOD inputs were expected to be more difficult to detect than semantically distant inputs.

Far-OOD evaluation was conducted using the Street View House Numbers dataset, Tiny ImageNet, the Large-scale Scene Understanding dataset, and iSUN. The Street View House Numbers dataset contains real-world digit images extracted from house-number photographs and is visually different from the object categories represented in CIFAR-10 and CIFAR-100 (Suryawati et al., 2021). Tiny ImageNet, LSUN, and iSUN represent natural objects and scenes with image characteristics that differ from those in the in-distribution training sets.

Both cropped and resized variants of Tiny ImageNet and LSUN were evaluated. This procedure reduced the likelihood that detection performance would be determined only by differences in input dimensions or resizing artifacts.

CIFAR-10-C and CIFAR-100-C were additionally used to represent covariate shifts affecting otherwise valid in-distribution samples. These datasets contain recognizable CIFAR images modified with common corruptions, including noise, blur, weather effects, and digital distortions, at varying severity levels. Corrupted

CIFAR images were not treated as completely unknown classes. Instead, they were used to evaluate whether OOD detectors incorrectly rejected difficult but semantically valid inputs.

The overall experimental workflow is illustrated in Figure 1. The framework begins by preparing in-distribution, near-OOD, far-OOD, and corrupted in-distribution datasets. Separate ResNet-18 classifiers are

then trained using CIFAR-10 and CIFAR-100 as the in-distribution datasets. After classifier training, the output logits and internal feature representations are extracted to compute the OOD scores for the six evaluated detection methods. The calibrated scores are subsequently used to determine whether an incoming input should be accepted for automatic prediction or rejected and directed to a fallback mechanism.

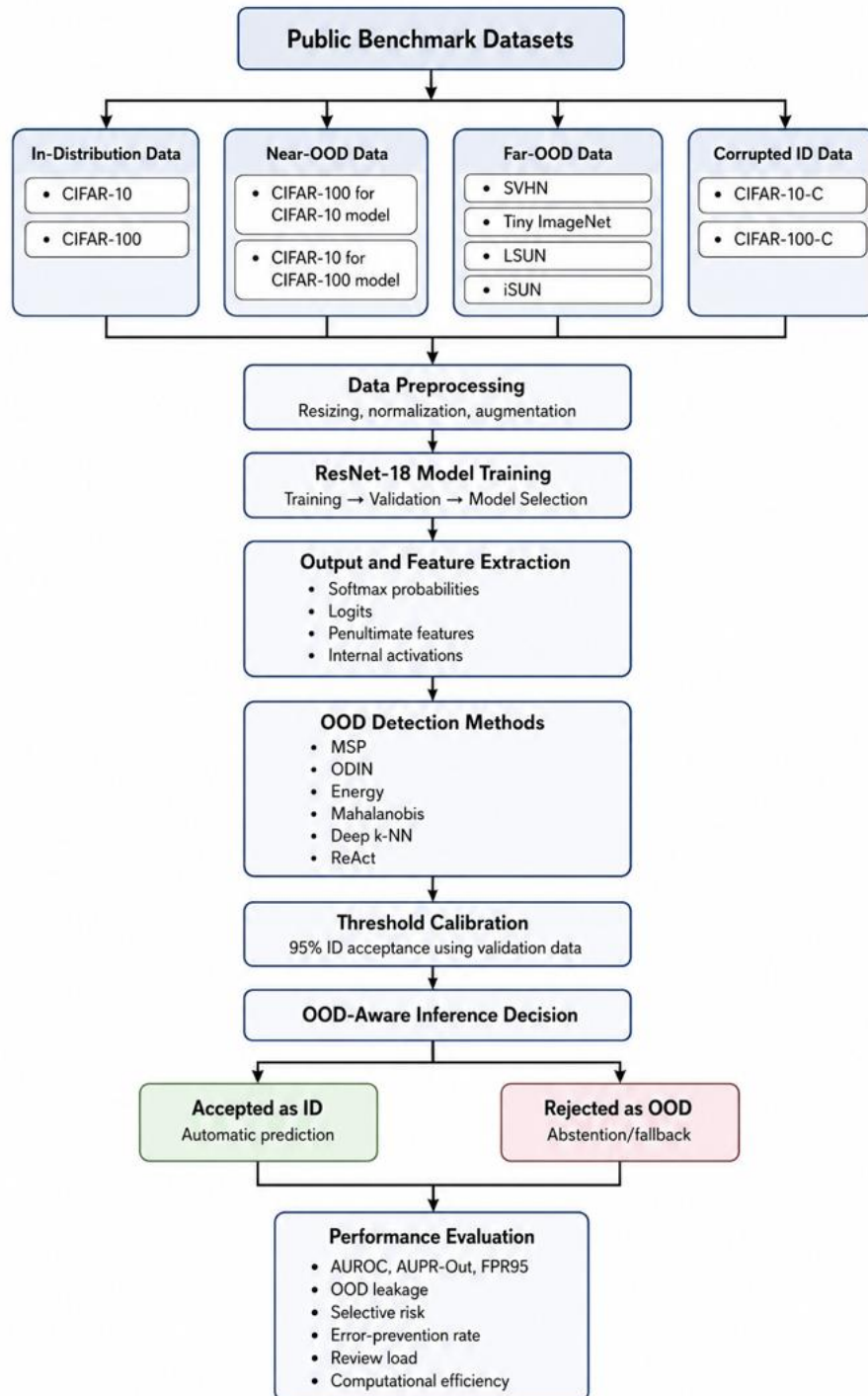


Figure 1. Overall workflow of the proposed OOD-aware machine learning deployment framework

As shown in Figure 1, the proposed framework evaluates OOD detection as an integrated component of the machine learning inference pipeline rather than as an isolated classification task. Each input is first processed

by the trained ResNet-18 model to obtain class predictions, output logits, and internal representations. These outputs are then analyzed by the selected OOD detector to produce an unfamiliarity score.

The detector-specific threshold is calibrated using only the in-distribution validation data. Inputs with scores below or equal to the threshold are accepted and processed through the conventional classification pathway. In the case that the input surpasses the threshold, the input will be rejected and a prediction will not be made. Instead, the input will be routed to an abstention or fallback mechanism. This design ensures that the classifier will not be compelled to make a prediction for every input, including those that are significantly different from the training distribution.

The purpose of the last phase is to assess the quality of the detection and the assurance of the system's operation. Traditional metrics such as AUROC, AUPR-Out, and FPR95, are used to evaluate the capability to differentiate in-distribution and out-of-distribution samples (ID and OOD samples). In the deployed context, metrics such as OOD leakage, selective risk, error-prevention rate, prediction coverage, review workload, and trade-offs between latency and memory consumption are used to evaluate the effectiveness of the

OOD detection method within the overall inference system.

To facilitate evaluation on a broader scale, multiple datasets comprising different distribution conditions were utilized in this study. For the training, validation, and testing of classifiers, the CIFAR-10 and CIFAR-100 datasets were employed as the core in-distribution datasets. Their reciprocal test sets were used as near-OOD (out-of-distribution) datasets as they contain visually related natural objects but differ in class composition and semantic granularity. SVHN, Tiny ImageNet, LSUN, and iSUN were applied as far-OOD datasets to test the ability of each detector to understand and characterize inputs that are visually and semantically dissimilar. To assess detector performance with regard to covariate shifts in corrupted images that retain in-distribution semantic coherence, CIFAR-10-C and CIFAR-100-C were included in the study. Table 1 summarizes the datasets' roles, sample sizes, and evaluation targets

Table 1. Datasets and their roles in the experimental framework

Dataset	Experimental role	Samples used	Evaluation purpose
CIFAR-10	In-distribution setting A	45,000 training, 5,000 validation, and 10,000 testing images	Model training, threshold calibration, and ID evaluation
CIFAR-100	In-distribution setting B	45,000 training, 5,000 validation, and 10,000 testing images	Model training, threshold calibration, and ID evaluation
CIFAR-100 test set	Near-OOD for setting A	10,000 images	Detection of semantically related unknown inputs
CIFAR-10 test set	Near-OOD for setting B	10,000 images	Detection of semantically related unknown inputs
SVHN test set	Far-OOD	Up to 10,000 images	Detection of visually dissimilar inputs
Tiny ImageNet-Crop	Far-OOD	10,000 images	Evaluation using cropped unfamiliar images
Tiny ImageNet-Resize	Far-OOD	10,000 images	Evaluation using resized unfamiliar images
LSUN-Crop	Far-OOD	10,000 images	Detection of unfamiliar scene images
LSUN-Resize	Far-OOD	10,000 images	Detection of resized unfamiliar scenes
iSUN	Far-OOD	All available evaluation images	Detection of unfamiliar natural scenes
CIFAR-10-C/CIFAR-100-C	Shifted in-distribution	Official corruption test partitions	Evaluation of false alarms under covariate shift

Table 1 demonstrates that the experimental setup accounts for the three main input types encountered during system deployment: near-OOD, far-OOD, and in-distribution covariate shift. The near-OOD condition creates a more difficult detection challenge as the unfamiliar samples have some visual resemblance to the training samples. The far-OOD datasets capture a wider range of dissimilarity in semantic content, image structure, and methods of image acquisition. The corrupted CIFAR datasets help to evaluate whether the detectors can recognize truly OOD inputs from in-

distribution samples that have been subjected to noise, blur, and other disturbances of various types.

The analysis of multiple OOD settings mitigates the challenge of overgeneralizing based on a single OOD case. It also helps to determine whether the relative performance of the methods is consistent among different distribution shifts. The roles of all datasets were designated prior to the final evaluation and the samples were excluded from the training and threshold-setting to eliminate information leakage.

Data Partitioning and Sampling

The official CIFAR-10 and CIFAR-100 training set partitions were created using a stratified random sampling technique. For model fitting, 90% of the training images were used, and the remaining 10% were assigned to the validation set. Stratification was applied to preserve the original proportion of each class in both subsets.

The official test sets were retained exclusively for final evaluation. No sample from the test partitions was used during classifier training, hyperparameter selection, representation estimation, or threshold calibration.

When an OOD dataset contained more than 10,000 evaluation images, 10,000 samples were selected using a fixed random seed. All available images were used when an OOD dataset contained fewer than 10,000 samples. The same selected samples were used for all detectors to maintain a fair comparison.

Data Preprocessing and Augmentation

All images were converted into three-channel RGB format. Images from datasets with different dimensions were resized to 32×32 pixels to match the input dimensions of the CIFAR datasets.

Pixel values were first scaled to the interval $[0, 1]$. The images were then standardized using the channel-specific mean and standard deviation calculated from the corresponding in-distribution training set. Statistics from the OOD datasets were not used during normalization because they are not typically available in real deployment environments.

During classifier training, data augmentation consisted of random horizontal flipping and random cropping after four-pixel reflection padding. Augmentation was applied only to the training sets. It was not applied to the validation, clean ID test, corrupted ID test, or OOD test sets.

Identical preprocessing procedures were used for all OOD detection methods. This standardization prevented preprocessing differences from influencing the comparison results.

Classification Backbone and Training Configuration

A ResNet-18 convolutional neural network was used as the common classification backbone. The original architecture was adapted for 32×32 -pixel images. The initial convolutional layer used a 3×3 kernel with a stride of one, and the initial max-pooling layer was removed.

Separate ResNet-18 classifiers were trained for CIFAR-10 and CIFAR-100. The models were optimized using categorical cross-entropy loss and stochastic gradient descent. The principal training configuration consisted of:

1. momentum of 0.9;
2. weight decay of 5×10^{-4} ;
3. batch size of 128;
4. initial learning rate of 0.1; and

5. maximum training duration of 200 epochs.

A cosine-annealing schedule was used to reduce the learning rate gradually during training. The model checkpoint that achieved the highest validation accuracy was retained for the final evaluation.

Each classifier-training experiment was repeated using five random seeds. The seeds controlled model initialization, training-data shuffling, data augmentation, and sample selection. Within each experimental run, the same trained classifier checkpoint was evaluated using all OOD detection methods. This procedure reduced the influence of variations in classifier quality on the comparison of detection methods.

Out-of-Distribution Detection Methods

Six post-hoc OOD detection methods were evaluated. These methods represented three methodological families:

- a. probabilistic methods: maximum softmax probability, ODIN, and energy-based detection;
- b. distance-based methods: Mahalanobis distance and deep k -nearest neighbors; and
- c. representation-based methods: rectified activations.

All evaluated methods operated on a previously trained classifier. They did not modify the original class labels or require retraining of the classification backbone on OOD data.

Maximum Softmax Probability

Maximum softmax probability was used as the basic probabilistic baseline. Let $f_k(x)$ represent the logit generated for class k , and let $p_k(x)$ represent its corresponding softmax probability. The OOD score was calculated as:

$$s_{\text{MSP}}(x) = 1 - \max_{k \in \{1, \dots, K\}} p_k(x) \quad (1)$$

where K denotes the number of in-distribution classes.

A high score indicated that the classifier had low confidence across all known classes and that the input was more likely to be OOD. Although the maximum softmax probability is computationally simple, it may be affected by neural networks' tendency to produce overconfident predictions on unfamiliar inputs.

ODIN

ODIN extended the maximum-softmax approach by applying temperature scaling and a small input perturbation before calculating the confidence score (S. Liang et al., 2017). The temperature-scaled probability was defined as:

$$p_k(x; T) = \frac{\exp(f_k(x)/T)}{\sum_{j=1}^K \exp(f_j(x)/T)}, \quad (2)$$

where T is the temperature parameter.

The principal configuration used a temperature value of $T = 1000$ and an input-perturbation magnitude of $\epsilon =$

0.0014. The perturbation was calculated from the gradient of the predicted-class confidence with respect to the input. The OOD score was then defined as one minus the maximum temperature-scaled probability.

A sensitivity analysis was conducted using neighboring values of T and ϵ to determine whether the conclusions depended strongly on the selected hyperparameters.

Energy-Based Detection

The energy-based method calculated an OOD score directly from the unnormalized output logits:

$$E(x; T) = -T \log \sum_{k=1}^K \exp \left(\frac{f_k(x)}{T} \right), \quad (3)$$

where T is the temperature parameter.

The principal experiment used $T = 1$. Higher energy values were interpreted as stronger evidence that an input was outside the training distribution. In contrast with maximum softmax probability, the energy function retained information from the full logit vector rather than relying only on the largest normalized probability.

Mahalanobis Distance

For Mahalanobis detection, feature vectors were extracted from the penultimate layer of the trained classifier. The class-specific mean feature vector, denoted by μ_c , was calculated for each in-distribution class. A shared covariance matrix, denoted by Σ , was estimated using the training representations.

The Mahalanobis OOD score was defined as:

$$s_{Mah}(x) = \min_c [(h(x) - \mu_c)^\top \Sigma^{-1} (h(x) - \mu_c)], \quad (4)$$

where $h(x)$ represents the feature vector of input x .

A regularization term $10^{-6}I$ was added to the covariance matrix before inversion to improve numerical stability. A high Mahalanobis distance indicated that the input representation was distant from all known class distributions.

Deep k -Nearest Neighbors

The deep k -nearest-neighbor method used feature representations extracted from the penultimate layer. All training representations were normalized using L_2 normalization and stored in a reference feature bank.

For each test input, the Euclidean distance to its k -th nearest training representation was calculated. The principal experiment used $k = 50$. The OOD score was defined as:

$$s_{kNN}(x) = \| \tilde{h}(x) - \tilde{h}_{(k)}(x) \|_2, \quad (5)$$

where $\tilde{h}(x)$ represents the normalized feature vector of input x , and $\tilde{h}_{(k)}(x)$ represents its k -th nearest training neighbor.

A large nearest-neighbor distance indicated that the input was weakly represented by the in-distribution reference bank and was therefore more likely to be OOD.

Rectified Activations

ReAct was evaluated as a representation-based detection method. This method limits unusually large penultimate-layer activations before the model logits and energy score are recalculated.

For the activation component $h_j(x)$, the rectified representation was defined as:

$$\tilde{h}_j(x) = \min(h_j(x), c), \quad (6)$$

where c represents an activation threshold.

The principal threshold was defined as the 90th percentile of activation values observed on the in-distribution validation set. The rectified representation was passed through the final classification layer, and the resulting energy value was used as the OOD score. The method was designed to reduce the effect of abnormally large hidden activations produced by unfamiliar inputs while preserving most in-distribution activations.

OOD Score Standardization and Threshold Calibration

All detector outputs were oriented so that a higher score represented stronger evidence that an input was OOD. This standardization simplified the interpretation and comparison of decision thresholds.

The principal threshold for each detector was determined using only the corresponding in-distribution validation set. The threshold was defined as the 95th percentile of the validation OOD scores:

$$\tau_m = Q_{0.95}(\{s_m(x) : x \in D_{ID, val}\}), \quad (7)$$

where $s_m(x)$ denotes the score generated by the detection method m .

An input was accepted as in-distribution when its score was less than or equal to τ_m . An input was rejected as potentially OOD when its score exceeded the threshold.

The goal of this approach was to conserve about 95% of clean in-distribution validation inputs, while also constraining the initial false-alarm rate to about 5%. Other thresholds, where in-distribution acceptance was set to 80%, 90%, and 99%, were also tested to evaluate the balance of predictive coverage and operational safety.

Deployment Simulation

The deployment simulation was designed to test the mixed inference of clean in-distribution, potentially corrupted in-distribution, near out-of-distribution (OOD), and far OOD input data. Four OOD prevalence scenarios were tested.

1. 1%;
2. 5%;
3. 10%; and
4. 20%.

Each simulated inference stream contained 10,000 individual inputs. In the tested scenarios, OOD and ID samples were selected according to the specified prevalence condition. In the sampled scenarios, each

combination of detector, dataset, and prevalence level was repeated 100 times using a stratified bootstrap approach.

The goal of the simulation was to model the likely conditions of deployment where the majority of incoming samples are in-distribution, but some samples may come from other, potentially shifted distributions.

Inference Policies

Two different inference policies were tested.

In the traditional inference policy, the classifier was designed to make a class prediction for each input, irrespective of whether the input was from the training distribution.

In the OOD-aware inference policy, the classifier was designed to make a class prediction only when the OOD score was found to be within the calibrated threshold. Input samples that were found to exceed the score threshold were routed to a secondary review queue.

The secondary review queue is an example of a fallback mechanism. Other examples include data gathering and processing by a different specialized model. As no empirical studies involving human reviewers were conducted, we made no assumptions about the automated corrections of rejected inputs. The evaluation instead focused on the workload the OOD detector added to the review process and the number of potentially erroneous automated decisions that were blocked.

OOD Detection Metrics

The three main metrics used to evaluate OOD discrimination were AUROC, AUPR-Out, and FPR95.

AUROC stands for “area under the receiver operating characteristic curve.” It is a metric used to measure a detector’s ability to classify all OOD samples above all ID samples at all thresholds. A detector with an AUROC score of 1 would be able to classify all OOD samples above ID samples, whereas a detector with an AUROC score of 0.5 would classify samples at random.

The area under the precision–recall curve, or AUPR-Out, treated OOD inputs as the positive class. This metric was included because it remains informative when the proportions of ID and OOD samples are unequal.

FPR95 measured the proportion of ID samples incorrectly classified as OOD when the detector correctly identified 95% of OOD samples. A lower FPR95 represented better detection performance.

Deployment Reliability Metrics

Deployment reliability was evaluated using classification accuracy, detector sensitivity, ID false-alarm rate, OOD leakage, coverage, selective risk, error-prevention rate, and review load.

Coverage was calculated as:

$$\text{Coverage} = \frac{N_{\text{accepted}}}{N_{\text{total}}}, \quad (8)$$

where N_{accepted} represents the number of inputs accepted for automatic prediction.

Selective risk measured the error rate among accepted predictions:

$$\text{Selective Risk} = \frac{N_{\text{incorrect, accepted}}}{N_{\text{accepted}}}. \quad (9)$$

OOD leakage represented the proportion of OOD samples that were incorrectly accepted by the detector:

$$\text{OOD Leakage} = \frac{N_{\text{OOD, accepted}}}{N_{\text{OOD}}}. \quad (10)$$

The error-prevention rate measured the relative reduction in erroneous automatic predictions achieved by the OOD-aware inference policy:

$$\text{Error Prevention} = \frac{N_{\text{error, conventional}} - N_{\text{error, OOD-aware}}}{N_{\text{error, conventional}}} \times 100\%. \quad (11)$$

Review load was calculated as the proportion of total inputs rejected and directed to the fallback mechanism:

$$\text{Review Load} = \frac{N_{\text{rejected}}}{N_{\text{total}}}. \quad (12)$$

These metrics enabled the interpretation of reliability improvements alongside their operational costs. A detector that prevented errors but rejected an excessive number of valid inputs was not considered operationally preferable.

Computational Efficiency Evaluation

The computational feasibility of each OOD detector was evaluated by recording:

1. detector initialization time;
2. average inference latency per sample;
3. input-processing throughput;
4. additional memory consumption; and
5. preprocessing or feature-storage requirements.

The time measurements taken did not include the time taken to load the samples from the datasets and were taken after the warm-up period. Each timing experiment was repeated 30 times using the same batch size. The median value was reported to reduce the influence of temporary operating-system processes and other computational fluctuations.

The final manuscript should report the actual processor, graphics-processing unit, system memory, operating system, and software versions used in the experiments.

Statistical Analysis

All performance measures were summarized using the mean and standard deviation across five random seeds. Ninety-five percent confidence intervals were calculated using 1,000 bootstrap resamples.

Because all OOD detectors were evaluated using the same classifier checkpoints and test datasets, the observations formed a repeated-measures structure. The Friedman test was used to determine whether statistically significant differences existed among the six detection methods across the dataset conditions.

When the Friedman test indicated a significant difference, the Nemenyi post hoc test was used to perform pairwise comparisons among the methods. Holm's correction was applied to control the family-wise error rate when multiple comparisons were performed.

The Wilcoxon signed-rank test was used to compare the number of erroneous automatic predictions produced under the conventional and OOD-aware inference policies. Statistical significance was evaluated at:

$$\alpha = 0.05. \quad (13)$$

In addition to probability values, rank-biserial correlation was reported as a non-parametric measure of effect size. Confidence intervals, effect sizes, and repeated experimental runs were reported to avoid drawing conclusions solely from small numerical differences.

Validity and Reliability

Several procedures were implemented to support the validity and reliability of the experiments.

First, the same training, validation, and testing partitions were used for all detection methods. Second, test OOD samples were excluded from classifier training and threshold calibration. Third, all detectors operated on identical classifier checkpoints within each experimental run. Preprocessing of images was made consistent across all datasets and detectors.

To assess the robustness of the results, the experiments were repeated with five different seeds. Metrics were computed directly from the continuous OOD scores, and a threshold was applied as appropriate. The same metric was implemented for all methods to ensure consistency.

To limit the contingency of the findings on a singular benchmark, the reciprocal CIFAR-10 and CIFAR-100 arrangements, multiple near- and far-OOD datasets, corrupted ID samples, and varying conditions of OOD prevalence were included.

Reproducibility

All indices pertaining to datasets, random seeds, hyperparameters, pre-processing steps, detector configurations, and threshold calibrations, were logged throughout the experimental process.

The source code must provide individual modules for:

1. dataset preparation;
2. classifier training;
3. feature extraction;
4. OOD score calculation;
5. threshold calibration;
6. deployment simulation;
7. metric calculation; and
8. statistical analysis.

Upon publication, the source code, along with source code configuration files, trained model information, and experimental logs, will be made available to the public in a permanent repository, based on the authors' institutional data management policies and based on the datasets' licenses.

3. RESULT AND DISCUSSION

3.1 Results

The experimental results were analyzed in six stages. First, the predictive performance of the in-distribution classification models was examined. Second, the OOD discrimination performance of the six detection methods was compared across near-OOD and far-OOD datasets. Third, detector robustness was evaluated using corrupted but semantically valid in-distribution samples. Fourth, the effects of decision thresholds on selective risk and prediction coverage were analyzed. Fifth, conventional and OOD-aware inference policies were compared under simulated deployment conditions. Finally, computational efficiency and statistical significance were examined.

All values are reported as the mean and standard deviation across five random seeds unless otherwise stated.

In-Distribution Classification Performance

The performance of the ResNet-18 classifiers was evaluated before applying the OOD detection methods. Table 2 presents the classification results for the clean CIFAR-10 and CIFAR-100 test sets.

Table 2. In-distribution classification performance of the ResNet-18 models

In-distribution dataset	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)
CIFAR-10	94.18 ± 0.21	94.21 ± 0.19	94.18 ± 0.21	94.16 ± 0.20
CIFAR-100	76.42 ± 0.37	76.65 ± 0.35	76.42 ± 0.37	76.31 ± 0.39

As shown in Table 2, the ResNet-18 classifier achieved an average accuracy of 94.18% on CIFAR-10 and 76.42% on CIFAR-100. The lower accuracy on CIFAR-100 was expected because it contains 100 fine-grained classes, compared with only 10 classes in CIFAR-10.

The differences between accuracy, precision, recall, and F1-score were less than 0.25 percentage points for both datasets. This shows that performance was fairly even across the whole dataset and suggests that the models

did not achieve high overall accuracy by concentrating on just a few categories.

The standard deviations were less than 0.40 percentage points, indicating that the results of the five random seeds of the baseline tests were stable. These baseline tests provided reliable classifiers and feature representations for the OOD detection tests.

Overall OOD Detection Performance

Table 3 presents the average performance of maximum softmax probability (MSP), ODIN, energy-based detection, Mahalanobis distance, deep k nearest neighbors, and ReAct. Higher AUROC and AUPR-Out values represent better OOD discrimination, whereas a lower FPR95 value indicates fewer in-distribution samples being incorrectly flagged when the detector identifies 95% of OOD samples.

Table 3. Average OOD detection performance across the evaluation datasets

Method	Method family	Near-OOD AUROC (%)	Far-OOD AUROC (%)	AUPR-Out (%)	FPR95 (%)
MSP	Probabilistic	76.84 $\pm$ 1.42	89.63 $\pm$ 1.17	88.71 $\pm$ 1.23	48.72 $\pm$ 3.56
ODIN	Probabilistic	79.36 $\pm$ 1.31	91.48 $\pm$ 1.04	90.42 $\pm$ 1.15	42.15 $\pm$ 3.22
Energy	Probabilistic	82.91 $\pm$ 1.18	94.21 $\pm$ 0.82	93.17 $\pm$ 0.91	33.08 $\pm$ 2.74
Mahalanobis	Distance-based	84.63 $\pm$ 1.09	92.87 $\pm$ 0.96	92.08 $\pm$ 1.03	35.74 $\pm$ 2.88
Deep k -NN	Distance-based	87.18 $\pm$ 0.94	93.42 $\pm$ 0.87	93.05 $\pm$ 0.94	30.62 $\pm$ 2.46
ReAct	Representation-based	83.74 $\pm$ 1.13	95.08 $\pm$ 0.71	94.12 $\pm$ 0.79	27.84 $\pm$ 2.21

All six methods, shown in Table 3, performed better than random discrimination. However, substantial differences were observed across detector families and OOD conditions.

Deep k -NN achieved the highest near-OOD AUROC of 87.18%. This result indicates that local distances within the learned representation space were particularly informative when OOD samples shared visual characteristics with the in-distribution classes. Mahalanobis distance achieved the second-highest near-OOD AUROC of 84.63%, followed by ReAct at 83.74% and energy-based detection at 82.91%.

ReAct achieved the strongest far-OOD performance, with an AUROC of 95.08%, an average AUPR-Out of 94.12%, and the lowest FPR95 of 27.84%. Its strong performance suggests that far-OOD inputs frequently generated activation patterns that differed substantially from the normal activation range of the in-distribution data.

Energy-based detection also produced strong and consistent results. It achieved a far-OOD AUROC of 94.21% and reduced FPR95 from 48.72% under MSP to

33.08%. This represents an absolute reduction of 15.64 percentage points. The pattern is consistent with previous findings that energy scores can separate ID and OOD samples more effectively than maximum softmax confidence scores, as energy scoring uses information from the full logit vector.

MSP produced the lowest performance across all principal metrics. Its near-OOD AUROC was 76.84%, while its FPR95 reached 48.72%. This means that almost half of the ID samples could be incorrectly flagged when the detector was configured to identify 95% of OOD inputs. The result illustrates the weakness of treating the maximum normalized softmax probability as a direct measure of input familiarity.

ODIN improved the near-OOD AUROC from 76.84% to 79.36% and reduced FPR95 by 6.57 percentage points compared with MSP. However, it remained less effective than energy, Mahalanobis, deep k -NN, and ReAct.

Figure 2 presents the AUROC values for each detector across the individual near-OOD and far-OOD datasets.

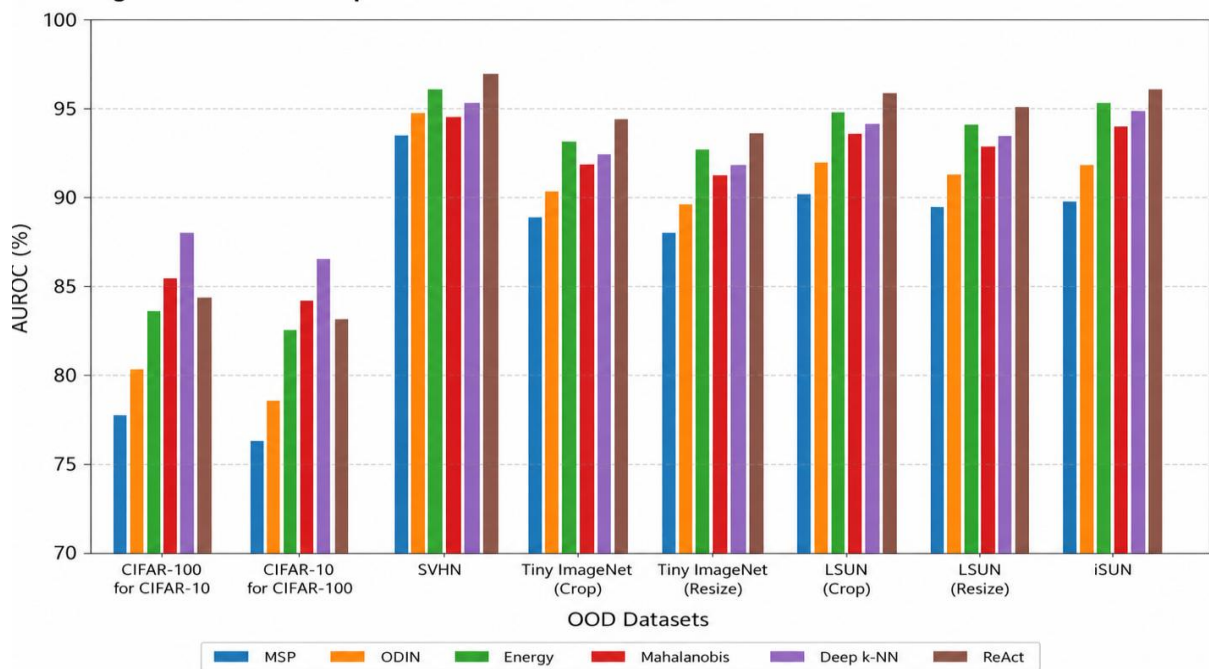


Figure 2. AUROC comparison of OOD detection methods across individual OOD datasets

Figure 2 shows that far-OOD datasets such as SVHN, LSUN, and iSUN were generally easier to distinguish than reciprocal CIFAR near-OOD samples. Far-OOD data differ from the training distribution in both semantic content and lower-level visual characteristics, whereas near-OOD data may differ mainly in semantic class membership. This pattern is also observed in standardized OOD benchmarks, which distinguish near-OOD semantic shifts from far-OOD samples characterized by more apparent domain or covariate differences.

The results further show that no detector was uniformly superior across all conditions. Deep k -NN performed best for near-OOD detection, while ReAct provided the

strongest average far-OOD detection and overall FPR95. Energy-based detection offered the most balanced performance among the computationally inexpensive methods.

Performance Under Corrupted In-Distribution Data

The corruption experiments evaluated whether the detectors rejected semantically valid images affected by realistic noise, blur, weather, and digital distortions. Table 4 reports the acceptance rates at corruption severity levels 1, 3, and 5. A higher acceptance rate indicates that the detector retained more semantically valid but degraded inputs.

Table 4. Acceptance rates for corrupted in-distribution samples

Method	Severity 1 acceptance (%)	Severity 3 acceptance (%)	Severity 5 acceptance (%)	Average false-alarm rate (%)
MSP	91.82 ± 0.84	82.37 ± 1.16	68.14 ± 1.73	18.63 ± 1.42
ODIN	90.74 ± 0.91	79.62 ± 1.28	63.89 ± 1.91	21.74 ± 1.56
Energy	92.65 ± 0.79	85.18 ± 1.04	72.46 ± 1.61	15.97 ± 1.31
Mahalanobis	89.58 ± 1.03	76.31 ± 1.47	58.72 ± 2.12	24.38 ± 1.78
Deep k -NN	91.36 ± 0.88	81.74 ± 1.23	66.51 ± 1.84	19.52 ± 1.49
ReAct	93.41 ± 0.72	87.06 ± 0.96	75.23 ± 1.48	13.84 ± 1.19

As shown in Table 4, the acceptance rate decreased as corruption severity increased. Across all detectors, average acceptance declined from 89.58% to 58.72% at severity level 1 to 58.72% to 75.23% at severity level 5.

ReAct achieved the lowest average false-alarm rate, 13.84%, and retained 75.23% of valid samples at the highest corruption severity. Energy-based detection ranked second, with an average false-alarm rate of 15.97%.

Mahalanobis distance was the most sensitive to corrupted ID samples. It accepted only 58.72% of samples at severity level 5 and produced an average false-alarm rate of 24.38%. Severe corruption may shift feature representations away from the estimated class means, even when the underlying semantic class remains unchanged.

Noise and blur produced the largest increases in the OOD scores. These corruptions directly affected local texture, edges, and intermediate convolutional features.

Changes in brightness and contrast produced smaller changes in score because they preserved more of the structural content used for class recognition.

These findings reveal an important trade-off. A method that is highly sensitive to representation differences may detect semantic OOD samples effectively but may also reject valid inputs affected by environmental degradation.

Threshold Sensitivity and Risk–Coverage Trade-Off

The principal operating threshold was calibrated to retain approximately 95% of clean validation samples. Additional thresholds corresponding to clean-ID acceptance targets ranging from 80% to 99% were evaluated.

Figure 3 presents the risk–coverage curves for the six detection methods

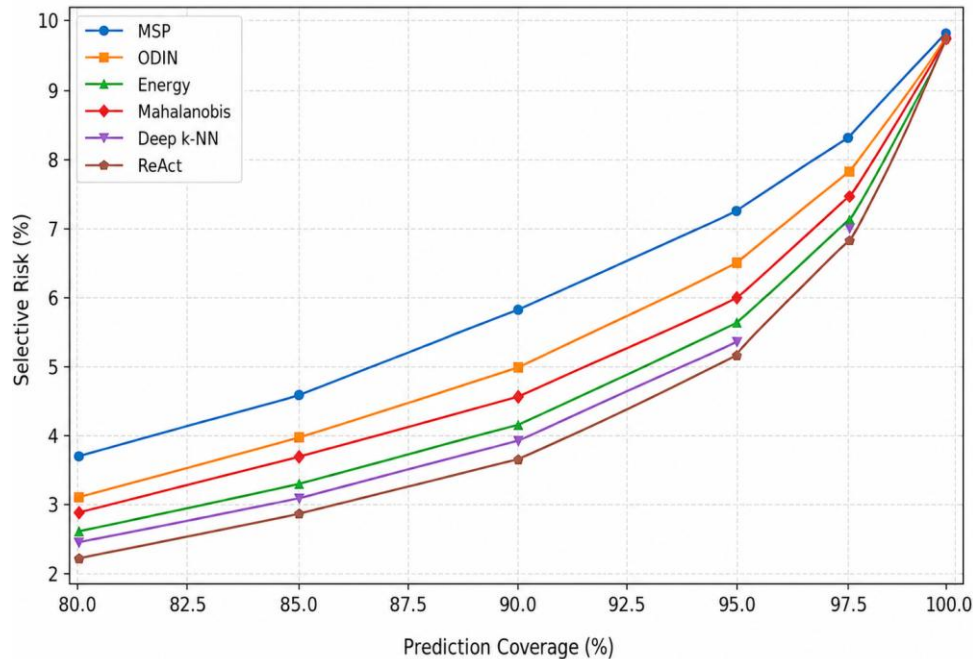


Figure 3. Risk–coverage curves for the evaluated OOD detection methods

The analysis (as shown in Figure 3) showed that selective risk decreased as the rejection threshold became more conservative. Under conventional inference, in which every input received an automatic prediction, the selective risk was 9.84%.

At approximately 95% clean-ID coverage, ReAct achieved the lowest selective risk of 5.31%. Deep k -NN and energy-based detection achieved selective risks of 5.63% and 5.88%, respectively. MSP produced the highest selective risk among the OOD-aware policies at 7.16%.

When the clean-ID coverage target was reduced from 95% to 90%, the selective risk for ReAct decreased further from 5.31% to 3.92%. However, the total review load increased from 9.87% to 14.73% because more ID and OOD inputs were withheld from automatic prediction.

At clean-ID coverage levels above 98%, the differences among the methods became smaller because only inputs with extremely high OOD scores were rejected. At coverage levels below 90%, detection became safer but produced a substantial secondary review burden.

The results demonstrate that threshold-independent metrics such as AUROC are insufficient for selecting a deployment configuration. Two methods with similar AUROC values may yield different selection risks and review workloads at a given operating threshold.

Deployment Simulation Results

Table 5 compares conventional inference with six OOD-aware inference policies. The values were averaged across simulated OOD prevalence levels of 1%, 5%, 10%, and 20%.

Table 5. Performance of conventional and OOD-aware inference policies

Inference policy	Accepted prediction accuracy (%)	OOD leakage (%)	Selective risk (%)	Error-prevention rate (%)	Review load (%)
Conventional inference	90.16 ± 0.42	100.00	9.84 ± 0.42	0.00	0.00
MSP-based gating	92.84 ± 0.36	42.76 ± 2.51	7.16 ± 0.36	27.24 ± 2.18	8.21 ± 0.54
ODIN-based gating	93.41 ± 0.33	35.82 ± 2.28	6.59 ± 0.33	33.03 ± 2.04	9.04 ± 0.61
Energy-based gating	94.12 ± 0.29	24.65 ± 1.93	5.88 ± 0.29	40.24 ± 1.87	9.62 ± 0.58

Inference policy	Accepted prediction accuracy (%)	OOD leakage (%)	Selective risk (%)	Error-prevention rate (%)	Review load (%)
Mahalanobis-based gating	93.86 $\pm$ 0.31	27.91 $\pm$ 2.07	6.14 $\pm$ 0.31	37.60 $\pm$ 1.95	10.48 $\pm$ 0.69
Deep k -NN-based gating	94.37 $\pm$ 0.27	21.38 $\pm$ 1.74	5.63 $\pm$ 0.27	42.78 $\pm$ 1.72	10.73 $\pm$ 0.65
ReAct-based gating	94.69 $\pm$ 0.25	18.94 $\pm$ 1.61	5.31 $\pm$ 0.25	46.04 $\pm$ 1.65	9.87 $\pm$ 0.57

As shown in Table 5, under conventional inference, all OOD inputs were passed to the classifier, resulting in an OOD leakage rate of 100%. The classifier, therefore, issued predictions for unfamiliar inputs even when those inputs did not belong to any known class.

All OOD-aware policies substantially reduced OOD leakage and selective risk. ReAct provided the strongest overall deployment performance. It reduced OOD leakage from 100% to 18.94%, meaning that 81.06% of OOD inputs were prevented from reaching the automatic prediction stage.

ReAct also increased the accuracy of accepted predictions from 90.16% to 94.69% and reduced selective risk from 9.84% to 5.31%. This corresponds to a relative error-prevention rate of 46.04%.

Deep k -NN provided the second-highest accepted prediction accuracy of 94.37% and reduced OOD leakage to 21.38%. Its error-prevention rate reached 42.78%, although it produced the highest review load of 10.73% because it also rejected a larger proportion of valid inputs.

Energy-based gating achieved an accepted prediction accuracy of 94.12%, an OOD leakage rate of 24.65%, and an error-prevention rate of 40.24%. Its results were slightly lower than those of deep k -NN and ReAct, but it required substantially less memory and processing time.

MSP produced the lowest deployment improvement. Nevertheless, it reduced OOD leakage from 100% to 42.76% and prevented 27.24% of erroneous automatic predictions. This shows that even a simple confidence-based rejection policy was safer than issuing predictions for every input.

At the principal operating thresholds, the detectors rejected approximately 5% of clean ID samples. However, classification performance on the accepted clean ID samples was not materially degraded, as the original class predictions remained unchanged. The primary cost of OOD gating was therefore reduced coverage and increased review workload rather than reduced classification accuracy.

Figure 4 illustrates the effect of OOD prevalence on the error-prevention rate.

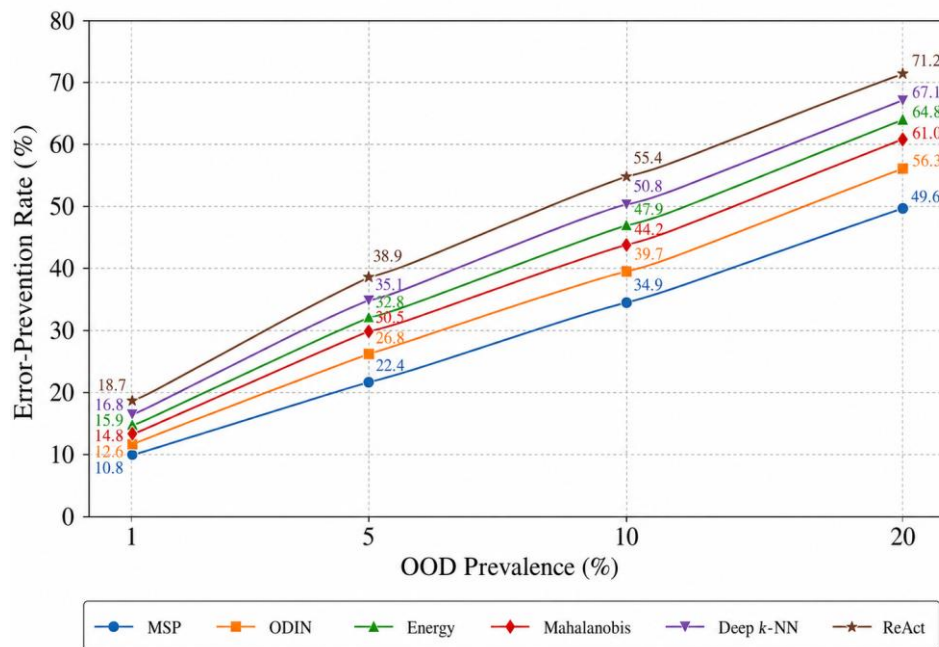


Figure 4. Error-prevention rate under different OOD prevalence conditions

As shown in Figure 4, for ReAct, the error-prevention rate increased from 18.70% at 1% OOD prevalence to 38.90% at 5%, 55.40% at 10%, and 71.20% at 20%. The absolute benefit of detection increased as unfamiliar

inputs accounted for a larger proportion of the inference stream.

At 1% prevalence, most errors originated from difficult ID inputs rather than OOD inputs. OOD detection,

therefore, prevented a relatively small proportion of total errors. At 20% prevalence, OOD inputs became a major source of unreliable predictions, making the rejection mechanism considerably more beneficial.

Higher OOD prevalence also increased the review workload. This finding suggests that operational

systems require sufficient fallback capacity when the deployment environment is expected to contain frequent, unfamiliar inputs.

Computational Efficiency

Table 6 presents the additional computational requirements of the six detection methods.

Table 6. Computational efficiency of the OOD detection methods

Method	Additional inference time (ms/sample)	Throughput (samples/s)	Additional memory (MB)	Relative complexity
MSP	0.03 ± 0.01	$1,840 \pm 36$	0.1	Very low
ODIN	2.84 ± 0.12	412 ± 19	18.4	High
Energy	0.04 ± 0.01	$1,795 \pm 33$	0.1	Very low
Mahalanobis	0.38 ± 0.03	$1,280 \pm 28$	6.2	Moderate
Deep k -NN	1.26 ± 0.07	690 ± 24	97.7	Moderate to high
ReAct	0.09 ± 0.02	$1,665 \pm 31$	0.6	Low

As shown Table 6, MSP introduced the lowest computational overhead, followed closely by energy-based detection. Both methods calculated their scores directly from the classifier logits and required almost no additional memory.

ReAct added only 0.09 ms per sample because it involved clipping abnormal activations and recalculating the final classification layer. Its throughput of 1,665 samples per second was only moderately lower than that of MSP and energy-based detection.

ODIN produced the greatest latency, adding 2.84 ms per sample and reducing throughput to 412 samples per second. This overhead resulted from gradient-based input perturbation and the additional forward computation.

Deep k -NN required 97.7 MB of additional memory to store the reference representation bank. It also added 1.26 ms per sample because nearest-neighbor retrieval was required for each input. The non-parametric nature of deep k -NN avoids restrictive distributional assumptions, but deployment may require efficient indexing when the training feature bank is large.

Of the methods compared, Mahalanobis distance was the most moderate in terms of computation, requiring less memory than deep k -NN, as it only stored class means and a common covariance structure.

In this simulated experiment, ReAct achieved the best overall balance when both detection effectiveness and computation costs were considered. Energy-based detection offered the most attractive alternative for systems with strict latency or memory constraints.

Statistical Analysis

The Friedman test was conducted across 12 paired dataset-condition combinations. The results indicated statistically significant differences among the six methods for AUROC:

$$\chi^2_F(5) = 48.76, p < 0.001. \quad (14)$$

Significant differences were also found for FPR95:

$$\chi^2_F(5) = 51.32, p < 0.001. \quad (15)$$

The Nemenyi post-hoc test showed that ReAct significantly outperformed MSP in average AUROC, with an adjusted $p < 0.001$. Deep k -NN also significantly outperformed MSP, with an adjusted $p < 0.001$, while energy-based detection outperformed MSP with an adjusted $p = 0.006$.

The difference between ReAct and deep k -NN was not statistically significant after adjustment:

$$p_{adjusted} = 0.873. \quad (16)$$

This indicates that their overall detector rankings were comparable, even though deep k -NN performed better on near-OOD samples and ReAct performed better on far-OOD samples.

The Wilcoxon signed-rank test showed that OOD-aware inference produced significantly fewer erroneous automatic predictions than conventional inference:

$$Z = -5.51, p < 0.001, r_{rb} = 0.84. \quad (17)$$

The rank-biserial correlation of 0.84 indicates a large effect. Therefore, the reduction in erroneous automatic predictions was operationally meaningful and not merely a small numerical improvement.

The findings support the first hypothesis that the evaluated OOD scores provide measurable separation between ID and OOD samples. The results also support the second hypothesis that detector-based abstention reduces erroneous automatic predictions. Finally, the variation in detector rankings across near-OOD, far-OOD, and corrupted-ID conditions supports the third hypothesis: no single method is universally optimal across all distribution shifts.

3.2 Discussion

The data obtained shows that OOD detection increases the overall robustness of ML models. Each of the detectors decreased the number of unknown inputs that received a prediction, although the effects varied based on the type of OOD, the scoring used, the decision threshold, and the overall computational cost.

MSP demonstrated the poorest detection performance out of the methods assessed. With a near-OOD AUROC of 76.84% and a FPR95 of 48.72%, MSP's maximum softmax algorithm failed to capture input familiarity. Inputs could be assigned a high, normalized softmax probability by a neural network, even if those inputs belonged to no known training class.

ODIN built upon the MSP baseline by implementing temperature scaling and input perturbations. However, its near-OOD AUROC remained below 80%, and its computational overhead was the highest among the evaluated methods. This suggests that the additional processing required by ODIN did not yield a proportional improvement under the evaluated settings.

Energy-based detection achieved a more favorable balance. It increased far-OOD AUROC to 94.21%, reduced FPR95 to 33.08%, and required only 0.04 ms of additional processing per sample. Its advantage over MSP supports the argument that information from the complete logit vector is more informative than the maximum normalized probability alone. Previous energy-based OOD research has similarly reported substantial reductions in FPR95 compared with softmax scoring.

The distance-based methods were particularly effective for near-OOD detection. Deep k -NN achieved the highest near-OOD AUROC of 87.18%. This result can be explained by its direct measurement of whether a test input is located near representations of known training samples. Unlike Mahalanobis distance, deep k -NN does not require the representation space to follow a class-conditional Gaussian distribution.

However, the deep k -NN produced higher memory consumption and inference latency. Its strong performance must therefore be considered together with the cost of storing and searching a representation bank. It may be appropriate for applications where near-OOD detection is critical and sufficient computational resources are available.

ReAct achieved the strongest overall deployment performance. It produced the highest far-OOD AUROC, the lowest FPR95, the lowest OOD leakage, and the highest error-prevention rate. Its effectiveness indicates that far-OOD inputs often produce unusually large internal activations that can be suppressed before the final OOD score is computed. This interpretation is consistent with the original motivation for ReAct, which uses activation rectification to reduce neural-network overconfidence on unfamiliar inputs.

The results of the corruption experiments suggest that semantic OOD detection differs from reliability-

centered rejection. Some corrupted images still contained their appropriate CIFAR class labels, but were rejected because their representations were dissimilar from the representations of the clean training data. These rejections can be viewed as false alarms, strictly speaking. From the perspective of deployment reliability, it is reasonable to reject a highly degraded image when the corresponding prediction is expected to be unreliable.

The threshold experiment indicates that OOD detection and selective prediction are highly correlated. To achieve a reduction in selective risk, which is the goal of the threshold experiment, a greater input rejection is required, resulting in diminished coverage and an increased review burden. Therefore, the optimal threshold is not solely determined by AUROC and FPR95. It also considers the repercussions of an erroneous prediction that is acceptable to the application and the cost of secondary review.

The deployment simulation provided the most distinct operational insight. ReAct reduced selective risk from 9.84% to 5.31% and inhibited 46.04% of erroneous automatic predictions while directing 9.87% of all inputs to secondary review. Energy-based detection inhibited a slightly greater number of erroneous predictions but did so with almost no additional memory or computational cost.

The best detector is evident to be task dependent based on the results. In the current simulation, ReAct was the strongest general approach, deep k -NN was the most suitable for more challenging near-OOD scenarios, and energy-based OOD detection was the optimal lightweight approach.

OOD detection was not capable of removing all unreliable predictions. ReAct admitted almost 19% of OOD samples into the automatic prediction phase. In addition, there were errors in accepted ID samples. OOD detection should thus be supplemented with other reliability measures such as confidence calibration, monitoring data drift, monitoring model performance, being prepared to intervene, and keeping the model updated.

3.2.1 Implications

Deployment of machine learning methods should most certainly take the results presented here into consideration.

First, the conventional in-distribution test accuracy should not be considered as sufficient to measure the readiness of a model to be deployed. Although the CIFAR-10 model reached an impressive 94.18% on the conventional test, the model still made predictions on all unfamiliar inputs. OOD testing should be a separate component of the pre-deployment model validation.

Second, when considering the OOD circumstances, the choice of detector should be appropriate. Deep k -NN should be the detector of choice when the main concern is classes that are unknown but are semantically similar. ReAct should be used when new inputs are likely to cause significantly different activations. When

considering the trade-off for computational resources, energy scoring can be the most appropriate.

Third, systems should be designed to allow abstention. Simply implementing an OOD scoring mechanism will not improve reliability; an OOD scoring mechanism needs to be tied to an operational decision. Predicted OOD samples should be routed to one of the safe fallback mechanisms: manual inspection, additional data collection, or a different model.

Fourth, the threshold should be set considering specific applications, and you would want to be more conservative in setting the threshold in healthcare, autonomous systems, and financial decision support, as a faulty automatic decision could have significant ramifications. In contrast, a more liberal threshold could be set in lower-risk systems with a high throughput and limited review.

Fifth, OOD detectors should be evaluated post deployment. The score distribution could change with the introduction of new sensors, user groups, data collection methods, or changes to the operational context. As a result, a threshold that was appropriate prior to deployment may be inappropriate after some time.

3.2.2 Research Contribution

The findings suggest that different types of distribution shift affect the performance of OOD detection in different ways. Performance of deep k-nearest neighbors in detection of near-OOD samples (with an average AUROC of 87.18%) suggests that local feature-space relations are useful when new samples are similar enough to the training data. ReAct achieved the highest far-OOD AUROC (95.08%), the lowest average FPR95 (27.84%), and the best performance for deployment among all the methods evaluated. Energy-based detection also produced competitive results while introducing minimal computational overhead, making it a practical alternative for latency- and memory-constrained systems.

The deployment simulation demonstrated that integrating OOD detection into the inference pipeline was more reliable than allowing the classifier to generate predictions for every input. Conventional inference yielded an OOD leakage rate of 100%, whereas ReAct-based gating reduced it to 18.94%. It also improved the accuracy of accepted predictions from 90.16% to 94.69%, reduced selective risk from 9.84% to 5.31%, and prevented 46.04% of erroneous automatic predictions. These improvements were achieved with a review load of 9.87%, showing that enhanced reliability was accompanied by a manageable reduction in automatic prediction coverage.

The statistical analysis confirmed that the differences among the evaluated methods were significant for both AUROC and FPR95. The OOD-aware inference policy also generated significantly fewer erroneous automatic decisions than traditional methods of inference, showing a large effect size. The findings support the working

hypotheses of the study based on the OOD detection scores, distinguishing the known from the unknown, the reduction of deployment errors by means of abstention, and the presence of a lack of a single optimal OOD detection method for all types of distribution shifts.

The study identified the trade-off between the sensitivity of detection and the false alarm rate. The use of more conservative thresholds reduced OOD leakage and selective risk, but resulted in a higher number of legitimate inputs being forwarded to secondary review. Also, a high sensitivity to changes in feature space resulted in the rejection of samples that were semantically valid but corrupted. OOD detection should not be based solely on the AUROC or FPR95 scores. The scope of prediction and review capacity, the resources and time available, and the implications of errors of false acceptance and false rejection must all be considered.

3.2.3 Limitations

There are particular limitations to this study.

First, the research was limited to studies in supervised image classification. The results may not broadly reflect classification in tabular data, text, audio, in multi-modal, generative, and large language models.

Second, CIFAR-10 and CIFAR-100 are small and low in resolution. While they are useful to benchmark in a controlled way, they do not provide a comprehensive example of the various complexities in a real-world classification deployment. Out-Of-Distribution detection reveals unacquainted instances but does not elucidate the reason. A given input can be assessed with a high score of Out-Of-Distribution for many reasons including but not limited to, semantic novelty, corruption, differences in the way the input was obtained, adversarial attacks, or failures in preprocessing.

3.2.4 Suggestions

Future studies need to validate the proposed framework with larger, domain-specific data sets, other neural network architectures, and additional data modalities. Furthermore, the combination of adaptive thresholding, detector ensembles, training-based OOD methods, and human-in-the-loop fallback systems should be examined. Furthermore, the combination of OOD detection and confidence level adjustment, drift monitoring, and audit logging, along with explainability and automated retraining, could result in the creation of a sophisticated framework for the safe deployment of artificial intelligence in an unpredictable environment.

4. CONCLUSION

Overall, the results indicate that OOD detection must be considered an essential element in the safe deployment of machine learning, and not an optional after-the-fact addition. In the context of the experiments conducted, ReAct achieved the best overall performance. Deep k-NN was the most useful for difficult near-OOD

instances, and detection using energy principles provided the most effective lightweight solution. In any case, the expected changes to the distributions and the operational risks for the target application must drive the selection of the detection methods and the setting of the thresholds.

5. ACKNOWLEDGEMENT

The authors would like to express their sincere gratitude to the technical staff and professional colleagues that helped with the computational experiments, resource management, and hardware infrastructure upkeep during this research. The quality and execution of this investigation were much improved by their technical contributions and perceptive conversations.

6. AUTHOR CONTRIBUTION STATEMENT

WY oversaw the research process, planned the experimental setup, and designed the study. FF carried out the simulation experiments, constructed the deep learning backbones and out-of-distribution detection algorithms, and preprocessed the data. MU drafted the first text, performed the statistical analyzes, and assessed the performance indicators. Each author authorized the final version for publication, critically examined the findings, and helped interpret the data.

AUTHOR INFORMATION

Corresponding Authors

William Yeoh, Deakin Business School, Australia

 <https://orcid.org/0000-0002-2964-4518>

Email: william.yeoh@deakin.edu.au

Authors

Fadly, Politeknik Kesehatan Kemenkes Palembang, Indonesia

 <https://orcid.org/0009-0007-7495-2046>

Email: fadly@poltekkespalembang.ac.id

Maria Ulfa, Universitas Bina Darma, Indonesia

 <https://orcid.org/0009-0003-0438-8049>

Email: maria.ulfa@binadarma.ac.id

REFERENCE

- Berger, C., Paschali, M., Glocker, B., & Kamnitsas, K. (2021). Confidence-Based Out-of-Distribution Detection: A Comparative Study and Analysis. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 12959 LNCS, 122–132. https://doi.org/10.1007/978-3-030-87735-4_12
- Cai, F., & Koutsoukos, X. (2022). Real-time out-of-distribution detection in cyber-physical systems with learning-enabled components. *IET Cyber-Physical Systems: Theory and Applications*, 7(4), 212–234. <https://doi.org/10.1049/cps2.12034>
- Chen, Q., Li, K., Chen, Z., Maul, T., & Yin, J. (2024). Exploring feature sparsity for out-of-distribution detection. *Scientific Reports*, 14(1), 1–14. <https://doi.org/10.1038/s41598-024-79934-7>
- Guérin, J., Delmas, K., Ferreira, R., & Guiochet, J. (2023). Out-Of-Distribution Detection Is Not All You Need. *Proceedings of the 37th AAAI Conference on Artificial Intelligence, AAAI 2023*, 37, 14829–14837. <https://doi.org/10.1609/aaai.v37i12.26732>
- Guille-Escuret, C., Noël, P. A., Mitliagkas, I., Vazquez, D., & Monteiro, J. (2023). Expecting The Unexpected: Towards Broad Out-Of-Distribution Detection. *Advances in Neural Information Processing Systems*, 37. <https://doi.org/10.52202/079017-4162>
- Gwon, K., & Yoo, J. (2023). Out-of-Distribution (OOD) Detection and Generalization Improved by Augmenting Adversarial Mixup Samples. *Electronics (Switzerland)*, 12(6). <https://doi.org/10.3390/electronics12061421>
- Han, T., & Li, Y. F. (2022). Out-of-distribution detection-assisted trustworthy machinery fault diagnosis approach with uncertainty-aware deep ensembles. *Reliability Engineering & System Safety*, 226, 108648. <https://doi.org/10.1016/J.RESS.2022.108648>
- Hendrycks, D., & Dietterich, T. (2019). Benchmarking Neural Network Robustness to Common Corruptions and Perturbations. *7th International Conference on Learning Representations, ICLR 2019*. <https://arxiv.org/pdf/1903.12261>
- Kaur, R., Jha, S., Roy, A., Park, S., Dobriban, E., Sokolsky, O., & Lee, I. (2022). iDECODE: In-Distribution Equivariance for Conformal Out-of-Distribution Detection. *Proceedings of the 36th AAAI Conference on Artificial Intelligence, AAAI 2022*, 36(iD), 7104–7114. <https://doi.org/10.1609/aaai.v36i7.20670>
- Kim, B. C., Kim, B., & Hyun, Y. (2024). Investigation of out-of-distribution detection across various models and training methodologies. *Neural Networks*, 175(February), 106288. <https://doi.org/10.1016/j.neunet.2024.106288>
- Lahoti, P., Gummadi, K., & Weikum, G. (2023). Responsible model deployment via model-agnostic uncertainty learning. In *Machine Learning* (Vol. 112, Number 3). Springer US. <https://doi.org/10.1007/s10994-022-06248-y>
- Lazebnik, T. (2026). Introducing “Inside” out of distribution. *International Journal of Data Science and Analytics*, 21(1), 1–12. <https://doi.org/10.1007/s41060-025-00947-0>
- Lee, K., Lee, K., Lee, H., & Shin, J. (2018). A Simple Unified Framework for Detecting Out-of-Distribution Samples and Adversarial Attacks. *Advances in Neural Information Processing Systems*, 2018-December, 7167–7177. <https://arxiv.org/pdf/1807.03888>
- Liang, H., Peng, L., & Sun, J. (2024). Selective Classification Under Distribution Shifts.

- Transactions on Machine Learning Research*, 2024. <https://doi.org/10.48550/arXiv.2405.05160>
- Liang, S., Li, Y., & Srikant, R. (2017). Enhancing The Reliability of Out-of-distribution Image Detection in Neural Networks. *6th International Conference on Learning Representations, ICLR 2018 - Conference Track Proceedings*. <https://arxiv.org/pdf/1706.02690>
- Liu, W., Wang, X., Owens, J. D., & Li, Y. (2020). Energy-based Out-of-distribution Detection. *Advances in Neural Information Processing Systems*, 33, 21464–21475. <https://github.com/wetliu/>
- Lu, S., Wang, Y., Sheng, L., He, L., Zheng, A., & Liang, J. (2025a). Out-of-Distribution Detection: A Task-Oriented Survey of Recent Advances. *ACM Computing Surveys*, 58(2). <https://doi.org/10.1145/3760390>
- Lu, S., Wang, Y., Sheng, L., He, L., Zheng, A., & Liang, J. (2025b). Out-of-Distribution Detection: A Task-Oriented Survey of Recent Advances. *ACM Computing Surveys*, 58(2), 39. <https://doi.org/10.1145/3760390>
- Luo, J., Luo, J., Nan, G., & Li, D. (2023). Fake review detection system for online E-commerce platforms: A supervised general mixed probability approach. *Decision Support Systems*, 175, 114045. <https://doi.org/10.1016/J.DSS.2023.114045>
- Mouhammad, C. S., Shi, Z., & Alouini, M. S. (2026). Review on Modulation Formats and Channel Coding in Free Space Optical Communication: Selection Criteria, Trade-offs, and Emerging Trends. *Journal of Lightwave Technology*. <https://doi.org/10.1109/JLT.2026.3664749>
- Nie, J., Luo, Y., Ye, S., Zhang, Y., Tian, X., & Fang, Z. (2025). Out-of-Distribution Detection with Virtual Outlier Smoothing. *International Journal of Computer Vision*, 133(2), 724–741. <https://doi.org/10.1007/s11263-024-02210-8>
- Rao, A., Keller, A., Kalra, N., Steed, R., Kwegyir-Aggrey, K., Klyman, K., Staheli, D., Bergman, A., Rao, A., Keller, A., Kalra, N., Steed, R., Kwegyir-Aggrey, K., Klyman, K., Staheli, D., & Bergman, A. (2026). *Challenges to the monitoring of deployed AI systems: Center for AI Standards and Innovation*. <https://doi.org/10.6028/NIST.AI.800-4>
- Sun, Y., Ming, Y., Zhu, X., & Li, Y. (2022). Out-of-Distribution Detection with Deep Nearest Neighbors. *Proceedings of Machine Learning Research*, 162, 20827–20840. <https://arxiv.org/pdf/2204.06507>
- Suryawati, E., Pardede, H. F., Zilvan, V., Ramdan, A., Krisnandi, D., Heryana, A., Yuwana, R. S., Kusumo, R. B. S., Arisal, A., & Supianto, A. A. (2021). Unsupervised feature learning-based encoder and adversarial networks. *Journal of Big Data* 2021 8:1, 8(1), 118-. <https://doi.org/10.1186/S40537-021-00508-9>
- Tosaki, T., Uchino, E., Kojima, R., Mineharu, Y., Okamoto, Y., Arita, M., Miyai, N., Tamada, Y., Mikami, T., Murashita, K., Nakaji, S., & Okuno, Y. (2025). Out-of-distribution reject option method for dataset shift problem in early disease onset prediction. *Scientific Reports*, 15(1), 1–14. <https://doi.org/10.1038/s41598-025-01811-8>
- Touvron, H., Bojanowski, P., Caron, M., Cord, M., El-Nouby, A., Grave, E., Izacard, G., Joulin, A., Synnaeve, G., Verbeek, J., & Jegou, H. (2023). ResMLP: Feedforward Networks for Image Classification with Data-Efficient Training. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(4), 5314–5321. <https://doi.org/10.1109/TPAMI.2022.3206148>
- Wang, H., Vaze, S., & Han, K. (2025). Dissecting Out-of-Distribution Detection and Open-Set Recognition: A Critical Analysis of Methods and Benchmarks. *International Journal of Computer Vision*, 133(3), 1326–1351. <https://doi.org/10.1007/s11263-024-02222-4>
- Wang, K., Ma, Q., Shen, C., & Lu, J. (2025). Application of Uncertainty to Out-of-Distribution Detection for Autonomous Driving Perception Safety. *IEEE Transactions on Intelligent Transportation Systems*, 26(8), 11276–11293. <https://doi.org/10.1109/TITS.2025.3557996>
- Wu, B., Jiang, J., Ren, H., Du, Z., Wang, W., Li, Z., Cai, D., He, X., Lin, B., & Liu, W. (2023). Towards In-Distribution Compatible Out-of-Distribution Detection. *The Thirty-Seventh AAAI Conference on Artificial Intelligence*, 39(9), 10333–10341. <https://doi.org/10.1609/aaai.v37i9.26230>
- Wu, Y. C., Lyu, S. H., Shang, H., Wang, X., & Qian, C. (2024). Confidence-aware Contrastive Learning for Selective Classification. *Proceedings of Machine Learning Research*, 235, 53706–53729. <https://doi.org/10.48550/arXiv.2406.04745>
- Xia, G., & Bouganis, C. S. (2024). Augmenting the Softmax with Additional Confidence Scores for Improved Selective Classification with Out-of-Distribution Data. *International Journal of Computer Vision*, 132(9), 3714–3752. <https://doi.org/10.1007/s11263-024-02029-3>
- Yang, J., Wang, P., Zou, D., Zhou, Z., Ding, K., Peng, W., Wang, H., Chen, G., Li, B., Sun, Y., Du, X., Zhou, K., Zhang, W., Hendrycks, D., Li, Y., & Liu, Z. (2022). OpenOOD: Benchmarking Generalized Out-of-Distribution Detection. *Advances in Neural Information Processing Systems*, 35. <https://arxiv.org/pdf/2210.07242>
- Yang, J., Zhou, K., Li, Y., & Liu, Z. (2024). Generalized Out-of-Distribution Detection: A Survey. *International Journal of Computer Vision*, 132(12), 5635–5662. <https://doi.org/10.1007/S11263-024-02117-4>
- Yu, Y., Shin, S., Ko, M., & Lee, K. (2024). Exploring using jigsaw puzzles for out-of-distribution detection. *Computer Vision and Image Understanding*, 241(February), 103968. <https://doi.org/10.1016/j.cviu.2024.103968>
- Zhao, Z., & Cao, L. (2024). Weighting non-IID batches for out-of-distribution detection. *Machine*

Learning, 113(10), 7371–7391.
<https://doi.org/10.1007/s10994-024-06605-z>