Distribution-Aware Feature Selection for Post-hoc Out-of-Distribution Detection

Max Gutbrod[1,2], David Rauber[1], and Christoph Palm[1,2]

Regensburg Medical Image Computing (ReMIC)[1] and Research Center of Health Sciences and Technology (RCHST)[2], OTH Regensburg, Regensburg, 93053, Germany

Abstract. Robust out-of-distribution (OOD) detection is crucial for deploying deep learning systems in high-stakes settings such as medical imaging and autonomous driving. Post-hoc feature-based detectors are attractive because they operate on pre-trained models, yet they often treat all feature dimensions as equally informative. We show that OOD-discriminative information in deep feature representations is often concentrated in subsets of features and is frequently axis-aligned. To exploit this structure, we introduce a distribution-aware feature selection strategy that ranks feature dimensions according to the discrepancy between in-distribution (ID) and OOD feature distributions, using the Wasserstein-1 distance as a principled metric.

To avoid the need for curated OOD validation data, we construct proxy-OOD data based on cross-domain mixup and evaluate adversarial perturbations as an alternative. Our method is a lightweight, broadly applicable add-on to existing feature-based OOD detectors that requires no retraining or changes to scoring functions. Evaluations across public benchmarks spanning medical and natural image domains show consistent performance improvements, while the reduced feature set lowers computational complexity and enables substantially faster inference. Our code is available at `https://github.com/remic-othr/mfs-ood`.

Keywords: OOD Detection · Feature Selection · Distributional Shift

1 Introduction

Out-of-distribution (OOD) samples – inputs drawn from distributions unseen during training – pose substantial risks when Machine Learning (ML) models are deployed in safety-critical domains such as autonomous driving and healthcare. ML models typically assume in-distribution (ID) inputs, making them unreliable under distribution shift [4, 36]. This challenge has long been recognized in ML research and is now echoed in the European Union's Artificial Intelligence (AI) Act, which requires high-risk AI systems to be resilient to unforeseen scenarios and to incorporate safeguards against undesirable behavior [1].

To allow fair comparisons among OOD detection methods, standardized frameworks were developed. OpenOOD [93, 98] provides benchmarks on natural images, while OpenMIBOOD [29] focuses on medical imaging. Both frameworks contain three OOD settings: covariate-shifted ID (cs-ID), near-OOD, and far-OOD, thereby providing a comprehensive basis for systematic evaluation.

OOD detection approaches can broadly be divided into training-based and post-hoc methods. While training-based methods modify the learning objective or model architecture, post-hoc approaches operate on a pre-trained network without requiring retraining. Among post-hoc approaches, feature-based methods have emerged as a strong and widely adopted family, leveraging intermediate representations of deep neural networks to distinguish ID from OOD samples. These methods operate on high-dimensional representations, often treating all feature dimensions as equally informative. However, prior work has shown that feature dimensions contribute unequally to OOD detection, and that retaining, reweighting, or projecting features can substantially improve performance [2,24,42,88,91,96]. These findings motivate studying how OOD-discriminative information is structured in learned feature representations and whether this structure can be exploited to identify informative feature dimensions.

Similar to other works [19,82,96], our approach relies on marginal feature selection (FS), which is most effective when OOD-discriminative signals are at least partly axis-aligned. Batch normalization (BN) has been shown to promote orthogonality and reduce correlations between feature directions in certain settings in deep random networks, which may encourage less redundant representations [15]. At the same time, dimension-wise nonlinearities such as ReLU operate independently on each feature and promote sparse activations [25]. While orthogonality alone does not imply axis alignment, the combination of decorrelation and sparsity may encourage representations in which discriminative information concentrates in a subset of features, which we investigate empirically.

To evaluate whether OOD-discriminative information is indeed concentrated in individual feature dimensions, we perform a rotation study on features extracted from several pre-trained CIFAR-10 models across five architectures [11]. We compare OOD detection performance using AUROC in three settings: (1) in the original feature space, (2) after applying our marginal FS (Sec. 4.1), and (3) after applying random Haar orthogonal rotations [80] to the feature space followed by marginal FS. These rotations preserve Euclidean distances and thus leave the detector's full-dimensional performance unchanged, while redistributing axis-aligned signals across multiple dimensions. Consequently, if AUROC after marginal FS degrades under rotation, this suggests that the representation relies on axis-aligned discriminatory structure. Figure 1 (left, top) provides empirical evidence on the evaluated datasets that OOD-discriminative information can concentrate in individual dimensions across several architectures. However, such axis-aligned discriminatory features are not guaranteed to transfer across OOD shifts and datasets. Therefore, in Fig. 1 (left, bottom) we evaluate AUROC performance when features are selected using one OOD dataset and evaluated on another. Supp. Sec. A describes the experimental setup in more detail.

The results indicate that OOD-discriminative information is often axis-aligned and can partially transfer across distribution shifts, a structure our Wasserstein-1 distance (WD)-based FS strategy exploits to identify discriminative feature di-

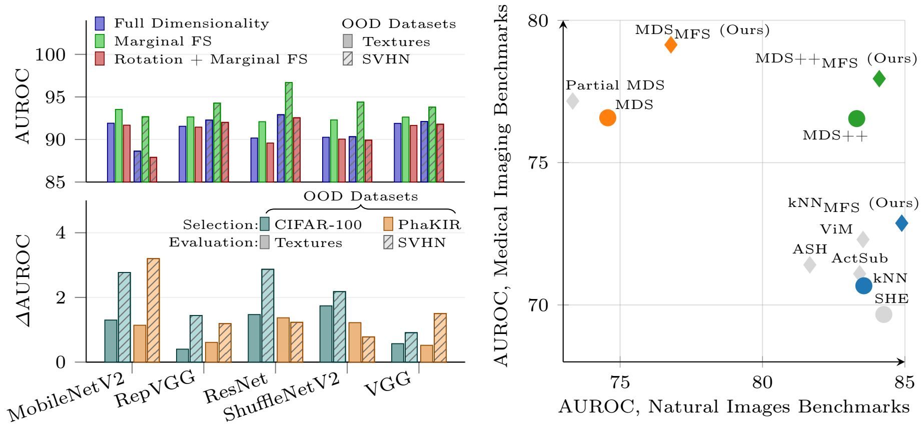

Fig. 1: Left (top): Marginal feature selection (FS) increases AUROC by retaining the features that are most discriminative between ID and OOD for each OOD dataset. Applying random orthogonal rotations before FS reduces AUROC, suggesting that OOD-discriminative information is largely axis-aligned. **Left (bottom)**: AUROC also improves when features are selected on one OOD dataset and evaluated on another, indicating that discriminative features can transfer across OOD shifts. ΔAUROC corresponds to AUROC after marginal FS minus AUROC obtained using all dimensions. In all cases, CIFAR-10 was used as ID dataset. **Right**: AUROC comparison across natural and medical imaging benchmarks, averaged over cs-ID, near-OOD, and far-OOD settings. Diamonds indicate methods using feature space transformations, while circles indicate those without. The proposed marginal FS improves results in both domains.

mensions for OOD detection. Estimating such separability, however, requires examples of distributional shift. In practical deployment, the target OOD distribution is unknown and labeled OOD validation data is unavailable. Consequently, post-hoc methods rely on proxy shifts for tasks such as threshold calibration and detector tuning. We construct these proxies via cross-domain mixup [97], which allows us to generate shifts of varying strength while using only a small cross-domain pool. Importantly, our strategy is designed to be broadly applicable to feature-based OOD detection methods, typically without requiring modifications to the base approach. To assess its effectiveness across different OOD detectors and data domains, we evaluated our FS strategy using three feature-based OOD detection methods on three medical imaging benchmarks and two natural image benchmarks. Figure 1 (right) shows that our WD-based marginal FS consistently improves OOD detection performance across these benchmarks. Our contributions are as follows:

– We provide empirical evidence that OOD-discriminative information in deep feature representations is often concentrated in a subset of feature dimensions and that these dimensions can transfer across distribution shifts.

- We introduce a distribution-aware feature selection strategy for post-hoc OOD detection that measures separability between ID and OOD feature distributions using the Wasserstein-1 distance.
- We propose a practical proxy-OOD validation strategy based on cross-domain mixup and evaluate adversarial perturbations as an alternative.
- We demonstrate consistent improvements across natural-image and medical-imaging benchmarks, while significantly reducing inference cost by removing non-informative feature dimensions.

2 Preliminaries

OOD Detection Problem Statement. Let $\mathcal{X} = \mathbb{R}^p$ denote the p-dimensional input space, $\mathcal{Y} = \{1, 2, \ldots, c\}$ the label space with c target classes, and $\mathcal{D} = \{(\mathbf{x}_i, y_i)\}_{i=1}^{N}$, with $\mathbf{x}_i \in \mathcal{X}$ and $y_i \in \mathcal{Y}$, the ID training dataset consisting of N samples. Let $f_{\mathcal{D}} : \mathcal{X} \to \mathcal{Y}$ be a deep neural network (DNN) trained on $\mathcal{D}$, composed of an encoder $h_\theta : \mathcal{X} \to \mathcal{Z} = \mathbb{R}^D$ and a classifier $g_\phi : \mathcal{Z} \to \mathcal{Y}$, such that $y = f(\mathbf{x}) = g_\phi(h_\theta(\mathbf{x}))$. The latent feature representation is $\mathbf{z}_i = h_\theta(\mathbf{x}_i) \in \mathcal{Z}$.

To perform OOD detection on $\mathbf{x}$, a scoring function $S : \mathbf{x} \mapsto \mathbb{R}$ assigns a scalar score. The binary classifier G separates ID from OOD using λ as threshold:

$$G(\mathbf{x}; \lambda) = \begin{cases} \text{ID} & S(\mathbf{x}) \geq \lambda, \\ \text{OOD} & S(\mathbf{x}) < \lambda. \end{cases} \tag{1}$$

For methods that operate on intermediate features, $\mathbf{z}$ is used in place of $\mathbf{x}$.

OOD Scenarios. Standardized benchmarks such as OpenOOD [93,98] and OpenMIBOOD [29] distinguish between three types of OOD scenarios, which provide a principled basis for evaluating OOD detection methods across varying degrees of distributional shift:

- cs-ID: Samples that retain the same labels but differ in input distribution, e.g., images acquired with a different device compared to the ID data.
- near-OOD: Samples with different underlying labels, typically accompanied by changes in input distribution.
- far-OOD: Samples from entirely different contexts, with both label changes and pronounced covariate shifts [94].

3 Related Work

We first provide a brief overview of the two main families of OOD detection: training-based and post-hoc approaches. We then review studies that transform feature representations to improve OOD detection.

Training-based Approaches. Training-based methods modify the model architecture [18,89] or loss function [31,37] to explicitly account for OOD inputs. During training, these approaches often require auxiliary OOD data, obtained from external datasets, adversarial generation, or data augmentation. Examples include extending the network with OOD-specific branches [18] or with auxiliary output classes [89]. In contrast, Hendrycks *et al.* [37] and Hein *et al.* [31] train with real OOD samples in the loss, while Yang *et al.* [95] train with mixup-generated samples [97] derived from ID data and additional affine transformations. While effective, these methods come at the cost of increased training complexity and require retraining before they can be implemented in existing systems.

Post-hoc Approaches. Post-hoc methods rely solely on outputs from a pre-trained network. They require no retraining and thus preserve the performance of the downstream task. These methods can be grouped into three categories: classification-based, feature-based, and hybrid methods [29].

- Classification-based methods rely solely on logits or softmax probabilities [8, 23, 28, 32, 33, 36, 50, 55, 56].
- Feature-based methods use representations from the penultimate layer [21, 49, 64, 83, 88, 99] or combine information from all intermediate layers [49, 76].
- Hybrid methods combine both types of information [19, 53, 54, 81, 82, 88, 92].

Among the feature-based approaches, three top-performing approaches across medical- and natural-image benchmarks are MDSEns/MDS [49], MDS++ [64], and kNN [83]. Lee *et al.* [49] compute the Mahalanobis distance score (MDS) between a test sample and the nearest class-conditional Gaussian distribution estimated from the ID training data on the penultimate layer. MDSEns extends this approach by incorporating intermediate layers, for which aggregation weights are tuned using an OOD validation set, approximated with adversarially perturbed ID data [26]. Mueller *et al.* introduce MDS++ [64], normalizing features prior to MDS computation. Sun *et al.* propose kNN [83], adopting a non-parametric approach using the k-th nearest ID neighbor distance as score.

Generative and unsupervised OOD detection. Generative and unsupervised approaches model ID structure without relying on classifier confidence, *e.g.*, through likelihood-ratio scores, autoencoder-style reconstruction errors, or self-supervised representation learning [5,70,77,84]. However, we omit this branch from our empirical comparison since it is complementary to our focus on post-hoc feature-based detectors for pretrained discriminative models.

Feature Space Transformation for OOD Detection. Prior work improves OOD detection by transforming feature representations, including marginally selecting informative subsets [19,82,96], jointly projecting features into lower-dimensional subspaces [24, 42, 88, 91, 101], and reshaping features [69, 81, 92].

Woodland *et al.* [91] evaluate a class-agnostic MDS [49] in combination with principal component analysis (PCA) in the context of medical imaging. Kamoi and Kobayashi [42] use PCA to select dimensions with the least explained variance. Ghosal *et al*. [24] propose a training-time subspace learning approach tailored to kNN-based OOD detection. Since it requires retraining the backbone, it is fundamentally different from our post-hoc setting and is therefore not included in the empirical comparison. In [101], Zongur *et al*. separate the feature space into task-relevant and residual subspaces. They then fuse a pruned energy score from the task-relevant space with a similarity-based density estimate from the residual components. Yuan *et al*. [96] use a discriminability score based on inter-class similarity and variance to mask less informative dimensions. Residual and ViM [88] project features into a subspace spanned by top eigenvectors of the ID covariance, assuming deviations from this subspace indicate OOD samples. ViM continues to use features from this subspace to compute an energy score based on logits. ReAct [81] clips features that show an abnormally high activation based on given ID activations. Djurisic *et al*. [19] mask features based on the lowest pth-percentile of feature activations. Instead of masking features, Xu *et al*. [92] use the pth-percentile as a dimensionality-wise scaling parameter. Similarly, Regmi [69] calculates an adaptive per-sample scaling parameter.

Although prior work has applied WD to OOD detection [21, 58, 90] and to dimensionality reduction outside the OOD setting [13, 14, 20], to the best of our knowledge, it has not been used for FS in OOD detection. In contrast to other methods that employ WD, we use it to identify discriminative feature dimensions rather than as an OOD scoring function. The resulting subset can then be used with feature-based OOD detectors without modifying their scoring functions.

4 Methodology

In this section, we present our method to select the most discriminative feature dimensions for OOD using the Wasserstein-1 distance (WD) (Supp. Sec. G) and describe how we construct OOD validation data.

4.1 Wasserstein Distance-based Feature Selection

Effective FS for OOD detection should yield feature spaces in which ID and OOD samples remain separable. Unlike dimensionality reduction (DR), which constructs new representations, FS preserves a subset of the original dimensions. Neither established DR techniques, such as PCA [39], stochastic neighbor embeddings (t-SNE) [60], and uniform manifold approximation and projection (UMAP) [61], nor existing FS strategies for OOD detection are explicitly optimized for this objective (cf. Sec. 3). To close this gap, we quantify ID/OOD separability using the WD, a distributional discrepancy measure that captures differences in feature distributions. It detects differences in spread, shape, or multimodality, remains well-defined for non-overlapping supports, and avoids strong parametric assumptions, making it well-suited for measuring separability.

Ideally, one would evaluate all joint combinations of feature dimensions that provide the best ID/OOD separability, but this is computationally intractable

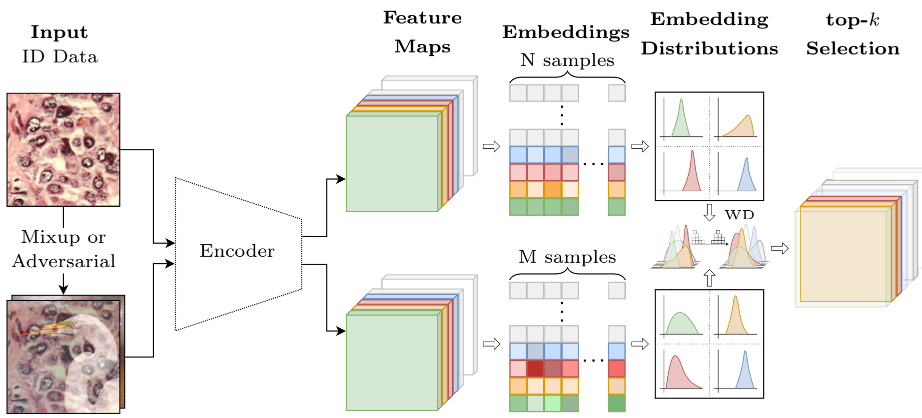

Fig. 2: Illustration of our strategy for identifying features that best differentiate between in-distribution (ID) and out-of-distribution (OOD) data. Samples from the ID dataset and synthetic samples, created via mixup [97] or adversarial perturbation [26], are input to a DNN. Feature maps are spatially reduced using average pooling. The resulting set of embeddings forms feature-wise distributions for both ID and synthetic OOD samples. The Wasserstein-1 distance between these distributions is used to identify the top-k most discriminative feature dimensions for OOD detection. **Credits**: Cancer cells, Dr. Cecil Fox, Public Domain; White Egret, Paul Brennan, CC0.

even for moderate feature dimensionality, as outlined in Supp. Sec. B. As a practical alternative, we compute the marginal WD for each feature independently. Let ρ and ν denote the empirical distributions of the ID and OOD feature representations in $\mathcal{Z}$, respectively. For a given dimension $d \in \{1, \dots, D\}$, let $\{a_1^{(d)}, \dots, a_N^{(d)}\}$ and $\{b_1^{(d)}, \dots, b_M^{(d)}\}$ denote the corresponding ID and OOD feature values. Sorting these values separately allows efficient computation of the WD via the area between their empirical cumulative distribution functions $F_\rho^{(d)}$ and $F_\nu^{(d)}$:

$$W_1^{(d)} = \int_{-\infty}^{\infty} \left| F_\rho^{(d)}(t) - F_\nu^{(d)}(t) \right| dt. \tag{2}$$

This calculation scales as $\mathcal{O}(D \cdot (N \log N + M \log M))$, dominated by sorting. To ensure comparability of WD values across feature dimensions, we normalize features using ID-based min-max scaling.

We pick the top-k dimensions $\mathcal{I}_k$ with the highest WD:

$$\tau = \operatorname{argsort}\left(W_1^{(1)}, \dots, W_1^{(D)}\right) \text{ in descending order, with } \mathcal{I}_k = \{\tau_1, \dots, \tau_k\}. \tag{3}$$

This marginal feature selection (MFS) strategy retains the dimensions that marginally contribute most to ID/OOD divergence, relying solely on the global, class-agnostic structure of the feature space. By removing features that contribute little to ID/OOD separability, MFS reduces noise in distance- and density-based scores, thereby improving detector discrimination. Crucially, MFS is performed once as a preprocessing step. During inference, detectors operate on the reduced feature space, which lowers computational cost. The retained fraction α is chosen by hyperparameter search over a held-out OOD validation set, selecting the value that maximizes validation AUROC. When no reduced fraction improves over the full space, the search selects α=1.0, so the method falls back to the full feature space and exactly preserves the original detector performance. Such results are marked with a gray background in our result tables. Importantly, our approach does not strictly require axis-aligned separability, but rather exploits it, when present. Because penultimate-layer dimensionality varies across architectures, we tune the fraction of retained feature dimensions α and compute $k=\lfloor \alpha D \rfloor$ rather than selecting fixed dimensions. An overview of the method is presented in Fig. 2.

4.2 OOD Validation Data

One component of our MFS strategy is the use of a validation distribution to identify discriminative feature dimensions. Ideally, this distribution should resemble the types of OOD data expected during deployment. However, curated OOD validation datasets are often unavailable.

A possible workaround is to generate adversarial samples from ID data, as proposed by Lee et al. [49], using Fast Gradient Sign Method (FGSM) [26]. This approach avoids external datasets and relies solely on ID samples, but it assumes that adversarial perturbations produce feature activations that reveal separability between ID and OOD data. For models trained with adversarial robustness in mind, however, changes in feature activations may be suppressed. In this case, adversarial samples may provide weak separability signals, limiting their effectiveness for MFS.

As a practical alternative, we construct proxy-OOD validation data using mixup [97]. Mixup is commonly applied between ID samples, producing interpolations that remain close to the ID manifold in feature space. While this can improve generalization, it may not induce sufficiently strong OOD shifts, particularly in models trained with mixup. To generate samples that move representations away from the ID manifold, we instead mix ID samples with cross-domain data readily available in standard OOD benchmarks. As a cross-domain source, we use natural images for medical models, and medical images for models trained on natural images, which avoids semantic overlap with ID classes.

To induce a range of domain shifts, we generate mixup samples with a randomly selected ID ratio $r \in \{0.25, 0.5, 0.75\}$, computing proxy-OOD samples as $\mathbf{x}_{\mathrm{OOD},i} = r_i \mathbf{x}_i + (1 - r_i)\mathbf{m}_j$, where $\mathbf{x}$ is the i-th sample in the batch, r_i a randomly selected ratio for the ID sample, and $\mathbf{m}_j$ a randomly drawn sample from the cross-domain pool.

Both adversarial samples and cross-domain mixup provide effective proxies but induce different shifts. The most suitable proxy may therefore depend on deployment conditions and model robustness. However, MFS does not need the proxy to identify the feature subset that would be optimal for any particular

Table 1: Performance comparison (AUROC↑/FPR@95↓) of all evaluated methods averaged across benchmarks and OOD categories. Improvements over the respective base OOD detector are bolded. Results for which α equals 1.0 have a gray background. *OASIS-3 uses LUNA16 3D scans as proxy data (cf. Sec. 5). Results for our variants are mean ± standard error (Supp. Sec. E). Best results for each column are underlined.

	Method			MIDOG		PhaKIR		OASIS-3*		CIFAR-10		ImageNet1k		Average	
				AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓
Direct Comparison	kNN	(2022)	[83]	70.64	67.88	42.41	94.93	98.96	5.28	88.02	44.15	79.10	58.80	75.83	54.21
	kNN_{MFS}	(Ours)		**76.66** ±0.140	**60.68** ±0.241	**42.41** ±0.000	**94.93** ±0.000	**99.55** ±0.015	**2.82** ±0.138	**88.41** ±0.001	**42.93** ±0.006	**81.37** ±0.023	**55.64** ±0.043	**77.68** ±0.028	**51.40** ±0.043
	MDS	(2018)	[49]	70.91	69.73	61.33	79.65	97.49	7.86	84.54	52.35	64.60	76.11	75.77	57.14
	MDS_{MFS}	(Ours)		**76.16** ±0.163	**58.77** ±0.206	**62.74** ±0.060	**78.91** ±0.141	**98.52** ±0.017	**6.10** ±0.082	**84.79** ±0.005	**51.93** ±0.012	**68.78** ±0.021	**70.61** ±0.035	**78.20** ±0.038	**53.27** ±0.049
	MDS++	(2025)	[64]	72.16	65.74	57.64	89.76	99.85	0.71	86.44	52.20	80.18	55.16	79.25	52.71
	$\text{MDS++}_{\text{MFS}}$	(Ours)		**76.78** ±0.225	**62.35** ±0.534	57.23 ±0.108	91.29 ±0.149	99.84 ±0.005	0.76 ±0.053	**86.82** ±0.012	**50.56** ±0.044	**81.38** ±0.004	**53.46** ±0.016	**80.41** ±0.049	**51.68** ±0.107
Subspace	Partial MDS	(2020)	[42]	71.88	67.58	62.18	79.49	97.43	7.92	82.66	56.56	64.01	76.74	75.63	57.66
	Residual	(2022)	[88]	72.79	66.23	63.81	79.08	97.95	6.78	82.74	56.70	58.80	84.93	75.22	58.74
	ViM	(2022)	[88]	72.16	66.24	50.27	94.77	94.47	14.52	86.69	50.39	80.37	54.73	76.79	56.13
	PCA MDS	(2024)	[91]	73.56	63.90	66.01	76.35	98.69	4.85	81.83	57.46	55.60	87.21	75.14	57.95
	ActSub	(2025)	[101]	69.94	70.83	52.15	81.98	91.19	26.98	80.70	64.27	86.12	44.76	76.02	57.76
Marginal Selection	ASH	(2022)	[19]	67.78	80.59	66.88	71.73	79.57	40.70	79.50	81.73	83.82	50.76	75.51	65.10
	DICE	(2022)	[82]	62.59	87.83	43.66	91.53	23.95	94.32	81.44	72.99	79.86	62.26	58.30	81.79
	DDCS	(2024)	[96]	69.97	76.45	49.27	81.63	63.81	62.92	81.03	73.30	80.91	57.64	69.00	70.39
Feature Shaping	ReAct	(2021)	[81]	65.35	79.76	35.04	97.43	64.16	65.56	84.13	67.10	82.07	55.46	66.15	73.06
	SCALE	(2023)	[92]	65.09	84.89	65.84	72.40	89.02	31.80	80.64	79.08	84.43	49.98	77.01	63.63
	AdaScale-A	(2025)	[69]	60.48	82.40	61.65	88.38	93.72	28.21	82.00	75.46	86.17	44.86	76.80	63.86
Unrel.	SHE	(2023)	[99]	70.26	68.05	44.64	96.39	94.10	14.30	86.43	48.12	82.09	53.22	75.50	56.01
	NCI	(2025)	[54]	69.71	69.74	52.28	81.83	66.55	57.18	85.32	57.35	83.05	52.59	71.38	63.74

unknown deployment OOD distribution but only to reveal features that are broadly useful for separating ID from OOD.

5 Experiments and Results

This section describes the choice of the selected post-hoc feature-based OOD detection methods and additional baselines, along with the experimental setup and metrics. We then report and discuss the results.

Evaluated Methods. We extend three base OOD detectors with our MFS strategy (Sec. 4.1), resulting in kNN_{MFS}, MDS_{MFS}, and $\text{MDS++}_{\text{MFS}}$. Primarily, we compare these methods to their original counterparts. Additionally, we compare them against several families of OOD detectors. First, we include subspace-based DR methods (Partial MDS [42], PCA MDS [91], ActSub [101], Residual [88], ViM [88]), which utilize joint projections of the feature space. Second, we evaluate marginal FS approaches (ASH [19], DDCS [96], DICE [82]). Third, we consider feature shaping methods (ReAct [81], SCALE [92], AdaSCALE-A [69]) that modify activations without explicit dimension masking. Finally, we include strong post-hoc baselines (SHE [99], NCI [54]) that are competitive but methodologically unrelated to our contribution. COMBOOD [67] is not included in our comparison as we were unable to match the reported performance (Supp. Sec. E.1). Similar discrepancies have also been reported in previous work [40,79].

Experimental Setup. We show the effectiveness of our MFS by evaluating two natural image benchmarks and three medical imaging benchmarks from OpenOOD [93] and OpenMIBOOD [29], respectively: CIFAR-10, ImageNet1k, MIDOG, PhaKIR, and OASIS-3. In total, this covers 16 natural [9,12,16,17,34,35,38,45,48,65,68,87,88,100] and 14 medical datasets [3,6,7,10,30,41,43,44,47,51,62,66,72–75,85,86] as presented in Supp. Tab. C1. To ensure comparability, we use the classification models provided by each framework, as listed in Supp. Tab. C2.

For cross-domain mixup, we use the official validation splits of PhaKIR and CIFAR-10 as auxiliary data. For the medical imaging benchmark OASIS-3, however, we deviate from the cross-domain setting as the underlying classifier operates on 3D volumes. Mixing 2D natural images into volumetric data is ill-defined: it would require either repeating a single image across slices or inserting different images per slice, both of which introduce artificial structures unrelated to plausible domain shifts. Moreover, even small cross-domain contributions induce large distribution shifts, preventing the generation of samples spanning different OOD severities. Instead, we use the in-domain far-OOD dataset LUNA16 [78], which consists of 3D lung CT scans and is not part of the OASIS-3 benchmark. These proxy-OOD data are used only for the MFS preprocessing step, i.e., to rank feature dimensions and select the retained feature fraction α. All subsequent method-specific hyperparameter tuning follows the default OpenOOD/OpenMIBOOD validation protocol.

Metrics. We evaluate all methods using two standard metrics: AUROC and false positive rate (FPR). AUROC quantifies overall discriminability between ID and OOD across all thresholds, with higher values indicating better performance. FPR@95 measures the false positive rate when 95 % of OOD samples are correctly detected, with lower values indicating better performance.

5.1 Benchmark Results

The results for all evaluated methods and benchmarks are shown in Tab. 1. We report the averaged results for each benchmark across the three evaluation settings cs-ID, near-OOD, and far-OOD, and present detailed results in Supp. Sec. E.

Comparison to Original Methods. Tab. 1 shows that the MFS variants consistently improve their corresponding base detectors. This trend holds for both kNN- and Mahalanobis-based methods and is reflected in higher AUROC and lower FPR@95 values. Notably, augmenting the simple kNN detector with MFS yields performance competitive with, and in some cases exceeding, recent state-of-the-art methods, achieving the lowest overall FPR@95 across benchmarks.

Performance gains are particularly pronounced for MDS, suggesting that MFS effectively suppresses noisy dimensions while preserving the covariance structure required by the method. Further, even the strong MDS++ baseline benefits from MFS, yielding the best overall detection performance across the evaluated benchmarks. Notably, $\text{MDS++}_{\text{MFS}}$ is the only evaluated method that decreases the FPR@95 to zero in the cs-ID regime for OASIS-3.

Besides these gains in AUROC and FPR@95, MFS reduces runtime on average by 58.89 % (kNN_{MFS}), 27.03 % (MDS_{MFS}), and 42.76 % ($\text{MDS++}_{\text{MFS}}$) compared to the original methods. For instance, by selecting only 1 % of the feature dimensions, kNN_{MFS} on MIDOG reduces runtime by ≈97 %, while simultaneously improving AUROC by 5.74 percentage points (p.p.) and FPR@95 by 6.75 p.p. The runtime performance results can be found in Supp. Sec. E.2.

Comparison to Remaining Baselines. Across benchmarks, our MFS variants outperform all remaining baselines on average, although a small number of individual baseline results are competitive. Several baselines exhibit strong dataset-dependent behavior, underscoring the importance of multi-benchmark evaluation across domains. ActSub [101] and AdaScale-A [69] achieve outstanding performance on ImageNet1k (IN1k), yet fail to maintain this performance level on other benchmarks. Conversely, Residual [88] and PCA MDS [91] perform well in the medical setting, but fail on IN1k. Among the three baselines using marginal FS strategies, which are conceptually closest to our approach, only ASH [19] shows comparatively robust performance across benchmarks. In contrast, DICE [82] and DDCS [96] exhibit highly variable performance across benchmarks.

To ensure a controlled comparison, we additionally tune all marginal selection and subspace baselines with the same proxy-OOD protocol used by our MFS variants. In the main results, feature selection for our variants is based on proxy-OOD data, while feature selection or subspace learning for the baselines uses real-OOD validation data. Tuning the baselines on real-OOD data rather than our proxy raises their AUROC by +0.65 points on average for subspace methods and +3.04 for marginal-selection methods. Conversely, granting our variants the same real-OOD validation set increases their AUROC by +2.41 (kNN_{MFS}), +1.02 (MDS_{MFS}), and +0.82 ($\text{MDS++}_{\text{MFS}}$). The reported gains of MFS are therefore conservative and would widen further if MFS also used real-OOD data for feature selection. Further details are provided in Supp. Sec. E.4.

Performance on the PhaKIR Benchmark. The PhaKIR benchmark reveals a failure mode of distance-based OOD detection in strongly anisotropic feature spaces. The analysis of the feature space of PhaKIR's classification model shows that ID/OOD separability is distributed across many feature dimensions, yet kNN_{MFS} fails to identify a viable subset. This behavior can be explained by variance dominance in the feature space. A large portion of the discriminative information aligns with high-variance directions, dominating kNN distances and suppressing contributions from ID/OOD separating low-variance components. Consequently, validation AUROC increases steadily as more features are retained, preventing subset selection (Supp. Fig. E5). Whitening removes this dominance by equalizing variance across dimensions, allowing the distributed discriminative signal to influence distances. The validation performance then saturates early, enabling kNN_{MFS} to select a subset and improve AUROC by ≈14 points. An analysis supporting this interpretation is provided in Supp. Sec. E.3.

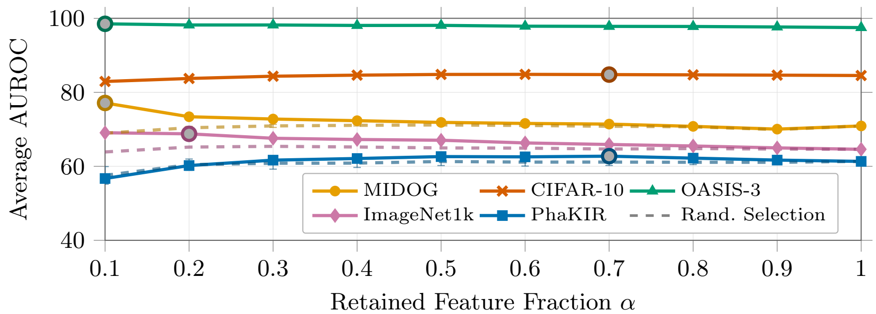

Fig. 3: Average test AUROC across cs-ID, near-OOD, and far-OOD settings for varying retained feature fractions α (0.1, ..., 1.0) on all benchmarks using $\mathrm{MDS}_{\mathrm{MFS}}$. In each benchmark, MFS selects a feature subset based on validation data (gray circles) at or close to the performance optimum. Additionally, ten randomly selected subsets are evaluated and averaged for each α for MIDOG, PhaKIR, and IN1k (dashed lines).

5.2 Influence of Retained Feature Fraction α

The sensitivity analysis in Fig. 3 examines the performance of $\mathrm{MDS}_{\mathrm{MFS}}$ across different retained feature fractions α. For most benchmarks, AUROC remains stable over a broad range of dimensionalities, indicating substantial redundancy in the learned representations for ID/OOD discrimination. On MIDOG, performance decreases as more dimensions are retained, suggesting that additional features introduce noise rather than useful signal. OASIS-3 shows a similar but weaker trend, with a slight decline toward higher dimensionalities. CIFAR-10 remains largely insensitive to α, demonstrating that strong reduction is possible without degrading detection performance. In contrast, PhaKIR improves steadily up to α=0.7, confirming that separability is distributed across many features and that aggressive reduction degrades performance significantly. Notably, in all cases, the selected α-value lies on or near the performance optima, indicating that our MFS strategy consistently identifies viable subsets.

5.3 Marginal Selection Metrics and Proxy-OOD Construction

The proposed marginal FS requires (i) a feature-wise measure of distribution shift and (ii) a proxy distribution that approximates OOD data, both of which we evaluate in this section. To assess the influence of the proxy shift, we evaluate adversarial perturbations and cross-domain mixup as OOD distributions for MFS (Sec. 4.2). We also compare against random feature subsets to test whether gains come from our MFS rather than merely reducing dimensionality.

Shift measures. We evaluate four feature-wise discrepancy measures: Jensen-Shannon (JS) divergence [52], Maximum Mean Discrepancy (MMD) [27] with a linear kernel, MMD with a radial basis function (RBF) kernel and the Wasserstein-1 distance (WD). JS is preferred over Kullback-Leibler (KL) [46] because

Table 2: (a) AUROC comparison of feature-wise discrepancy measures – Jensen–Shannon (JS), MMD with linear and RBF kernels, and Wasserstein-1 Distance (WD) – for marginal FS and (b) proxy-OOD strategies: Adversarial (AD) and Mixup (MU). Results are averaged across medical and natural image benchmarks.

(a) Measure comparison

	kNN_{MFS}		MDS_{MFS}		$\text{MDS++}_{\text{MFS}}$	
	Med.	Nat.	Med.	Nat.	Med.	Nat.
JS	71.96	84.35	78.06	75.90	77.73	83.77
MMD_{lin}	72.86	84.72	79.10	**76.80**	77.16	84.02
MMD_{rbf}	**72.90**	84.73	79.11	76.72	77.42	83.96
WD	72.87	**84.89**	**79.14**	76.79	**77.95**	**84.10**

(b) Proxy-OOD comparison

	kNN_{MFS}		MDS_{MFS}		$\text{MDS++}_{\text{MFS}}$	
	Med.	Nat.	Med.	Nat.	Med.	Nat.
AD	72.79	84.84	76.42	**77.19**	76.85	84.01
MU	**72.87**	**84.89**	**79.14**	76.79	**77.95**	**84.10**

it is symmetric and smooths KL using a mixture distribution, which yields a stable measure that remains well-defined when supports do not overlap. The linear-kernel MMD_{lin} reduces to the difference between mean feature embeddings, whereas the RBF variant captures non-linear distributional differences. The RBF's bandwidth parameter is set using the median heuristic [63]. As shown in Tab. 2a, all measures yield comparable performance, indicating that MFS does not depend strongly on a specific shift metric. The near-identical performance of WD and both MMD variants suggests that the ID/OOD shift may be largely driven by location differences rather than more complex distributional differences, such as multimodality. We favor WD due to its parameter-free formulation and lower computational complexity; $\mathcal{O}(D \cdot (N \log N + M \log M))$ vs. $\mathcal{O}(D \cdot (N + M)^2)$ for MMD_{rbf}. Nevertheless, MMD_{lin} provides a computationally inexpensive alternative if the one-time setup cost for MFS is critical.

Comparison to random subset selection. To verify that performance gains stem from discriminative features rather than reduced dimensionality alone, we compare MDS_{MFS} to random feature subsets for MIDOG, PhaKIR, and IN1k. As shown in Fig. 3, MFS consistently outperforms random subsets by a large margin on MIDOG and IN1k. On PhaKIR, however, random subsets in the range between 0.1 and 0.2 slightly outperform MFS. This behavior reflects the interaction between distributed separability and variance-dominated distance geometry (Supp. Sec. E.3). Although informative low-variance directions exist for ID/OOD separation, selected subsets still include high-variance components that dominate kNN distances and suppress these informative features. In contrast, random selection occasionally removes many dominant directions, allowing the informative low-variance signal to influence the score. Comparing the distribution of the WD across features between PhaKIR and IN1k (Supp. Fig. E3) suggests that both the spread of discriminative information across features and its alignment with high-variance components (Supp. Fig. E4) influence gains achievable with MFS.

Proxy-OOD construction. As shown in Tab. 2b, both proxies yield comparable performance, with mixup providing a small advantage. This suggests that both constructions produce suitable proxy shifts for FS, indicating that the method is not strongly dependent on the specific proxy-OOD generation strategy. However, cross-domain mixup offers a practical default because it induces distribution shifts independent of model robustness and common training procedures. Importantly, gains across all OOD settings (cs-ID, near-OOD, far-OOD, Supp. Tab. E1) confirm that both proxies can generalize effectively for FS guidance. The proxy-stability analysis in Supp. Sec. F supports this, showing substantial overlap between selected features across different cross-domain proxy datasets.

Overall, these results suggest that performance gains are closely tied to the ability to identify informative feature dimensions, rather than to reduced dimensionality alone. Given its consistent performance across benchmarks, we use WD-based MFS with mixup-based proxy data as our default configuration.

5.4 Comparing Marginal Selection and Joint Subspace Learning

While previous results suggest that OOD-discriminative information is largely axis-aligned, such structure may also be distributed across combinations of features. To examine this possibility, we evaluate a Wasserstein Discriminant Analysis-inspired [22] joint subspace learning (SL) approach that learns projections maximizing ID/OOD separability. Full details on this approach can be found in Supp. Sec. D. Because the SL approach mixes feature dimensions, strong performance points to the presence of useful joint directions, whereas limited gains suggest that separability is predominantly axis-aligned. As shown in Supp. Tab. D2, MFS consistently improves performance across detectors and benchmarks, while SL yields mixed results. While it occasionally improves performance over its respective base detector, it often matches (α=1.0) or underperforms the original feature space and rarely surpasses MFS. These findings further support the interpretation that, for the considered benchmarks, discriminative structure is predominantly axis-aligned, and that selecting informative dimensions is more reliable than learning joint projections using a WD-based objective.

5.5 Generalization beyond Convolutional Backbones

Our evaluation focused on convolutional models, which are standard in the OpenOOD and OpenMIBOOD frameworks. However, since our method operates post-hoc on extracted features, it is, in principle, architecture-agnostic. To analyze the performance of MFS on a transformer architecture, we additionally evaluate Swin-T Transformers [59]. For comparability, we use the publicly available pre-trained checkpoint utilized in OpenOOD for IN1k, and the checkpoint from [57] for CIFAR-10, as OpenOOD did not evaluate Swin-T on this benchmark. The OpenMIBOOD benchmarks are not included due to the unavailability of public transformer-based models. For the Swin-T architecture, MFS yields consistent AUROC improvements over the base detectors (+0.21 for kNN, +0.25 for MDS, and +1.71 for MDS++). These gains are primarily driven by improved performance on CIFAR-10, whereas results on IN1k remain largely unchanged. Notably, the SL variants are competitive with MFS for this architecture, suggesting that Swin-T representations may contain more exploitable joint discriminative structure than those of common convolutional backbones. Results for all evaluated methods are shown in Supp. Tab. E2.

As a further preliminary check beyond discriminative backbones, we also apply MFS to representations from a generative Stable-Diffusion 2.1 VAE [71]. Despite having only four pooled channel dimensions, MFS still yields small AUROC gains for MDS on both MIDOG (+0.22 AUROC points) and CIFAR-10 (+1.08 AUROC points). These results suggest that MFS may also benefit generative and low-dimensional representations, motivating a broader study in future work. More details are shown in Supp. Sec. E.5.

6 Conclusion

We presented a distribution-aware marginal feature selection strategy for post-hoc OOD detection that identifies feature dimensions maximizing ID/OOD separability using the Wasserstein-1 distance. Extensive experiments across natural-image and medical-imaging benchmarks show that discriminative information is often concentrated in a subset of feature dimensions and can transfer across distribution shifts.

The proposed method is broadly applicable, requires no retraining, and integrates seamlessly with existing feature-based detectors. By operating on a reduced set of informative features, it improves detection performance while substantially accelerating inference. For example, augmenting the kNN detector reduces FPR@95 by 2.81 percentage points on average while lowering average runtime by 58.9 %. Performance gains diminish when discriminative information is broadly distributed or dominated by variance structure, suggesting benefits from whitening or hybrid strategies in strongly anisotropic feature spaces. Overall, distribution-aware feature selection provides a simple, efficient, and broadly applicable way to strengthen post-hoc OOD detection.

Acknowledgements

The authors declare that they have no competing interests. This work was supported by the Bavarian Academic Forum for Health. This study uses only publicly available and/or access-controlled datasets. Ethics approval and informed consent were handled by the original dataset providers where applicable. Data were provided in part by OASIS-3: Longitudinal Multimodal Neuroimaging: Principal Investigators: T. Benzinger, D. Marcus, J. Morris; NIH P30 AG066444, P50 AG00561, P30 NS09857781, P01 AG026276, P01 AG003991, R01 AG043434, UL1 TR000448, R01 EB009352. AV-45 doses were provided by Avid Radiopharmaceuticals, a wholly owned subsidiary of Eli Lilly.

References

1. Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act). *Official Journal of the European Union*, L 168, pp. 1–159 (2024), recital 75
2. Ahn, Y.H., Park, G., Kim, S.T.: Line: Out-of-distribution detection by leveraging important neurons. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023. pp. 19852–19862. IEEE (2023). https://doi.org/10.1109/CVPR52729.2023.01901, accessed: 2026-06-25
3. Allan, M., Kondo, S., Bodenstedt, S., Leger, S., Kadkhodamohammadi, R., Luengo, I., Fuentes, F., Flouty, E., Mohammed, A., Pedersen, M., et al.: 2018 robotic scene segmentation challenge. ArXiv preprint **abs/2001.11190** (2020), https://arxiv.org/abs/2001.11190, accessed: 2026-06-25
4. Amodei, D., Olah, C., Steinhardt, J., Christiano, P., Schulman, J., Mané, D.: Concrete problems in ai safety. ArXiv preprint **abs/1606.06565** (2016), https://arxiv.org/abs/1606.06565, accessed: 2026-06-25
5. An, J., Cho, S.: Variational autoencoder based anomaly detection using reconstruction probability. Special lecture on IE **2**(1), 1–18 (2015)
6. Atkinson Amorim, J.G., Matias, A., Bottamedi, T., Sanches, V., Costa, A.F., Onofre, F., Onofre, A., Wangenheim, A.: CCAgT: Images of Cervical Cells with AgNOR Stain Technique (2022), accessed: 2026-06-25
7. Aubreville, M., Wilm, F., Stathonikos, N., Breininger, K., Donovan, T.A., Jabari, S., Veta, M., Ganz, J., Ammeling, J., van Diest, P.J., et al.: A comprehensive multi-domain dataset for mitotic figure detection. Scientific data **10**(1), 484 (2023)
8. Bendale, A., Boult, T.E.: Towards open set deep networks. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016. pp. 1563–1572. IEEE Computer Society (2016). https://doi.org/10.1109/CVPR.2016.173, accessed: 2026-06-25
9. Bitterwolf, J., Müller, M., Hein, M.: In or out? fixing imagenet out-of-distribution detection evaluation. In: Krause, A., Brunskill, E., Cho, K., Engelhardt, B., Sabato, S., Scarlett, J. (eds.) International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA. Proceedings of Machine Learning Research, vol. 202, pp. 2471–2506. PMLR (2023), https://proceedings.mlr.press/v202/bitterwolf23a.html, accessed: 2026-06-25
10. Bodenstedt, S., Allan, M., Agustinos, A., Du, X., Garcia-Peraza-Herrera, L., Kenngott, H., Kurmann, T., Müller-Stich, B., Ourselin, S., Pakhomov, D., et al.: Comparative evaluation of instrument segmentation and tracking methods in minimally invasive surgery. ArXiv preprint **abs/1805.02475** (2018), https://arxiv.org/abs/1805.02475, accessed: 2026-06-25
11. Chen, Y.: Pytorch cifar models. https://github.com/chenyaofo/pytorch-cifar-models, accessed: 2026-06-25
12. Cimpoi, M., Maji, S., Kokkinos, I., Mohamed, S., Vedaldi, A.: Describing textures in the wild. In: 2014 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2014, Columbus, OH, USA, June 23-28, 2014. pp. 3606–3613. IEEE Computer Society (2014). https://doi.org/10.1109/CVPR.2014.461, accessed: 2026-06-25

13. Clark, R.A., Needham, T., Weighill, T.: Generalized dimension reduction using semi-relaxed gromov-wasserstein distance. In: Walsh, T., Shah, J., Kolter, Z. (eds.) AAAI-25, Sponsored by the Association for the Advancement of Artificial Intelligence, February 25 - March 4, 2025, Philadelphia, PA, USA. pp. 16082–16090. AAAI Press (2025). https://doi.org/10.1609/AAAI.V39I15.33766, accessed: 2026-06-25
14. Collas, A., Vayer, T., Flamary, R., Breloy, A.: Entropic wasserstein component analysis. In: 2023 IEEE 33rd International Workshop on Machine Learning for Signal Processing (MLSP). pp. 1–6. IEEE (2023)
15. Daneshmand, H., Joudaki, A., Bach, F.R.: Batch normalization orthogonalizes representations in deep random networks. In: Ranzato, M., Beygelzimer, A., Dauphin, Y.N., Liang, P., Vaughan, J.W. (eds.) Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual. pp. 4896–4906 (2021), https://proceedings.neurips.cc/paper/2021/hash/26cd8ecadce0d4efd6cc8a8725cbd1f8-Abstract.html, accessed: 2026-06-25
16. Deng, J., Dong, W., Socher, R., Li, L., Li, K., Li, F.: Imagenet: A large-scale hierarchical image database. In: 2009 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2009), 20-25 June 2009, Miami, Florida, USA. pp. 248–255. IEEE Computer Society (2009). https://doi.org/10.1109/CVPR.2009.5206848, accessed: 2026-06-25
17. Deng, L.: The mnist database of handwritten digit images for machine learning research [best of the web]. IEEE signal processing magazine **29**(6), 141–142 (2012)
18. DeVries, T., Taylor, G.W.: Learning confidence for out-of-distribution detection in neural networks. ArXiv preprint **abs/1802.04865** (2018), https://arxiv.org/abs/1802.04865, accessed: 2026-06-25
19. Djurisic, A., Bozanic, N., Ashok, A., Liu, R.: Extremely simple activation shaping for out-of-distribution detection. In: The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023. OpenReview.net (2023), https://openreview.net/pdf?id=ndYXTEL6cZz, accessed: 2026-06-25
20. Eufrazio, R.P., Montesuma, E.F., Cavalcante, C.C.: A dimensionality reduction technique based on the gromov-wasserstein distance. In: International Conference on Geometric Science of Information. pp. 111–120. Springer (2025)
21. Feng, C., Chen, W., Ke, A., Ren, Y., Xie, X., Zhou, S.K.: Detecting out-of-distribution samples via conditional distribution entropy with optimal transport. ArXiv preprint **abs/2401.11726** (2024), https://arxiv.org/abs/2401.11726, accessed: 2026-06-25
22. Flamary, R., Cuturi, M., Courty, N., Rakotomamonjy, A.: Wasserstein discriminant analysis. Machine Learning **107**(12), 1923–1945 (2018)
23. Gal, Y., Ghahramani, Z.: Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In: Balcan, M., Weinberger, K.Q. (eds.) Proceedings of the 33nd International Conference on Machine Learning, ICML 2016, New York City, NY, USA, June 19-24, 2016. JMLR Workshop and Conference Proceedings, vol. 48, pp. 1050–1059. JMLR.org (2016), http://proceedings.mlr.press/v48/gal16.html, accessed: 2026-06-25
24. Ghosal, S.S., Sun, Y., Li, Y.: How to overcome curse-of-dimensionality for out-of-distribution detection? In: Wooldridge, M.J., Dy, J.G., Natarajan, S. (eds.) Thirty-Eighth AAAI Conference on Artificial Intelligence, AAAI 2024, Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence, IAAI 2024,

Fourteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2014, February 20-27, 2024, Vancouver, Canada. pp. 19849–19857. AAAI Press (2024). https://doi.org/10.1609/AAAI.V38I18.29960, accessed: 2026-06-25

25. Glorot, X., Bordes, A., Bengio, Y.: Deep sparse rectifier neural networks. In: Gordon, G., Dunson, D., Dudík, M. (eds.) Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics. Proceedings of Machine Learning Research, vol. 15, pp. 315–323. PMLR, Fort Lauderdale, FL, USA (2011), https://proceedings.mlr.press/v15/glorot11a.html, accessed: 2026-06-25
26. Goodfellow, I.J., Shlens, J., Szegedy, C.: Explaining and harnessing adversarial examples. In: Bengio, Y., LeCun, Y. (eds.) 3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings (2015)
27. Gretton, A., Borgwardt, K.M., Rasch, M.J., Schölkopf, B., Smola, A.J.: A kernel two-sample test. J. Mach. Learn. Res. **13**, 723–773 (2012), https://dl.acm.org/doi/10.5555/2503308.2188410, accessed: 2026-06-25
28. Guo, C., Pleiss, G., Sun, Y., Weinberger, K.Q.: On calibration of modern neural networks. In: Precup, D., Teh, Y.W. (eds.) Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017. Proceedings of Machine Learning Research, vol. 70, pp. 1321–1330. PMLR (2017), http://proceedings.mlr.press/v70/guo17a.html, accessed: 2026-06-25
29. Gutbrod, M., Rauber, D., Nunes, D.W., Palm, C.: Openmibood: Open medical imaging benchmarks for out-of-distribution detection. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2025, Nashville, TN, USA, June 11-15, 2025. pp. 25874–25886. Computer Vision Foundation / IEEE (2025). https://doi.org/10.1109/CVPR52734.2025.02410, accessed: 2026-06-25
30. Hajj, H.A., Lamard, M., Conze, P., Roychowdhury, S., Hu, X., Marsalkaite, G., Zisimopoulos, O., Dedmari, M.A., Zhao, F., Prellberg, J., Sahu, M., Galdran, A., Araújo, T., Vo, D.M., Panda, C., Dahiya, N., Kondo, S., Bian, Z., Quellec, G.: CATARACTS: challenge on automatic tool annotation for cataract surgery. Medical Image Anal. **52**, 24–41 (2019). https://doi.org/10.1016/J.MEDIA.2018.11.008, accessed: 2026-06-25
31. Hein, M., Andriushchenko, M., Bitterwolf, J.: Why relu networks yield high-confidence predictions far away from the training data and how to mitigate the problem. In: IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019. pp. 41–50. Computer Vision Foundation / IEEE (2019). https://doi.org/10.1109/CVPR.2019.00013, accessed: 2026-06-25
32. Hendrycks, D., Basart, S., Mazeika, M., Mostajabi, M., Steinhardt, J., Song, D.: A benchmark for anomaly segmentation. ArXiv preprint **abs/1911.11132** (2019), https://arxiv.org/abs/1911.11132, accessed: 2026-06-25
33. Hendrycks, D., Basart, S., Mazeika, M., Zou, A., Kwon, J., Mostajabi, M., Steinhardt, J., Song, D.: Scaling out-of-distribution detection for real-world settings. In: Chaudhuri, K., Jegelka, S., Song, L., Szepesvári, C., Niu, G., Sabato, S. (eds.) International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA. Proceedings of Machine Learning Research, vol. 162, pp. 8759–8773. PMLR (2022), https://proceedings.mlr.press/v162/hendrycks22a.html, accessed: 2026-06-25

34. Hendrycks, D., Basart, S., Mu, N., Kadavath, S., Wang, F., Dorundo, E., Desai, R., Zhu, T., Parajuli, S., Guo, M., Song, D., Steinhardt, J., Gilmer, J.: The many faces of robustness: A critical analysis of out-of-distribution generalization. In: 2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021. pp. 8320–8329. IEEE (2021). https://doi.org/10.1109/ICCV48922.2021.00823, accessed: 2026-06-25
35. Hendrycks, D., Dietterich, T.G.: Benchmarking neural network robustness to common corruptions and perturbations. In: 7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019. OpenReview.net (2019), https://openreview.net/forum?id=HJz6tiCqYm, accessed: 2026-06-25
36. Hendrycks, D., Gimpel, K.: A baseline for detecting misclassified and out-of-distribution examples in neural networks. In: 5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings. OpenReview.net (2017), https://openreview.net/forum?id=Hkg4TI9xl, accessed: 2026-06-25
37. Hendrycks, D., Mazeika, M., Dietterich, T.G.: Deep anomaly detection with outlier exposure. In: 7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019. OpenReview.net (2019), https://openreview.net/forum?id=HyxCxhRcY7, accessed: 2026-06-25
38. Horn, G.V., Aodha, O.M., Song, Y., Cui, Y., Sun, C., Shepard, A., Adam, H., Perona, P., Belongie, S.J.: The inaturalist species classification and detection dataset. In: 2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018. pp. 8769–8778. IEEE Computer Society (2018). https://doi.org/10.1109/CVPR.2018.00914, accessed: 2026-06-25
39. Hotelling, H.: Analysis of a complex of statistical variables into principal components. Journal of educational psychology **24**(6), 417 (1933)
40. Humblot-Renaux, G., Franchi, G., Escalera, S., Moeslund, T.B.: Cooked: Ensemble-based ood detection in the era of zero-shot clip. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops. pp. 261–271 (2025)
41. Jha, D., Smedsrud, P.H., Riegler, M.A., Halvorsen, P., De Lange, T., Johansen, D., Johansen, H.D.: Kvasir-seg: A segmented polyp dataset. In: MultiMedia modeling: 26th international conference, MMM 2020, Daejeon, South Korea, January 5–8, 2020, proceedings, part II 26. pp. 451–462. Springer (2020)
42. Kamoi, R., Kobayashi, K.: Why is the mahalanobis distance effective for anomaly detection? ArXiv preprint **abs/2003.00402** (2020), https://arxiv.org/abs/2003.00402, accessed: 2026-06-25
43. Kavur, A.E., Gezer, N.S., Baris, M., Aslan, S., Conze, P., Groza, V., Pham, D.D., Chatterjee, S., Ernst, P., Özkan, S., Baydar, B., Lachinov, D.A., Han, S., Pauli, J., Isensee, F., Perkonigg, M., Sathish, R., Rajan, R., Sheet, D., Dovletov, G., Speck, O., Nürnberger, A., Maier-Hein, K.H., Akar, G.B., Ünal, G.B., Dicle, O., Selver, M.A.: CHAOS challenge - combined (CT-MR) healthy abdominal organ segmentation. Medical Image Anal. **69**, 101950 (2021). https://doi.org/10.1016/J.MEDIA.2020.101950, accessed: 2026-06-25
44. Kavur, A.E., Selver, M.A., Dicle, O., Barış, M., Gezer, N.S.: CHAOS - Combined (CT-MR) Healthy Abdominal Organ Segmentation Challenge Data (2019)
45. Krizhevsky, A., Hinton, G.: Learning multiple layers of features from tiny images. Tech. Rep. 0, University of Toronto, Toronto, Ontario (2009)

46. Kullback, S., Leibler, R.A.: On information and sufficiency. The annals of mathematical statistics **22**(1), 79–86 (1951)
47. LaMontagne, P.J., Benzinger, T.L., Morris, J.C., Keefe, S., Hornbeck, R., Xiong, C., Grant, E., Hassenstab, J., Moulder, K., Vlassenko, A.G., et al.: Oasis-3: longitudinal neuroimaging, clinical, and cognitive dataset for normal aging and alzheimer disease. medrxiv pp. 2019–12 (2019)
48. Le, Y., Yang, X.: Tiny imagenet visual recognition challenge. CS 231N **7**(7), 3 (2015)
49. Lee, K., Lee, K., Lee, H., Shin, J.: A simple unified framework for detecting out-of-distribution samples and adversarial attacks. In: Bengio, S., Wallach, H.M., Larochelle, H., Grauman, K., Cesa-Bianchi, N., Garnett, R. (eds.) Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, December 3-8, 2018, Montréal, Canada. pp. 7167–7177 (2018), https://proceedings.neurips.cc/paper/2018/hash/abdeb6f575ac5c6676b747bca8d09cc2-Abstract.html, accessed: 2026-06-25
50. Liang, S., Li, Y., Srikant, R.: Enhancing the reliability of out-of-distribution image detection in neural networks. In: 6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings. OpenReview.net (2018), https://openreview.net/forum?id=H1VGkIxRZ, accessed: 2026-06-25
51. Liew, S.L., Lo, B.P., Donnelly, M.R., Zavaliangos-Petropulu, A., Jeong, J.N., Barisano, G., Hutton, A., Simon, J.P., Juliano, J.M., Suri, A., et al.: A large, curated, open-source stroke neuroimaging dataset to improve lesion segmentation algorithms. Scientific data **9**(1), 320 (2022)
52. Lin, J.: Divergence measures based on the shannon entropy. IEEE Transactions on Information theory **37**(1), 145–151 (2002)
53. Liu, L., Qin, Y.: Fast decision boundary based out-of-distribution detector. In: Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024. OpenReview.net (2024), https://openreview.net/forum?id=eQaOb4r6YC, accessed: 2026-06-25
54. Liu, L., Qin, Y.: Detecting out-of-distribution through the lens of neural collapse. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2025, Nashville, TN, USA, June 11-15, 2025. pp. 15424–15433. Computer Vision Foundation / IEEE (2025). https://doi.org/10.1109/CVPR52734.2025.01437, accessed: 2026-06-25
55. Liu, W., Wang, X., Owens, J.D., Li, Y.: Energy-based out-of-distribution detection. In: Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., Lin, H. (eds.) Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual (2020), https://proceedings.neurips.cc/paper/2020/hash/f5496252609c43eb8a3d147ab9b9c006-Abstract.html, accessed: 2026-06-25
56. Liu, X., Lochman, Y., Zach, C.: GEN: pushing the limits of softmax-based out-of-distribution detection. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023. pp. 23946–23955. IEEE (2023). https://doi.org/10.1109/CVPR52729.2023.02293, accessed: 2026-06-25
57. Liu, Y., Sangineto, E., Bi, W., Sebe, N., Lepri, B., Nadai, M.D.: Efficient training of visual transformers with small datasets. In: Ranzato, M., Beygelzimer, A., Dauphin, Y.N., Liang, P., Vaughan, J.W. (eds.) Advances in Neu-

ral Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual. pp. 23818–23830 (2021), `https://proceedings.neurips.cc/paper/2021/hash/c81e155d85dae5430a8cee6f2242e82c-Abstract.html`, accessed: 2026-06-25
58. Liu, Y., Tang, H., Zhang, H., Qin, J., Li, Z.: Ot-detector: Delving into optimal transport for zero-shot out-of-distribution detection. In: Kwok, J. (ed.) Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence, IJCAI-25. pp. 1648–1656. International Joint Conferences on Artificial Intelligence Organization (2025). `https://doi.org/10.24963/ijcai.2025/184`, accessed: 2026-06-25
59. Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B.: Swin transformer: Hierarchical vision transformer using shifted windows. In: 2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021. pp. 9992–10002. IEEE (2021). `https://doi.org/10.1109/ICCV48922.2021.00986`, accessed: 2026-06-25
60. Van der Maaten, L., Hinton, G.: Visualizing data using t-sne. Journal of machine learning research **9**(11) (2008)
61. McInnes, L., Healy, J., Melville, J.: Umap: Uniform manifold approximation and projection for dimension reduction. ArXiv preprint **abs/1802.03426** (2018), `https://arxiv.org/abs/1802.03426`, accessed: 2026-06-25
62. Menze, B.H., Jakab, A., Bauer, S., Kalpathy-Cramer, J., Farahani, K., Kirby, J.S., Burren, Y., Porz, N., Slotboom, J., Wiest, R., Lanczi, L., Gerstner, E.R., Weber, M., Arbel, T., Avants, B.B., Ayache, N., Buendia, P., Collins, D.L., Cordier, N., Corso, J.J., Criminisi, A., Das, T., Delingette, H., Demiralp, Ç., Durst, C.R., Dojat, M., Doyle, S., Festa, J., Forbes, F., Geremia, E., Glocker, B., Golland, P., Guo, X., Hamamci, A., Iftekharuddin, K.M., Jena, R., John, N.M., Konukoglu, E., Lashkari, D., Mariz, J.A., Meier, R., Pereira, S., Precup, D., Price, S.J., Raviv, T.R., Reza, S.M.S., Ryan, M.T., Sarikaya, D., Schwartz, L.H., Shin, H., Shotton, J., Silva, C.A., Sousa, N.J., Subbanna, N.K., Székely, G., Taylor, T.J., Thomas, O.M., Tustison, N.J., Ünal, G.B., Vasseur, F., Wintermark, M., Ye, D.H., Zhao, L., Zhao, B., Zikic, D., Prastawa, M., Reyes, M., Leemput, K.V.: The multimodal brain tumor image segmentation benchmark (BRATS). IEEE Trans. Medical Imaging **34**(10), 1993–2024 (2015). `https://doi.org/10.1109/TMI.2014.2377694`, accessed: 2026-06-25
63. Muandet, K., Fukumizu, K., Sriperumbudur, B., Schölkopf, B.: Kernel mean embedding of distributions: A review and beyond. Foundations and Trends in Machine Learning **10**(1-2), 1–141 (2017). `https://doi.org/10.1561/2200000060`, accessed: 2026-06-25
64. Müller, M., Hein, M.: Mahalanobis++: Improving OOD Detection via Feature Normalization. In: Forty-second International Conference on Machine Learning (2025)
65. Netzer, Y., Wang, T., Coates, A., Bissacco, A., Wu, B., Ng, A.Y., et al.: Reading digits in natural images with unsupervised feature learning. In: NIPS workshop on deep learning and unsupervised feature learning. vol. 2011, p. 7. Granada (2011)
66. Pogorelov, K., Randel, K.R., Griwodz, C., Eskeland, S.L., de Lange, T., Johansen, D., Spampinato, C., Dang-Nguyen, D.T., Lux, M., Schmidt, P.T., et al.: Kvasir: A multi-class image dataset for computer aided gastrointestinal disease detection. In: Proceedings of the 8th ACM on Multimedia Systems Conference. pp. 164–169 (2017)

67. Rajasekaran, M., Sajol, M.S.I., Berglind, F., Mukhopadhyay, S., Das, K.: Combood: A semiparametric approach for detecting out-of-distribution data for image classification. In: Proceedings of the 2024 SIAM International Conference on Data Mining (SDM). pp. 643–651. SIAM (2024)
68. Recht, B., Roelofs, R., Schmidt, L., Shankar, V.: Do imagenet classifiers generalize to imagenet? In: Chaudhuri, K., Salakhutdinov, R. (eds.) Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA. Proceedings of Machine Learning Research, vol. 97, pp. 5389–5400. PMLR (2019), http://proceedings.mlr.press/v97/recht19a.html, accessed: 2026-06-25
69. Regmi, S.: Adascale: Adaptive scaling for ood detection. ArXiv preprint **abs/2503.08023** (2025), https://arxiv.org/abs/2503.08023, accessed: 2026-06-25
70. Ren, J., Liu, P.J., Fertig, E., Snoek, J., Poplin, R., DePristo, M.A., Dillon, J.V., Lakshminarayanan, B.: Likelihood ratios for out-of-distribution detection. In: Wallach, H.M., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E.B., Garnett, R. (eds.) Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada. pp. 14680–14691 (2019), https://proceedings.neurips.cc/paper/2019/hash/1e79596878b2320cac26dd792a6c51c9-Abstract.html, accessed: 2026-06-25
71. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022. pp. 10674–10685. IEEE (2022). https://doi.org/10.1109/CVPR52688.2022.01042, accessed: 2026-06-25
72. Rueckert, T., Maerkl, R., Rauber, D., Klausmann, L., Gutbrod, M., Rueckert, D., Feussner, H., Wilhelm, D., Palm, C.: Video Dataset for Surgical Phase, Keypoint, and Instrument Recognition in Laparoscopic Surgery (PhaKIR) (2025), https://arxiv.org/abs/2511.06549, accessed: 2026-06-25
73. Rueckert, T., Rauber, D., Maerkl, R., Klausmann, L., Yildiran, S.R., Gutbrod, M., Nunes, D.W., Moreno, A.F., Luengo, I., Stoyanov, D., Toussaint, N., Cho, E., Kim, H.B., Choo, O.S., Kim, K.Y., Kim, S.T., Arantes, G., Song, K., Zhu, J., Xiong, J., Lin, T., Kikuchi, S., Matsuzaki, H., Kouno, A., Manesco, J.R.R., Papa, J.P., Choi, T., Jeong, T.K., Park, J., Alabi, O., Wei, M., Vercauteren, T., Wu, R., Xu, M., Wang, A., Bai, L., Ren, H., Yamlahi, A., Hennighausen, J., Maier-Hein, L., Kondo, S., Kasai, S., Hirasawa, K., Yang, S., Wang, Y., Chen, H., Rodríguez, S., Aparicio, N., Manrique, L., Lyons, J.C., Hosie, O., Ayobi, N., Arbeláez, P., Li, Y., Khalil, Y.A., Nasirihaghighi, S., Speidel, S., Rueckert, D., Feussner, H., Wilhelm, D., Palm, C.: Comparative validation of surgical phase recognition, instrument keypoint estimation, and instrument instance segmentation in endoscopy: Results of the PhaKIR 2024 challenge. Medical Image Analysis **109**, 103945 (2026). https://doi.org/10.1016/j.media.2026.103945, https://doi.org/10.1016/j.media.2026.103945
74. Rueckert, T., Rieder, M., Rauber, D., Xiao, M., Humolli, E., Feussner, H., Wilhelm, D., Palm, C.: Augmenting instrument segmentation in video sequences of minimally invasive surgery by synthetic smoky frames. In: International Journal of Computer Assisted Radiology and Surgery. vol. 18, pp. S54 – S56 (2023)

75. Saikia, A.R., Bora, K., Mahanta, L.B., Das, A.K.: Comparative assessment of cnn architectures for classification of breast fnac images. Tissue and Cell **57**, 8–14 (2019)
76. Sastry, C.S., Oore, S.: Detecting out-of-distribution examples with gram matrices. In: Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event. Proceedings of Machine Learning Research, vol. 119, pp. 8491–8501. PMLR (2020), `http://proceedings.mlr.press/v119/sastry20a.html`, accessed: 2026-06-25
77. Sehwag, V., Chiang, M., Mittal, P.: SSD: A unified framework for self-supervised outlier detection. In: 9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021. OpenReview.net (2021), `https://openreview.net/forum?id=v5gjXpmR8J`, accessed: 2026-06-25
78. Setio, A.A.A., Traverso, A., de Bel, T., Berens, M.S.N., van den Bogaard, C., Cerello, P., Chen, H., Dou, Q., Fantacci, M.E., Geurts, B., van der Gugten, R., Heng, P., Jansen, B., de Kaste, M.M.J., Kotov, V., Lin, J.Y., Manders, J.T.M.C., Sóñora-Mengana, A., García-Naranjo, J.C., Papavasileiou, E., Prokop, M.: Validation, comparison, and combination of algorithms for automatic detection of pulmonary nodules in computed tomography images: The LUNA16 challenge. Medical Image Anal. **42**, 1–13 (2017). `https://doi.org/10.1016/J.MEDIA.2017.06.015`, accessed: 2026-06-25
79. Sevillano-García, I., Luengo, J., Herrera, F.: Stood-x: Explainable out-of-distribution detection via nonparametric statistical testing on large-scale datasets. Pattern Recognition p. 113254 (2026)
80. Stewart, G.W.: The efficient generation of random orthogonal matrices with an application to condition estimators. SIAM Journal on Numerical Analysis **17**(3), 403–409 (1980). `https://doi.org/10.1137/0717034`, accessed: 2026-06-25
81. Sun, Y., Guo, C., Li, Y.: React: Out-of-distribution detection with rectified activations. In: Ranzato, M., Beygelzimer, A., Dauphin, Y.N., Liang, P., Vaughan, J.W. (eds.) Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual. pp. 144–157 (2021), `https://proceedings.neurips.cc/paper/2021/hash/01894d6f048493d2cacde3c579c315a3-Abstract.html`, accessed: 2026-06-25
82. Sun, Y., Li, Y.: DICE: leveraging sparsification for out-of-distribution detection. In: Avidan, S., Brostow, G.J., Cissé, M., Farinella, G.M., Hassner, T. (eds.) Computer Vision - ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23-27, 2022, Proceedings, Part XXIV. Lecture Notes in Computer Science, vol. 13684, pp. 691–708. Springer (2022). `https://doi.org/10.1007/978-3-031-20053-3_40`, accessed: 2026-06-25
83. Sun, Y., Ming, Y., Zhu, X., Li, Y.: Out-of-distribution detection with deep nearest neighbors. In: Chaudhuri, K., Jegelka, S., Song, L., Szepesvári, C., Niu, G., Sabato, S. (eds.) International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA. Proceedings of Machine Learning Research, vol. 162, pp. 20827–20840. PMLR (2022), `https://proceedings.mlr.press/v162/sun22d.html`, accessed: 2026-06-25
84. Tack, J., Mo, S., Jeong, J., Shin, J.: CSI: novelty detection via contrastive learning on distributionally shifted instances. In: Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., Lin, H. (eds.) Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual (2020), `https://proceedings.neurips.`

cc/paper/2020/hash/8965f76632d7672e7d3cf29c87ecaa0c-Abstract.html, accessed: 2026-06-25

85. Tobon-Gomez, C., Geers, A.J., Peters, J., Weese, J., Pinto, K., Karim, R., Ammar, M., Daoudi, A., Margeta, J., Sandoval, Z.L., Stender, B., Zheng, Y., Zuluaga, M.A., Betancur, J., Ayache, N., Chikh, M.A., Dillenseger, J., Kelm, B.M., Mahmoudi, S., Ourselin, S., Schlaefer, A., Schaeffter, T., Razavi, R., Rhode, K.S.: Benchmark for algorithms segmenting the left atrium from 3d CT and MRI datasets. IEEE Trans. Medical Imaging **34**(7), 1460–1473 (2015). https://doi.org/10.1109/TMI.2015.2398818, accessed: 2026-06-25
86. Twinanda, A.P., Shehata, S., Mutter, D., Marescaux, J., de Mathelin, M., Padoy, N.: Endonet: A deep architecture for recognition tasks on laparoscopic videos. IEEE Trans. Medical Imaging **36**(1), 86–97 (2017). https://doi.org/10.1109/TMI.2016.2593957, accessed: 2026-06-25
87. Vaze, S., Han, K., Vedaldi, A., Zisserman, A.: Open-set recognition: A good closed-set classifier is all you need. In: The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022. OpenReview.net (2022), https://openreview.net/forum?id=5hLP5JY9S2d, accessed: 2026-06-25
88. Wang, H., Li, Z., Feng, L., Zhang, W.: Vim: Out-of-distribution with virtual-logit matching. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022. pp. 4911–4920. IEEE (2022). https://doi.org/10.1109/CVPR52688.2022.00487, accessed: 2026-06-25
89. Wang, Y., Li, B., Che, T., Zhou, K., Liu, Z., Li, D.: Energy-based open-world uncertainty modeling for confidence calibration. In: 2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021. pp. 9282–9291. IEEE (2021). https://doi.org/10.1109/ICCV48922.2021.00917, accessed: 2026-06-25
90. Wang, Y., Sun, W., Jin, J., Kong, Z., Yue, X.: Wood: Wasserstein-based out-of-distribution detection. IEEE Transactions on Pattern Analysis and Machine Intelligence **46**(2), 944–956 (2023)
91. Woodland, M., Patel, N., Castelo, A., Al Taie, M., Eltaher, M., Yung, J.P., Netherton, T.J., Calderone, T.L., Sanchez, J.I., Cleere, D.W., et al.: Dimensionality reduction and nearest neighbors for improving out-of-distribution detection in medical image segmentation. Machine Learning for Biomedical Imaging **2**(UNSURE2023 special issue), 2006–2052 (2024)
92. Xu, K., Chen, R., Franchi, G., Yao, A.: Scaling for training time and post-hoc out-of-distribution detection enhancement. In: The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024. OpenReview.net (2024), https://openreview.net/forum?id=RDSTjtnqCg, accessed: 2026-06-25
93. 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.: Openood: Benchmarking generalized out-of-distribution detection. In: Koyejo, S., Mohamed, S., Agarwal, A., Belgrave, D., Cho, K., Oh, A. (eds.) Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022 (2022), http://papers.nips.cc/paper_files/paper/2022/hash/d201587e3a84fc4761eadc743e9b3f35-Abstract-Datasets_and_Benchmarks.html, accessed: 2026-06-25

94. Yang, J., Zhou, K., Li, Y., Liu, Z.: Generalized out-of-distribution detection: A survey. International Journal of Computer Vision pp. 1–28 (2024)
95. Yang, T., Huang, Y., Xie, Y., Liu, J., Wang, S.: Mixood: Improving out-of-distribution detection with enhanced data mixup. ACM Trans. Multimedia Comput. Commun. Appl. **19**(5) (2023)
96. Yuan, Y., He, R., Dong, Y., Han, Z., Yin, Y.: Discriminability-driven channel selection for out-of-distribution detection. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024. pp. 26171–26180. IEEE (2024). `https://doi.org/10.1109/CVPR52733.2024.02473`, accessed: 2026-06-25
97. Zhang, H., Cissé, M., Dauphin, Y.N., Lopez-Paz, D.: mixup: Beyond empirical risk minimization. In: 6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings. OpenReview.net (2018), `https://openreview.net/forum?id=r1Ddp1-Rb`, accessed: 2026-06-25
98. Zhang, J., Yang, J., Wang, P., Wang, H., Lin, Y., Zhang, H., Sun, Y., Du, X., Zhou, K., Zhang, W., Li, Y., Liu, Z., Chen, Y., Li, H.: OpenOOD v1.5: Enhanced benchmark for out-of-distribution detection. In: NeurIPS 2023 Workshop on Distribution Shifts: New Frontiers with Foundation Models (2024)
99. Zhang, J., Fu, Q., Chen, X., Du, L., Li, Z., Wang, G., Liu, X., Han, S., Zhang, D.: Out-of-distribution detection based on in-distribution data patterns memorization with modern hopfield energy. In: The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023. OpenReview.net (2023), `https://openreview.net/pdf?id=KkazG4lgKL`, accessed: 2026-06-25
100. Zhou, B., Lapedriza, A., Khosla, A., Oliva, A., Torralba, A.: Places: A 10 million image database for scene recognition. IEEE transactions on pattern analysis and machine intelligence **40**(6), 1452–1464 (2017)
101. Zöngür, B., Hesse, R., Roth, S.: Activation subspaces for out-of-distribution detection. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 3509–3519 (2025)