A noisy elephant in the room: Is your out-of-distribution detector robust to label noise?

Galadrielle Humblot-Renaux[1] Sergio Escalera[1,2] Thomas B. Moeslund[1]
[1]Visual Analysis and Perception lab, Aalborg University, Denmark
[2]Department of Mathematics and Informatics, University of Barcelona and Computer Vision Center, Spain
gegeh@create.aau.dk sescalera@ub.edu tbm@create.aau.dk

Abstract

The ability to detect unfamiliar or unexpected images is essential for safe deployment of computer vision systems. In the context of classification, the task of detecting images outside of a model's training domain is known as out-of-distribution (OOD) detection. While there has been a growing research interest in developing post-hoc OOD detection methods, there has been comparably little discussion around how these methods perform when the underlying classifier is not trained on a clean, carefully curated dataset. In this work, we take a closer look at 20 state-of-the-art OOD detection methods in the (more realistic) scenario where the labels used to train the underlying classifier are unreliable (e.g. crowd-sourced or web-scraped labels). Extensive experiments across different datasets, noise types & levels, architectures and checkpointing strategies provide insights into the effect of class label noise on OOD detection, and show that poor separation between incorrectly classified ID samples vs. OOD samples is an overlooked yet important limitation of existing methods. Code: https://github.com/glhr/ood-labelnoise

1. Introduction

In many real-world applications where deep neural networks are deployed "in the wild", it is desirable to have models that not only correctly classify samples drawn from the distribution of labeled data but also flag unexpected inputs as out-of-distribution (OOD). This has motivated the development of a wide range of OOD detection methods and benchmarks for computer vision [47, 66]. In particular, *post-hoc* OOD detection methods have shown wide appeal: compared to training-based methods, post-hoc OOD detectors can be applied on top of existing image classifiers without the need for re-training, have little to no architecture constraints, do not compromise classification performance, and achieve strong performance in large-scale settings [67].

Existing OOD benchmarks place significant emphasis on carefully designing the selection of OOD datasets used for evaluation [4, 14, 60, 67]. In contrast, the role of the in-distribution (ID) dataset used for training the underlying classifier is seldom discussed. Among the most popular choices of ID dataset are MNIST, CIFAR10, CIFAR100 and ImageNet [30, 67] - all of which have been carefully curated and reliably annotated. Yet, in practice, the collection of labelled datasets involves a trade-off between acquisition time/cost and annotation quality - human inattention, mis-clicking, limited expertise, crowd-sourcing, automated annotation, and other cost-saving measures inevitably introduce labelling errors [54]. Besides, some images are inherently ambiguous to label even for the most knowledgeable and careful of annotators [28]. Considering how pervasive the problem of label noise is in real-world image classification datasets, its effect on OOD detection is crucial to study.

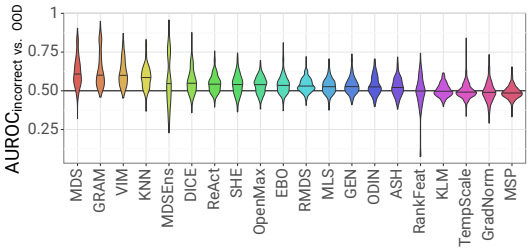

Figure 1. Can state-of-the-art OOD detectors tell incorrectly classified ID images apart from OOD inputs? Not really. Here we compare their performance across 396 trained classifiers.

To address this gap, we systematically analyse the label noise robustness of a wide range of OOD detectors, ranging from the widely adopted Maximum Softmax Probability (MSP) baseline [14, 19], to distance-based methods operating in feature space [32, 57], to more recent, complex methods such as SHE [70] and ASH [10]. In particular:

1. We present the first study of post-hoc OOD detection in the presence of noisy classification labels, examining the performance of 20 state-of-the-art methods under different types and levels of label noise in the training data. Our study includes multiple classification architectures

and datasets, ranging from the beloved CIFAR10 to the more difficult Clothing1M, and shows that even at a low noise rate, the label noise setting poses an interesting challenge for many methods.
2. We revisit the notion that OOD detection performance correlates with ID accuracy [14, 60], examining when and why this relation holds. Robustness to inaccurate classification requires that OOD detectors effectively separate mistakes on ID data from OOD samples - yet most existing methods confound the two (Figure 1).
3. Our analysis includes key takeaways and recommendations for future evaluation and development of OOD detection methods considering an unreliable label setting.

2. Problem set-up

In this work, we tackle the question: *what happens when post-hoc OOD detectors are applied on top of a classifier trained with unreliable labels - a common setting in practice?* We introduce the main relevant concepts below.

Classifier We study OOD detection in the context of supervised image classification, where a discriminative model $h : \mathcal{X} \rightarrow \mathcal{Y}$ is trained on a dataset of N labelled examples $D_{train} = \{(x_i, y_i)\}_{i=1}^{N} \in \mathcal{X} \times \mathcal{Y}$, where each x is an input image and each y is the corresponding class from the label space $\mathcal{Y}$. A common choice would for example be CIFAR10 [31]. $P_{train}(X, Y)$ defines the underlying training data distribution. The classifier is evaluated on a test set D_{test} drawn from the same distribution $P_{test}(X, Y) = P_{train}(X, Y)$.

OOD detector Post-hoc OOD detection equips the trained classifier h with a scoring function $o : X \rightarrow \mathbb{R}$ aiming to distinguish *usual* examples drawn from $P_{test}(X)$ (ID samples) and *anomalous* (OOD) examples drawn from a disjoint, held-out distribution $P_{out}(X)$. In practice, a collection of auxiliary datasets with minimal semantic overlap (e.g. CIFAR10 $\rightarrow$ SVHN [39]) is commonly used for evaluation [67]. Ideally, the score assigned to ID samples should be consistently lower (or higher) than for OOD samples, such that anomalous inputs can easily be flagged.

Label noise We consider a noisy label setting, where the classifier h does not have access to the true target values y_i during training, but rather learns from a noisy dataset $D_{noisy} = \{(x_i, \hat{y}_i)\}_{i=1}^{N}$, where the target labels are corrupted: $\exists\, i$ such that $\hat{y}_i \neq y_i$. In this work, we consider only closed-set label noise, where $D_{noisy} \in \mathcal{X} \times \mathcal{Y}$ (that is, the noisy labels lie in the same label space as the true labels [46]). The noise level ϵ is given by $P(y \neq \hat{y})$, the probability that an observed label is incorrect. Common models for studying and simulating label noise are (we refer to [13] for a detailed taxonomy):

1. Noisy Completely at Random (NCAR) or uniform label noise: labels are flipped at a constant rate ϵ, regardless of class or image.
2. Noisy at Random (NAR) or class-conditional label noise: a constant noise rate across all images of the same class, but different classes may have different noise rates.
3. Noisy Not at Random (NNAR) or instance-dependent label noise: noisy labels are jointly determined by the true class and the associated image.

In practice, *real* (as opposed to synthetically generated) label noise occurring from an imperfect annotation pipeline follows complex patterns, and is thus best represented by an instance-dependent model: some *classes* are more likely to be mislabeled than others, and so are some *images* (e.g. ambiguous or rare samples) [50, 64].

3. Related work

Studying label noise The effect of unreliable labels on supervised learning is a well-studied problem in deep learning [51] and computer vision [1, 28], as errors or inconsistencies are a natural part of label collection in many real applications. Though increased dataset size can help [45], noisy labels degrade classification performance, especially in the later stages of training where over-parameterized models are prone to memorizing them [36, 69]. The precise effects of label noise have been shown to depend on the noise model and distribution [41]. A recent CIFAR classification benchmark suggests that models trained on real, instance-dependent noisy labels are significantly more prone to memorization than those trained on synthetic class-conditional labels with the same overall noise rate [64]. We therefore consider real noisy labels in our benchmark (stemming from human annotation error), which we compare to two sets of synthetic noisy labels (uniform and class-conditional). We also compare the effect of validation-based early stopping vs. converging on the training set.

Effect of label noise on reliability Existing studies of label noise are largely focused on classification accuracy, and few works address the other side of the coin: *reliability*. We look at reliability from the angle of OOD detection performance - to the best of our knowledge, there is currently no comparable study of OOD detection under a noisy label setting. Most closely related to our work are perhaps the experiments in [42] and the analysis in [40]. [42] evaluates the effect of synthetic uniform label noise on MC-dropout and deep ensembles' uncertainty estimates, showing a significant degradation in OOD detection performance with increasing noise levels - in comparison, we study post-hoc OOD detection (with a wider variety of architectures, datasets, and methods) and consider real noisy datasets. [40] studies label noise robustness in terms of model calibration, showing that early stopping, while beneficial in terms of accuracy, offers no reliability guarantees.

Benchmarking OOD detection robustness Previous works have investigated the limits of state-of-the-art OOD detection methods in various challenging settings, such as

semantic similarity between ID vs. OOD classes [12, 14, 60], fine-grained ID labels [23], large-scale datasets [22] and adversarial attacks [49, 68]. In contrast, we focus on robustness to *degraded classification performance on the ID dataset* due to noisy labels, which to the best of our knowledge has comparably received little attention.

Relation between ID classification and OOD detection performance In the standard clean label setting, a strong relationship between ID classification and OOD detection performance has been observed in prior work. [60] studies the relation between closed-set (ID) classification and open-set recognition performance (AUROC), and finds open-set recognition performance to be highly correlated with classification accuracy. [14] observes a similar trend for out-of-distribution detection performance across a large variety of pre-trained deep learning architectures, using the MSP as OOD score. Both works only consider clean training datasets, and a small subset of methods. We study the extent to which this relation holds across a wider range of OOD detection methods and noisy datasets, and provide a very simple explanation for why some methods like MSP reach such a high correlation.

4. OOD detection methods

We evaluate 20 post-hoc OOD detection methods from the OpenOOD benchmark [67] - currently the most comprehensive open-source benchmark available. Here we present and broadly categorize these methods based on how their scoring function is designed.

Softmax-based OOD detection revolves around the idea that ID samples are associated with higher-confidence, lower-entropy predictions than OOD samples. The baseline Maximum Softmax Probability (MSP) [19] simply takes the Softmax "confidence" of the predicted class as OOD score. While MSP implicitly assumes a Softmax temperature of 1, TempScaling [15] treats the temperature as a hyper-parameter, softening or sharpening the Softmax probabilities (essentially modulating categorical entropy), with the aim of improving calibration. ODIN [35] combines temperature scaling with input perturbation - "pushing" the input image a little in the direction that increases the MSP. In contrast, the Generalized ENtropy (GEN) score considers the full predictive distribution and captures how much it deviates from a one-hot distribution.

Logit-based OOD detection bypasses the squashing effect of Softmax normalization. The Maximum Logit Score (MLS) [21] directly takes the logit of the predicted class. In a similar vein, energy-based OOD detection (EBO) was first proposed in [37]: a score is derived by applying the LogSumExp function to the logits - essentially a smooth version of the MLS, with an additional temperature parameter. Several post-hoc methods using an energy score have since followed suit, proposing various modifications to the network [55] or features [10, 53, 56] before extracting an energy score: REACT [56] clips activations at an upper bound, RankFeat [53] subtracts the rank-1 matrix from activations, DICE [55] applies weight sparsification such that only the strongest contributors remain, and ASH [10] sparsifies activations based on a pruning percentile.

Distance-based OOD detection aims to capture how much a test sample deviates from the ID dataset. The Mahalanobis distance score (MDS) [32] method fits a Gaussian distribution to the features of each class in the ID dataset; at test-time, the OOD score is taken as the distance to the closest class. The same authors also proposed MDSEnsemble [32], which computes an MDS score not just from the features extracted before the final layer, but also at earlier points in the network, and aggregates them. Alternatively, the Relative Mahalanobis distance score (RMDS) [44] was proposed as a simple fix to MDS, which additionally fits a class-independent Gaussian to the entire ID dataset to compute a background score which is subtracted from the class-specific MDS score. Among other distance-based methods which rely on class-wise statistics, KLMatching (KLM) [21] takes the smallest KL Divergence between a test sample's Softmax probabilities and the mean Softmax probability vector for each ID class. OpenMax [3] operates in logit space, fitting a class-wise Weibull distribution to the distances of ID samples from the mean logit vector. Simplified Hopfield Energy (SHE) [70] computes the inner product between a test sample's features and the mean ID feature of the predicted class. GRAM [48] computes the Gram matrices of intermediate feature representations throughout the network, comparing them with the range of values observed for each class in the ID data. In contrast, deep k-nearest neighbor (KNN) [57] proposes a simple approach with no distributional assumptions - computing its score as the Euclidean distance to the closest samples from the ID set, regardless of class. Lastly, Virtual-logit Matching (VIM) [62] combines a logit energy score with a class-agnostic term capturing how features deviate from a principal subspace defined by the training set.

Gradient-based OOD detection: GradNorm [24] is the only method in OpenOOD which directly derives its score from the gradient space, claiming that gradient magnitude is higher for ID inputs. The KL divergence between predicted Softmax probabilities and a uniform target is backpropagated to obtain gradients w.r.t the last layer parameters, followed by an L_1 norm to obtain the magnitude.

5. Experiments

We summarize our experimental set-up below, and refer to the supplementary for further details.

ID Datasets We select popular image classification datasets from the label noise literature featuring real noisy labels alongside clean reference labels, spanning different

	ID dataset	classes	# images (train/val/test)	resolution	noise rate
[31]	**CIFAR-10**	10	50,000/1,000/9,000	32x32	0%
[64]	CIFAR-10N-Agg				9.01%
	CIFAR-10N-Rand1				17.23%
	CIFAR-10N-Worst				40.21%
[31]	**CIFAR-100-Fine**	100	50,000/1,000/9,000	32x32	0%
[64]	CIFAR-100N-Fine				40.20%
[31]	**CIFAR-100-Coarse**	20	50,000/1,000/9,000	32x32	0%
[64]	CIFAR-100N-Coarse				26.40%
[65]	**Clothing1M**	14	24,637/7,465/5,395	256x256	0%
	Clothing1M-Noisy				38.26%

Table 1. Dataset overview. Clean ones are shown in bold. The training set (clean or noisy labels) is used to train the classifier; the validation set (clean labels) is used for early stopping; the test set (clean labels) is used for evaluating classification and OOD detection performance. We always use clean labels for evaluation.

input sizes, number of classes, and sources & levels of label noise - see Table 1. The recently released CIFAR-N dataset [64] provides noisy re-annotations of CIFAR-10 and CIFAR-100 collected via crowd-sourcing: each image was annotated by 3 people, and different noisy label sets were created for different label selection methods (majority voting, random selection, or worst label selection). Note that CIFAR-100-Fine and CIFAR-100-Coarse contain the same set of images - only the class definitions and labels differ. Clothing1M [65] is a large-scale dataset collected by scraping shopping websites. Although the raw Clothing1M contains over a million images, we consider the smaller subset of images for which there is both a noisy and clean label.

Synthetic noise For each real noisy label set, using the corresponding clean labels, we additionally create 2 synthetic counterparts with the same overall noise rate: one following a uniform (NCAR, class-independent) label noise model, and the other following a class-conditional label noise model with the exact same noise transition matrix as the real noise. We name these synthetic variants SU (Synthetic Uniform noise) and SCC (Synthetic Class-Conditional noise) - for example, from CIFAR-10N-Agg we create 2 synthetic versions, SU and SCC.

OOD Datasets For fair comparison, we use the same selection of OOD datasets for all models - the OOD datasets are therefore chosen such that there is minimal semantic overlap with any of the ID datasets. We include MNIST [9], SVHN [39], Textures [6] as they are commonly used as examples of "far"-OOD [67] (very different appearance and semantics than the ID dataset). As examples of more natural images, we also include EuroSAT-RGB [18], Food-101 [5], a sub-set of the Stanford Online Products [52], and a 12-class sub-set of ImageNet. Since some methods require an OOD validation set for hyperparameter tuning, half of these classes are randomly selected and held-out for this purpose. The other 6 ImageNet classes, and the other OOD datasets make up the OOD test set.

Evaluation metrics We evaluate OOD detectors' ability to separate ID vs. OOD samples in terms of the Area Under the Receiver Operating Characteristic Curve (AUROC), where images from the ID test set (e.g. CIFAR10 test set) are considered positive samples, and those from the OOD test set (e.g. SVHN test set) as negatives. This is the most commonly reported metric in the literature [25], and we denote it as $\text{AUROC}_{\text{ID vs. OOD}}$. In addition, unlike previous works, we separately measure the $\text{AUROC}_{\text{correct vs. OOD}}$ (and $\text{AUROC}_{\text{incorrect vs. OOD}}$), where only correctly (or incorrectly) classified samples from the ID test set are considered - ideally, performance should be high on both metrics.

Architectures We include 3 architecture families: CNNs, MLPs and transformers. We select lightweight architectures which have shown competitive results when trained on small-scale datasets: ResNet18 [17], MLP-Mixer [58] and Compact Transformers [16]. Following the OpenOOD benchmark [67], we do not adopt any advanced training strategies besides standard data augmentation. For each training dataset, we repeat training with 3 random seeds, and save 2 model checkpoints: an *early* checkpoint (based on best validation accuracy) and the *last* checkpoint (after a pre-defined number of epochs has elapsed, allowing for convergence - differs per architecture).

Bird's eye view To summarize, we train 3 different classifier architectures on 22 datasets (4 clean, 6 with real label noise, 12 with synthetic label noise), with 3 random seeds and 2 checkpoints saved per model - adding up to 396 distinct classifiers. On top of each classifier, 20 different OOD detection methods are applied and evaluated on 7 OOD datasets. Throughout the paper, OOD detection performance is taken as the median across the 7 OOD datasets (see the supplementary for results and a discussion of the median vs. mean OOD detection performance).

Statistical significance tests When comparing pairs of methods or settings, we use the Almost Stochastic Order (ASO) test [7, 11] as implemented by Ulmer et al. [59]. This statistical test was specifically designed to compare deep learning models, making no distributional assumptions. We apply ASO with a significance level $\alpha = 0.05$ and report $\epsilon_{\text{min}}[\text{A>B}]$. If $\epsilon_{\text{min}}[\text{A>B}] \geq 0.5$ we cannot claim that method A is better than method B; the smaller ϵ_{min}, the more confident we can be that method A is superior.

6. Analysis

We explore the effect of label noise on OOD detection, starting with an overall view of performance trends in Section 6.1, then looking at OOD detection in relation to classification performance in Section 6.2, delving into what works (and what doesn't) in Section 6.3, and raising important considerations about how/whether to make use of a clean validation set in Section 6.4. Section 6.5 extends results to a more practical setting. More detailed analyses and additional supporting figures are in the supplementary.

training labels	CIFAR10										CIFAR100-Coarse				CIFAR100-Fine				Clothing1M			
		Agg			Rand1			Worst															
method	**clean**	**N**	**SCC**	**SU**	**N**	**SCC**	**SU**	**N**	**SCC**	**SU**	**clean**	**N**	**SCC**	**SU**	**clean**	**N**	**SCC**	**SU**	**clean**	**N**	**SCC**	**SU**
GRAM	**94.45**	**89.49**	**89.12**	**90.7**	88.82	**89.19**	90.52	88.6	88.73	**87.98**	82.07	80.05	**82.1**	79.2	82.93	76.31	80.24	82.64	**91.04**	**89.07**	**94.71**	**95.37**
MDS	96.07	87.93	92.4	92.97	92.37	89.25	**87.45**	**86.74**	**86.49**	89.2	80.07	78.89	82.84	80.12	80.1	74.96	74.48	73.49	87.12	**90.98**	**88.58**	**92.38**
VIM	**95.65**	**89.9**	**91.81**	**92.3**	88.75	**88.6**	84.49	**86.31**	**87.23**	**88.75**	84.29	76.61	80.04	78	81.37	75.31	73.24	**73.94**	**88.99**	83.09	87.17	90.14
MDSEns	92.57	83.89	83.62	83.79	81.8	83.06	80.36	82.95	84.02	84.11	79.25	78.31	77.41	73.6	84.85	77.47	78.43	79.85	95.36	95.44	95.78	95.69
KNN	93.63	90.07	88.75	90.14	87.74	86.66	85.11	86.3	83.73	84.14	**84.16**	74.4	**80.39**	75.82	83.29	75.67	76.48	71.35	85.32	85.5	84.59	80.87
RMDS	92.92	89.38	87.94	88.14	89.07	85.73	**87.04**	84.03	81.99	82.35	82.14	75.93	77.36	74.81	83.28	76	75.75	73.48	75.81	71.43	78.22	66.66
DICE	90	83.33	84.18	86.24	88.52	81	86.21	82.79	79.58	79.07	**82.79**	77.68	75.01	70.43	82.52	76.51	73.92	68.41	84.96	75.72	86.69	82.89
ReAct	90.91	87.32	86.63	82.16	89.74	82.96	81.97	84.5	78.41	80.11	**82.79**	73.09	73.62	70.38	**83.76**	73.57	73.71	67.55	82.57	73.1	80.22	76.58
GEN	91.86	85.99	85.44	82.08	**89.86**	82.89	80.84	83.75	81.81	80.57	82.69	73.25	71.47	70.99	81.34	73.4	73.1	67.11	83.91	73.57	79.79	76.78
EBO	91.31	84.87	85.73	81.62	**89.88**	81.93	77.88	83.04	81.38	77.4	82.74	72.99	70.93	67.85	81.41	73.65	73.01	67.42	85.19	76.32	85.31	76.64
SHE	89.6	87.81	84.33	86.48	86.63	83.16	83.04	83.24	80.06	78.98	80.42	71.8	80	70.11	80.38	69.63	69.56	66.68	82.29	78.07	78.4	77.73
ODIN	91.47	87.71	86.31	82.48	89.79	82.21	80.68	84.09	82.13	80.09	81.42	73.1	70.88	69.97	**83.59**	74.85	72.19	67.16	83.38	71.59	77.73	75.47
MLS	91.26	84.76	85.59	81.57	88.81	82.31	78	83.54	82.01	80.01	82.66	72.92	70.85	69.19	81.45	73.64	72.5	67.03	83.3	72.75	77.74	75.54
TempScale	91.67	85.76	85.04	82.25	85.07	78.1	79.78	82.85	80.51	80.13	81.63	71.67	69.94	69.35	80.75	72.66	71.28	66.98	79.82	68.77	86.28	74.45
ASH	88.33	84.75	82.37	81.66	82.29	76.47	72.48	85.27	78.26	75.48	82.78	71.19	73.21	68.43	82.74	74.13	70.95	67.48	81.53	74.89	76.63	75.74
OpenMax	90.5	86.12	83.46	82.26	83.05	82.87	79.12	80.39	75.68	77.41	81.14	76.69	72.65	69.17	80.13	72.82	75.6	67.66	71.74	69.23	72.55	74.36
MSP	91.34	85.19	84.87	82.41	85.13	82.21	80.68	82.48	79.4	80.09	80.51	70.42	68.88	69.97	79.85	70.92	69.61	66.92	77.57	66.02	72.44	73.89
KLM	90.84	83.7	82.15	81.69	80.13	81.88	80.64	74.79	75.93	76.83	79.37	70.2	69.24	67.81	79.58	71.52	69.89	67.25	77.26	65.49	65.22	62.01
GradNorm	86.22	79.55	77.1	77.84	81.66	79.88	76.96	84.11	72.96	72.88	69.29	66.27	73.17	67.44	70.95	65.65	71.26	63.29	79.71	75.32	73.93	77.52
RankFeat	81.83	83.53	75.86	73.12	78.29	75.07	79.14	82.57	77.25	75.75	73.15	64.6	69.54	65.44	68.77	76.07	70.24	67.64	69.16	75.85	69.93	73.37

Table 2. **Best-case** OOD detection performance ($\text{AUROC}_{\text{ID vs. OOD}}$ in %) per method (that is, after selecting the best architecture-seed-checkpoint combination for each training label set). N, SCC, and SU refer to the real and synthetic noisy label sets described in Section 5. The top-3 for each training dataset are highlighted in bold, and the top-1 is underlined. In red are scores $< 75\%$ and in orange scores between 75 and 80%. Rows are sorted based on the total performance across columns.

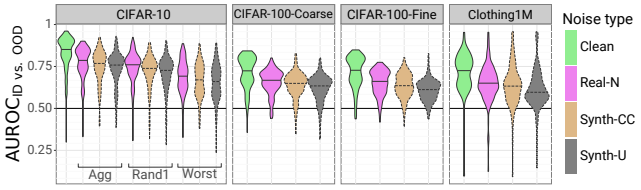

Figure 2. Distribution of OOD detection performance across methods & models when training the classifier on different label sets.

6.1. Where there's noise there's trouble

Figure 2 gives an overview of OOD detection performance for different training datasets and label noise settings. We see a clear drop in overall OOD detection performance when label noise is introduced in the training dataset, compared to training on a cleanly labelled dataset (green). Even with only 9% of incorrect CIFAR10 labels (CIFAR-Agg labels sets), the median $\text{AUROC}_{\text{ID vs. OOD}}$ across all models drops by over 5%. In Table 2, for each method we report the *best-case* OOD detection performance for a given training label set. While most methods are able to reach 80% $\text{AUROC}_{\text{ID vs. OOD}}$ with a classifier trained on clean labels, the number of competitive methods falls with increasing label noise, especially at noise rates $> 20\%$. GRAM, KNN, MDS, MDSEnsemble and VIM are the only methods able to reach 90+% AUROC on at least one of the noisy datasets. **Takeaway: Enter the elephant** Label noise in the classifier's training data makes it more difficult for post-hoc OOD detection methods to flag unfamiliar samples at test-time, even in small-scale settings like CIFAR10.

6.2. Does accuracy tell the whole story?

The most obvious effect of label noise in the training data is a decrease in classification performance on ID test data. At the same time, previous works have remarked a strong relation between classification performance and OOD detection for popular post-hoc methods like MSP [14] and MLS [60]. We dig deeper. When does this relation hold and why?

For which methods does this relation hold? In Figure 3, we quantify the relationship between ID accuracy and $\text{AUROC}_{\text{ID vs. OOD}}$ in terms of Spearman correlation ρ. We find that correlation varies widely across methods, being the strongest for MSP, and is generally weaker for those which operate earlier in the network. We also note that for all methods except KNN and RMDS, the label noise setting makes OOD detection performance *less predictable* - and so does early stopping (cf. Section 6.4). This points to the distribution of ID scores playing an important role in OOD detection performance.

When it does - why? We provide a simple observation which is lacking in prior work: methods whose OOD detection performance predictably degrades along-

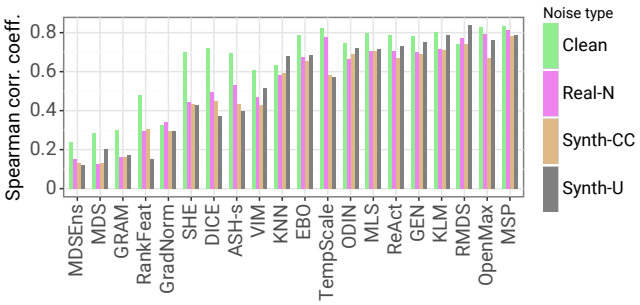

Figure 3. Does OOD detection performance ($\text{AUROC}_{\text{ID vs. OOD}}$) correlate with ID classification performance (accuracy)? We measure the rank correlation across different architectures, seeds, checkpoints, and datasets for different label sets. All results shown here are statistically significant ($p << 0.001$).

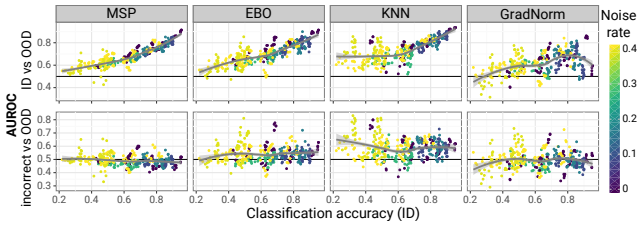

Figure 4. Relationship between ID classification performance and OOD detection performance, considering all ID test samples (top) or only incorrectly classified ones (bottom) in the AUROC metric. Each point corresponds to a single model.

side classification accuracy are characterized by a high $\text{AUROC}_{\text{correct vs. OOD}}$ and a low $\text{AUROC}_{\text{incorrect vs. OOD}}$. On clean, easy datasets like CIFAR10, they exhibit strong OOD detection performance as there are few incorrectly predicted ID samples in the test set (thus the $\text{AUROC}_{\text{incorrect vs. OOD}}$ term is negligible in the overall $\text{AUROC}_{\text{ID vs. OOD}}$) - however, when the number of incorrect prediction grows, the low $\text{AUROC}_{\text{incorrect vs. OOD}}$ becomes a more significant factor. Importantly and as exemplified by Figure 4, for all methods, $\text{AUROC}_{\text{incorrect vs. OOD}}$ is not (or only weakly, $\rho < 0.2$) correlated with classification accuracy. MSP is the most clear-cut example, with a median $\text{AUROC}_{\text{incorrect vs. OOD}}$ of around 0.5 across all dataset-architecture-seed-checkpoint combinations (bottom left of Figure 4) - that is, MSP often is no better (or worse) than a random detector at separating ID mistakes and OOD inputs, no matter how accurate the underlying classifier is. The Top-4 methods in Table 2 are the only ones with a median $\text{AUROC}_{\text{incorrect vs. OOD}} \geq 0.6$ - none of the other methods exceed a median $\text{AUROC}_{\text{incorrect vs. OOD}}$ of 0.55 - see Figure 1.

Takeaway: Would your OOD detector be better off as a failure detector? Accuracy correlating with OOD detection performance is partly symptomatic of many seemingly effective methods being unable to separate *incorrectly classified ID samples* from OOD samples - a bottleneck for robustness to imperfect classification. Claims that post-hoc OOD detection can be improved by simply improving the underlying classifier [60] overlook this fundamental issue.

It's not just about the noise rate We find that for a fixed noise rate in a given dataset, different types/models of label noise yield comparable classification accuracy ($\epsilon_{\text{min}} \geq 0.5$ for all pair-wise comparisons), yet have different effects on OOD performance. Indeed, real label noise is better handled than the same level of synthetic by most methods, with SU labels being the most challenging - this trend is clear in Figure 2. Figure 5 shows an example of how different noise types and checkpointing strategies shape the magnitude and spread of logits. Intuitively, when the noise is spread randomly across samples (SU noise model), it is more difficult to learn which kinds of images or classes to be uncer-

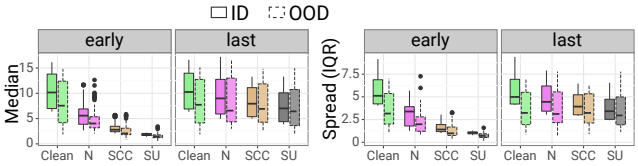

Figure 5. Max Logit ID and OOD score statistics across models trained on Clothing1M, for different noise types & checkpointing.

tain about, leading to consistently lower-confidence predictions across all ID samples (low median, low spread). Conversely, when label noise is more concentrated for certain classes (SCC) and/or for certain features (real noise), the classifier can learn to be more confident in some parts of the input space than others (higher median, higher spread).

Takeaway: Faking it is better than ignoring it Uniform (synthetic) label noise in the training data tends to degrade OOD detection more strongly than class-dependant (synthetic) and instance-dependent (real) label noise. We encourage the use of synthetic uniform labels to evaluate the worst-case performance of OOD detectors, as they can be easily generated for any image classification dataset.

6.3. Design features which hurt or help

Why are the winners the best? In terms of design features, the methods with the strongest performance in a label noise setting have a distance-based scoring function, and take features as input rather than class probabilities. SHE is the only OOD detector satisfying both criteria which doesn't sit at the top of the pile in Table 2 - we attribute its lower performance to two factors: it summarizes the ID dataset only with class-wise means (which may be overly reductive in a label noise setting where variance is larger), and it only considers correctly predicted samples when computing them (which may be small in number if the classifier is inaccurate or the number of classes is high). In contrast, GRAM which includes higher-order raw moments to describe ID data statistics is the top-1 method in Table 2. In the comparison of Figure 6, GRAM and MDSEnsemble - the only methods in our benchmark which incorporate features at different depths in the network - stand out as having the "flattest" accuracy-AUROC curves, which is especially beneficial when the training dataset is inherently difficult (e.g. CIFAR100 due to fine-grained labels or Clothing1M due to the image complexity and diversity). However, we note that the performance of MDSEnsemble and GRAM is highly architecture-dependent - the best OOD detection performance is achieved with a ResNet18 classifier, while MLPMixer and CCT architectures give sub-par results (often sub-50% ie. even worse than a random detector). Whether this large performance variation is due to the layer types, feature dimensionality or other factors, and whether it can be remedied warrants further investigation.

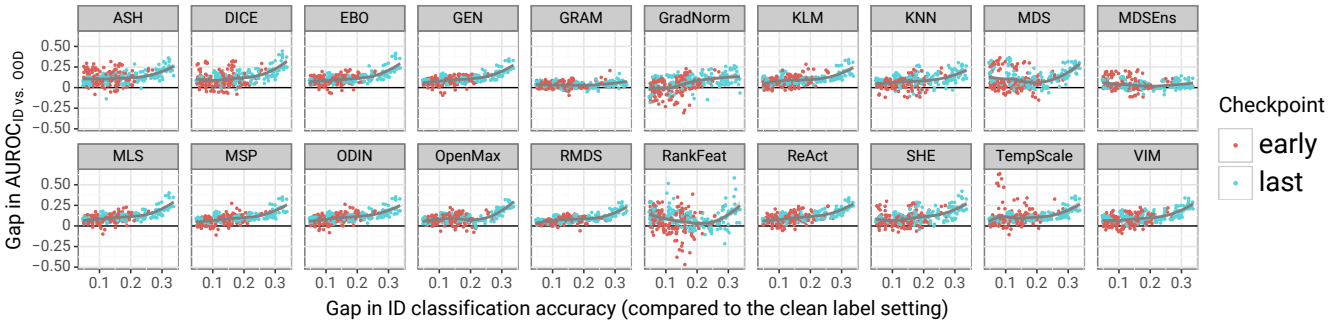

Figure 6. Relation between the drop in accuracy caused by noisy labels and the resulting drop in OOD detection performance across all 20 methods. Each point corresponds to a single model trained with noisy labels.

Takeaway: Distance is healthy Out of the 20 post-hoc OOD detectors in our benchmark, distance-based OOD detectors operating in feature space appear the most promising to cope with the problem of unreliable predictions. Intuitively, distance-based methods are more dissociated from the classifier's prediction, and more dependent on the content/appearance of ID images. In contrast, we did not find compelling evidence that methods targeting class logits or class probabilities for OOD detection are better suited for the noisy label setting.

Are there tricks that work? We consider 3 popular "tricks" aiming to better separate ID vs. OOD samples in logit or probability space - temperature scaling, input perturbation and sparsification - and assess their effectiveness in a noisy label setting (excluding cleanly trained models). To isolate the effect of Softmax temperature scaling and input perturbation, we introduce $\text{ODIN}_{\text{notemp}}$ (ODIN with temperature T fixed to 1) and $\text{ODIN}_{\text{nopert}}$ (perturbation magnitude m set to 0). We find that scaling T by maximizing likelihood on ID validation labels is detrimental ($\epsilon_{\min}[\text{MSP}>\text{TempScale}] = 0.15$), however picking T based OOD validation detection performance does make a statistically significant (though not practically significant) difference ($\epsilon_{\min}[\text{ODIN}_{\text{nopert}}>\text{MSP}] = 0.05$). Input perturbation does not help in a label noise setting: looking at the optimal m selected during $\text{ODIN}_{\text{notemp}}$'s automatic parameter tuning, we observe that as label noise rate increases, the more likely that $m = 0$ is picked (no perturbation). As for feature or weight sparsification, we note that REACT and DICE are the most promising logit-based methods in the $\text{AUROC}_{\text{incorrect vs. OOD}}$ ranking of Figure 1.

6.4. Let's not forget about the validation set

Picking a model checkpoint While it is well-understood that early stopping is beneficial to classification accuracy when training a classifier with noisy labels [34], we investigate whether this extends to OOD detection performance. We compare the OOD detection performance for the 2 checkpointing strategies, and find that for almost all methods, early stopping is beneficial ($\epsilon_{\min}[\text{early}>\text{last}] < 0.5$). However, looking at Figure 6, we note that early stopping may increase the *rate* at which OOD detection performance drops due to label noise for a given drop in accuracy - to an extreme in the case of TempScaling. A closer look at Figure 5 gives some insight into its effect on the logits.

What about OOD detector parameter tuning? Many of the methods in our benchmark involve a set-up step where dataset-specific parameters are computed (e.g. statistics for ID samples) and/or a tuning step where hyperparameters are tuned to maximize OOD detection performance on a held-out validation OOD set. The set of (hyper)parameters for each method is outlined in the supplementary. Among these methods, some make use of *classification labels* during set-up/tuning - e.g. to compute statistics for each class. In a label noise setting, this raises the question of whether to use a *clean* validation set or the *noisy* training set for set-up/tuning, or whether this makes a difference. We compare both settings for the 6 methods in our benchmark making use of class labels during set-up: MDS, RMDS, MDSEnsemble, GRAM, OpenMax and SHE, with results visualized in the supplementary. For SHE which computes the mean of features for each class during set-up, there is no statistically significant difference between using clean validation labels or potentially noisy training labels, although the latter may be better in some cases ($\epsilon_{\min}[\text{SHE}_{\text{val}}>\text{SHE}_{\text{train}}] = 1$ and $\epsilon_{\min}[\text{SHE}_{\text{train}}>\text{SHE}_{\text{val}}] = 0.63$). For methods based on the Malahanobis score, using noisy training labels to compute class-wise feature means and tied covariance is better ($\epsilon_{\min}[\text{MDS}_{\text{train}}>\text{MDS}_{\text{val}}] = 0.19$ and $\epsilon_{\min}[\text{RMDS}_{\text{train}}>\text{RMDS}_{\text{val}}] = 0$) - intuitively, the class-specific statistics are more accurate with more data. Common to these 3 methods is that the OOD score at test-time does not depend on the *predicted* class (likely to be incorrect in a label noise setting), but is rather based on distance to the *closest* class in feature space (regardless of what class is predicted). OpenMax computes the mean logit per class, only considering correctly predicted samples (labels are used to check correctness) - using a potentially

noisy training set yields consistently better performance ($\epsilon_{\min}[\text{OpenMax}_{\text{train}} > \text{OpenMax}_{\text{val}}] = 0$). Lastly, and in contrast to the other methods, GRAM benefits from using clean validation samples rather than a large number of noisy training samples for computing class-specific bounds of feature correlations ($\epsilon_{\min}[\text{GRAM}_{\text{val}} > \text{GRAM}_{\text{train}}] = 0.23$). However, the performance gap between the two settings is small.

Takeaway: Clean isn't always better or possible The use of clean vs. noisy labels during label-based parameter tuning is an important consideration. For distance-based methods which compute class-wise statistics, it appears that quantity often trumps quality, even when over 30% of training labels are incorrect. This is promising for applications where a clean validation set is not available (e.g. medical imaging where labels are inherently subjective [28]).

6.5. What about a more realistic setting?

We have thus far studied OOD detection in a simple (but standard [67]) setting where the base classifier is trained from scratch, and where there is strong semantic and covariate shift between ID and OOD images. Yet in practice, pre-training is widely adopted, and distribution shifts may be much more subtle. We therefore extend our study of label noise to *fine-grained semantic shift detection* with a base classifier that has been *pre-trained on ImageNet* [8] before being trained on a dataset of interest. We follow the Semantic Shift Benchmark (SSB), where the goal is to detect *unknown classes* from a known dataset (e.g. held-out bird species from the CUB [61] dataset or held-out aircraft model variants from FGVC-Aircraft [38]). Using SSB splits, we train ResNet50s (pre-trained) on half of the classes from CUB/FGVC-Aircraft (448x448 images), and we evaluate post-hoc OOD detection performance on known classes from the test set **(ID)** vs. the remaining unseen classes **(OOD)** split into 3 increasingly difficult sets. Since clean vs. real noisy label pairs are not available, we inject synthetic label noise in the training set (SU noise model) and follow the same evaluation procedure as in previous sections. Fig. 7 summarizes its detrimental effect on fine-grained semantic shift detection across the 20 studied OOD detection methods: increasing label noise and "difficulty" of the OOD set act as orthogonal bottlenecks to detection performance. Increased label noise pulls $\text{AUROC}_{\text{ID vs. OOD}}$ and $\text{AUROC}_{\text{incorrect vs. OOD}}$ to 50%.

Takeaway: Limitations of post-hoc OOD detectors extend beyond toy settings Even in a more realistic setting where the base classifier has first been pre-trained on ImageNet and OOD samples are similar in appearance to the ID dataset, all 20 methods poorly separate incorrectly classifed ID samples from OOD samples, and degrade when the classifier has been trained on noisy labels.

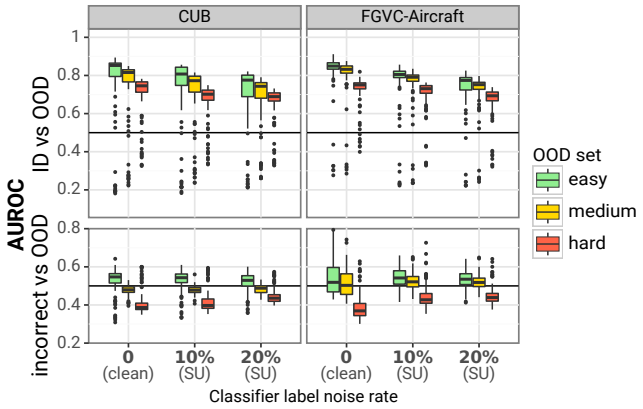

Figure 7. Each boxplot shows the performance distribution across 6 classifiers (3 seeds, 2 checkpoints) $\times$ 20 post-hoc methods, considering all ID test samples (top) or only incorrectly classified ones (bottom) in the AUROC metric.

7. Zooming out

Study limitations and possible extensions We have focused on post-hoc OOD detection methods due to their pragmatic appeal and to maintain experimental feasibility. Extending this study to training-based OOD detection methods [71] would of course be valuable. Aligning with OOD benchmarks [67], we also trained the base classifiers with a standard discriminative objective. Alternative supervision schemes may also be considered, and the effect of pre-training (and on what?) would be interesting to further analyse in a label noise setting, as it been shown to improve post-hoc OOD detection performance [2, 20, 33]. Lastly, the potential of noisy label removal [29, 43] or noise-robust learning [27, 63] techniques from the label noise literature (designed with classification performance in mind) for improving OOD detection would be a natural next step.

Conclusion We have explored the intersection between classification label noise and OOD detection, and conducted extensive experiments to extract new insights into the limitations of existing post-hoc methods. Our findings also echo the need to re-think the aims and evaluation of OOD detection in the context of safe deployment [26] (e.g. do we really want to exclude ID misclassifications from detection?). We hope that this work paves the way for future investigations which prioritize the robustness and applicability of OOD detection models in practical, imperfect classification scenarios which account for data uncertainty.

8. Acknowledgements

This work was supported by the Danish Data Science Academy, which is funded by the Novo Nordisk Foundation (NNF21SA0069429) and VILLUM FONDEN (40516). Thanks to the Pioneer Centre for AI (DNRF grant P1).

References

[1] Görkem Algan and Ilkay Ulusoy. Image classification with deep learning in the presence of noisy labels: A survey. *Knowledge-Based Systems*, 215:106771, 2021. 2

[2] Anders Johan Andreassen, Yasaman Bahri, Behnam Neyshabur, and Rebecca Roelofs. The evolution of out-of-distribution robustness throughout fine-tuning. *Transactions on Machine Learning Research*, 2022. 8

[3] Abhijit Bendale and Terrance E. Boult. Towards open set deep networks. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1563–1572, 2016. 3

[4] Julian Bitterwolf, Maximilian Mueller, and Matthias Hein. In or out? fixing imagenet out-of-distribution detection evaluation. In *ICML*, 2023. 1

[5] Lukas Bossard, Matthieu Guillaumin, and Luc Van Gool. Food-101 – mining discriminative components with random forests. In *European Conference on Computer Vision*, 2014. 4

[6] Mircea Cimpoi, Subhransu Maji, Iasonas Kokkinos, Sammy Mohamed, and Andrea Vedaldi. Describing textures in the wild. In *2014 IEEE Conference on Computer Vision and Pattern Recognition*, pages 3606–3613, 2014. 4

[7] Eustasio Del Barrio, Juan A Cuesta-Albertos, and Carlos Matrán. An optimal transportation approach for assessing almost stochastic order. In *The Mathematics of the Uncertain*, pages 33–44. Springer, 2018. 4

[8] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, 2009. 8

[9] Li Deng. The mnist database of handwritten digit images for machine learning research [best of the web]. *IEEE Signal Processing Magazine*, 29(6):141–142, 2012. 4

[10] Andrija Djurisic, Nebojsa Bozanic, Arjun Ashok, and Rosanne Liu. Extremely simple activation shaping for out-of-distribution detection. In *The Eleventh International Conference on Learning Representations*, 2023. 1, 3

[11] Rotem Dror, Segev Shlomov, and Roi Reichart. Deep dominance - how to properly compare deep neural models. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 2773–2785. Association for Computational Linguistics, 2019. 4

[12] Stanislav Fort, Jie Ren, and Balaji Lakshminarayanan. Exploring the limits of out-of-distribution detection. In *Advances in Neural Information Processing Systems*, 2021. 3

[13] Benoit Frenay and Michel Verleysen. Classification in the presence of label noise: A survey. *IEEE Transactions on Neural Networks and Learning Systems*, 25(5):845–869, 2014. 2

[14] Ido Galil, Mohammed Dabbah, and Ran El-Yaniv. A framework for benchmarking class-out-of-distribution detection and its application to imagenet. In *The Eleventh International Conference on Learning Representations*, 2023. 1, 2, 3, 5

[15] Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning*, pages 1321–1330. PMLR, 2017. 3

[16] Ali Hassani, Steven Walton, Nikhil Shah, Abulikemu Abuduweili, Jiachen Li, and Humphrey Shi. Escaping the big data paradigm with compact transformers. 2021. 4

[17] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 4

[18] Patrick Helber, Benjamin Bischke, Andreas Dengel, and Damian Borth. Eurosat: A novel dataset and deep learning benchmark for land use and land cover classification. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 2019. 4

[19] Dan Hendrycks and Kevin Gimpel. 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. 1, 3

[20] Dan Hendrycks, Kimin Lee, and Mantas Mazeika. Using pre-training can improve model robustness and uncertainty. In *Proceedings of the 36th International Conference on Machine Learning*, pages 2712–2721. PMLR, 2019. 8

[21] Dan Hendrycks, Steven Basart, Mantas Mazeika, Andy Zou, Joseph Kwon, Mohammadreza Mostajabi, Jacob Steinhardt, and Dawn Song. Scaling out-of-distribution detection for real-world settings. In *Proceedings of the 39th International Conference on Machine Learning*, pages 8759–8773. PMLR, 2022. 3

[22] Dan Hendrycks, Steven Basart, Mantas Mazeika, Andy Zou, Joe Kwon, Mohammadreza Mostajabi, Jacob Steinhardt, and Dawn Song. Scaling out-of-distribution detection for real-world settings. *ICML*, 2022. 3

[23] R. Huang and Y. Li. Mos: Towards scaling out-of-distribution detection for large semantic space. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8706–8715, Los Alamitos, CA, USA, 2021. IEEE Computer Society. 3

[24] Rui Huang, Andrew Geng, and Yixuan Li. On the importance of gradients for detecting distributional shifts in the wild. In *Advances in Neural Information Processing Systems*, 2021. 3

[25] Galadrielle Humblot-Renaux, Sergio Escalera, and Thomas B. Moeslund. Beyond auroc & co. for evaluating out-of-distribution detection performance. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 3881–3890, 2023. 4

[26] Paul F Jaeger et al. A call to reflect on evaluation practices for failure detection in image classification. In *ICLR*, 2023. 8

[27] Lu Jiang, Di Huang, Mason Liu, and Weilong Yang. Beyond synthetic noise: Deep learning on controlled noisy labels. In *Proceedings of the 37th International Conference on Machine Learning*, pages 4804–4815. PMLR, 2020. 8

[28] Davood Karimi, Haoran Dou, Simon K. Warfield, and Ali Gholipour. Deep learning with noisy labels: Exploring techniques and remedies in medical image analysis. *Medical Image Analysis*, 65:101759, 2020. 1, 2, 8

[29] Taehyeon Kim, Jongwoo Ko, sangwook Cho, JinHwan Choi, and Se-Young Yun. Fine samples for learning with noisy labels. In *Advances in Neural Information Processing Systems*, pages 24137–24149. Curran Associates, Inc., 2021. 8

[30] Konstantin Kirchheim, Marco Filax, and Frank Ortmeier. Pytorch-ood: A library for out-of-distribution detection based on pytorch. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 4351–4360, 2022. 1

[31] Alex Krizhevsky and Geoffrey Hinton. Learning multiple layers of features from tiny images. Technical Report 0, University of Toronto, Toronto, Ontario, 2009. 2, 4

[32] Kimin Lee, Kibok Lee, Honglak Lee, and Jinwoo Shin. A simple unified framework for detecting out-of-distribution samples and adversarial attacks. In *Advances in Neural Information Processing Systems*. Curran Associates, Inc., 2018. 1, 3

[33] Jingyao Li, Pengguang Chen, Zexin He, Shaozuo Yu, Shu Liu, and Jiaya Jia. Rethinking out-of-distribution (ood) detection: Masked image modeling is all you need. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11578–11589, 2023. 8

[34] Mingchen Li, Mahdi Soltanolkotabi, and Samet Oymak. Gradient descent with early stopping is provably robust to label noise for overparameterized neural networks. In *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, pages 4313–4324. PMLR, 2020. 7

[35] Shiyu Liang, Yixuan Li, and R. Srikant. Enhancing the reliability of out-of-distribution image detection in neural networks. In *International Conference on Learning Representations*, 2018. 3

[36] Sheng Liu, Jonathan Niles-Weed, Narges Razavian, and Carlos Fernandez-Granda. Early-learning regularization prevents memorization of noisy labels. *Advances in Neural Information Processing Systems*, 33, 2020. 2

[37] Weitang Liu, Xiaoyun Wang, John Owens, and Yixuan Li. Energy-based out-of-distribution detection. In *Advances in Neural Information Processing Systems*, pages 21464–21475. Curran Associates, Inc., 2020. 3

[38] S. Maji, J. Kannala, E. Rahtu, M. Blaschko, and A. Vedaldi. Fine-grained visual classification of aircraft. Technical report, 2013. 8

[39] Yuval Netzer, Tao Wang, Adam Coates, Alessandro Bissacco, Bo Wu, and Andrew Y. Ng. Reading Digits in Natural Images with Unsupervised Feature Learning. In *NIPS Workshop on Deep Learning and Unsupervised Feature Learning 2011*, 2011. 2, 4

[40] Amanda Olmin and Fredrik Lindsten. Robustness and reliability when training with noisy labels. In *Proceedings of The 25th International Conference on Artificial Intelligence and Statistics*, pages 922–942. PMLR, 2022. 2

[41] Diane Oyen, Michal Kucer, Nick Hengartner, and Har Simrat Singh. Robustness to label noise depends on the shape of the noise distribution. In *Advances in Neural Information Processing Systems*, 2022. 2

[42] Chao Pan, Bo Yuan, Wei Zhou, and Xin Yao. Towards robust uncertainty estimation in the presence of noisy labels. In *Artificial Neural Networks and Machine Learning – ICANN 2022*, pages 673–684, Cham, 2022. Springer International Publishing. 2

[43] Geoff Pleiss, Tianyi Zhang, Ethan Elenberg, and Kilian Q Weinberger. Identifying mislabeled data using the area under the margin ranking. In *Advances in Neural Information Processing Systems*, pages 17044–17056. Curran Associates, Inc., 2020. 8

[44] Jie Ren, Stanislav Fort, Jeremiah Liu, Abhijit Guha Roy, Shreyas Padhy, and Balaji Lakshminarayanan. A simple fix to mahalanobis distance for improving near-ood detection, 2021. 3

[45] David Rolnick, Andreas Veit, Serge Belongie, and Nir Shavit. Deep learning is robust to massive label noise, 2018. 2

[46] Ragav Sachdeva, Filipe R. Cordeiro, Vasileios Belagiannis, Ian Reid, and Gustavo Carneiro. Evidentialmix: Learning with combined open-set and closed-set noisy labels. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 3607–3615, 2021. 2

[47] Mohammadreza Salehi, Hossein Mirzaei, Dan Hendrycks, Yixuan Li, Mohammad Hossein Rohban, and Mohammad Sabokrou. A unified survey on anomaly, novelty, open-set, and out of-distribution detection: Solutions and future challenges. *Transactions on Machine Learning Research*, 2022. 1

[48] Chandramouli Shama Sastry and Sageev Oore. Detecting out-of-distribution examples with Gram matrices. In *Proceedings of the 37th International Conference on Machine Learning*, pages 8491–8501. PMLR, 2020. 3

[49] Vikash Sehwag, Arjun Nitin Bhagoji, Liwei Song, Chawin Sitawarin, Daniel Cullina, Mung Chiang, and Prateek Mittal. Analyzing the robustness of open-world machine learning. In *Proceedings of the 12th ACM Workshop on Artificial Intelligence and Security*, page 105–116, New York, NY, USA, 2019. Association for Computing Machinery. 3

[50] Hwanjun Song, Minseok Kim, and Jae-Gil Lee. SELFIE: Refurbishing unclean samples for robust deep learning. In *ICML*, 2019. 2

[51] Hwanjun Song, Minseok Kim, Dongmin Park, Yooju Shin, and Jae-Gil Lee. Learning from noisy labels with deep neural networks: A survey. *IEEE Transactions on Neural Networks and Learning Systems*, pages 1–19, 2022. 2

[52] Hyun Oh Song, Yu Xiang, Stefanie Jegelka, and Silvio Savarese. Deep metric learning via lifted structured feature embedding. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 4

[53] Yue Song, Nicu Sebe, and Wei Wang. Rankfeat: Rank-1 feature removal for out-of-distribution detection. In *Advances in Neural Information Processing Systems*, 2022. 3

[54] Alexander Sorokin and David Forsyth. Utility data annotation with amazon mechanical turk. In *2008 IEEE Computer*

Society Conference on Computer Vision and Pattern Recognition Workshops, pages 1–8, 2008. 1

[55] Yiyou Sun and Yixuan Li. Dice: Leveraging sparsification for out-of-distribution detection. In *Computer Vision – ECCV 2022*, pages 691–708, Cham, 2022. Springer Nature Switzerland. 3

[56] Yiyou Sun, Chuan Guo, and Yixuan Li. React: Out-of-distribution detection with rectified activations. In *Advances in Neural Information Processing Systems*, pages 144–157. Curran Associates, Inc., 2021. 3

[57] Yiyou Sun, Yifei Ming, Xiaojin Zhu, and Yixuan Li. Out-of-distribution detection with deep nearest neighbors. In *Proceedings of the 39th International Conference on Machine Learning*, pages 20827–20840. PMLR, 2022. 1, 3

[58] Ilya Tolstikhin, Neil Houlsby, Alexander Kolesnikov, Lucas Beyer, Xiaohua Zhai, Thomas Unterthiner, Jessica Yung, Andreas Peter Steiner, Daniel Keysers, Jakob Uszkoreit, Mario Lucic, and Alexey Dosovitskiy. MLP-mixer: An all-MLP architecture for vision. In *Advances in Neural Information Processing Systems*, 2021. 4

[59] Dennis Ulmer, Christian Hardmeier, and Jes Frellsen. deepsignificance-easy and meaningful statistical significance testing in the age of neural networks. *arXiv preprint arXiv:2204.06815*, 2022. 4

[60] Sagar Vaze, Kai Han, Andrea Vedaldi, and Andrew Zisserman. Open-set recognition: A good closed-set classifier is all you need. In *International Conference on Learning Representations*, 2022. 1, 2, 3, 5, 6

[61] C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie. The caltech-ucsd birds-200-2011 dataset. Technical Report CNS-TR-2011-001, California Institute of Technology, 2011. 8

[62] H. Wang, Z. Li, L. Feng, and W. Zhang. Vim: Out-of-distribution with virtual-logit matching. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4911–4920, Los Alamitos, CA, USA, 2022. IEEE Computer Society. 3

[63] Hongxin Wei, Huiping Zhuang, Renchunzi Xie, Lei Feng, Gang Niu, Bo An, and Yixuan Li. Mitigating memorization of noisy labels by clipping the model prediction. In *Proceedings of the 40th International Conference on Machine Learning*. JMLR.org, 2023. 8

[64] Jiaheng Wei, Zhaowei Zhu, Hao Cheng, Tongliang Liu, Gang Niu, and Yang Liu. Learning with noisy labels revisited: A study using real-world human annotations. In *International Conference on Learning Representations*, 2022. 2, 4

[65] Tong Xiao, Tian Xia, Yi Yang, Chang Huang, and Xiaogang Wang. Learning from massive noisy labeled data for image classification. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2691–2699, 2015. 4

[66] Jingkang Yang, Kaiyang Zhou, Yixuan Li, and Ziwei Liu. Generalized out-of-distribution detection: A survey. *arXiv preprint arXiv:2110.11334*, 2021. 1

[67] Jingkang Yang, Pengyun Wang, Dejian Zou, Zitang Zhou, Kunyuan Ding, WENXUAN PENG, Haoqi Wang, Guangyao Chen, Bo Li, Yiyou Sun, Xuefeng Du, Kaiyang Zhou, Wayne Zhang, Dan Hendrycks, Yixuan Li, and Ziwei Liu. OpenOOD: Benchmarking generalized out-of-distribution detection. In *Thirty-sixth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2022. 1, 2, 3, 4, 8

[68] Xiaoyong Yuan, Pan He, Qile Zhu, and Xiaolin Li. Adversarial examples: Attacks and defenses for deep learning. *IEEE Transactions on Neural Networks and Learning Systems*, 30 (9):2805–2824, 2019. 3

[69] Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning requires rethinking generalization. In *International Conference on Learning Representations*, 2017. 2

[70] Jinsong Zhang, Qiang Fu, Xu Chen, Lun Du, Zelin Li, Gang Wang, xiaoguang Liu, Shi Han, and Dongmei Zhang. Out-of-distribution detection based on in-distribution data patterns memorization with modern hopfield energy. In *The Eleventh International Conference on Learning Representations*, 2023. 1, 3

[71] Jingyang Zhang, Jingkang Yang, Pengyun Wang, Haoqi Wang, Yueqian Lin, Haoran Zhang, Yiyou Sun, Xuefeng Du, Kaiyang Zhou, Wayne Zhang, Yixuan Li, Ziwei Liu, Yiran Chen, and Hai Li. Openood v1.5: Enhanced benchmark for out-of-distribution detection. *arXiv preprint arXiv:2306.09301*, 2023. 8