# Why Out-of-Distribution Detection Experiments Are Not Reliable - Subtle Experimental Details Muddle the OOD Detector Rankings

**Kamil Szyc**[1] **Tomasz Walkowiak**[1] **Henryk Maciejewski**[1]

[1]Faculty of Information and Communication Technology, Wrocław University of Science and Technology, Wrocław, Poland

## Abstract

Reliable detection of out-of-distribution (OOD) instances is becoming a critical requirement for machine learning systems deployed in safety-critical applications. Recently, many OOD detectors have been developed in the literature, and their performance has been evaluated using empirical studies based on well-established benchmark datasets. However, these studies do not provide a conclusive recommendation because the performance of OOD detection depends on the benchmark datasets. In this work, we want to question the reliability of the OOD detection performance numbers obtained from many of these empirical experiments. We report several experimental conditions that are not controlled and lead to significant changes in OOD detector performance and rankings of OOD methods. These include the technicalities related to how the DNN was trained (such as seed, train/test split, etc.), which do not change the accuracy of closed-set DNN models but may significantly change the performance of OOD detection methods that rely on representation from these DNNs. We performed extensive sensitivity studies in image and text domains to quantify the instability of OOD performance measures due to unintuitive experimental factors. These factors need to be more rigorously controlled and accounted for in many current OOD experiments. Experimental studies in OOD detection should improve methodological standards regarding experiment control and replication.

## 1 INTRODUCTION

As machine learning systems are increasingly used in real-world, high-stakes applications, such as autonomous vehicles or clinical use, safety is becoming the critical requirement for ML technologies Hendrycks et al. [2021], Nicora et al. [2022], McAllister et al. [2017]. To ensure safe and robust ML recognition, ML models must reliably detect out-of-distribution (OOD) examples.

A variety of state-of-the-art OOD detectors have been developed in recent literature. They either rely on representations learned by closed-set classifiers (post-hoc methods) or require additional training (e.g., contrastive learning or training with exposure to outlier examples).

Progress in OOD detection is justified empirically: a new method is considered the new SoTA if it performs better on well-established OOD benchmarks, with often only slight improvement over previous methods. For instance, the well-known MDS method based on Mahalanobis distance (Lee et al. [2018]) outperforms the LID method (Ma et al. [2018]) by ca. 1 p.p. (in AUC measure, with SVHN as in-distribution (ID)), and ODIN (Liang et al. [2017]), by 2.5 p.p. on CIFAR-10 as ID; or recent SoTA VIM method (Wang et al. [2022]) outperforms the second best MDS by ca. 0.2 to 4 p.p. (with ImageNet-1K as ID). See more examples in Table 1.

These studies do not lead to conclusive results concerning the objective ranking of OOD detectors, as the performance of OOD detectors depends on the benchmark, with no universal SoTA method available Tajwar et al. [2021], Yang et al. [2022].

In this work, we want to raise the question about the reliability of the current measures of progress in OOD detection methods. The question is motivated by the problem with reproducibility of OOD performance metrics: trying to reproduce the OOD benchmark results for a given DNN architecture and ID/OOD datasets often leads to performance figures different than reported in the original paper (unless the original code is used intact). We make an (unintuitive) observation that OOD performance numbers are sensitive to subtle variations/changes in the experiments. We aim to identify these factors of variability in OOD detection performance. We show that in many current OOD detection empirical studies, these factors are not rigorously controlled,

*Proceedings of the 39th Conference on Uncertainty in Artificial Intelligence* (UAI 2023), PMLR 216:2078–2088.

which impairs the reliability of OOD detection performance measures reported in many recent works.

Our main contributions are the following:

- We show that OOD performance numbers are sensitive to experimental factors, which do not affect the closed-set accuracy of the DNN models used in the study but significantly change results for OOD detection. These subtle experimental details include the seed, train/test split, etc., used in training the DNN model.
- We analyze the experimental settings/factors and quantify the instability of OOD performance measures as a function of these factors. In these analyses, we considered OOD detection in the image and text recognition domains, with representations generated by CNNs or transformers (BERT).
- We suggest an explanation of the nature of this problem (as these factors barely affect discriminative models but vastly change generative models implied by OOD detectors).
- We formulate conclusions for improving the reliability of OOD detection benchmarks. We postulate that the OOD experimental factors we identify should be rigorously reported along with the results. Also, OOD experimental studies should not rely on single-point results but should average over these factors of instability for a more reliable comparison of OOD detectors.

Code is available at `https://github.com/TrustAIRiders/OoD-instability`.

## 2 RELATED WORK

The problem of reliability of performance measures reported in ML empirical studies was previously raised by Recht et al. [2018, 2019], Engstrom et al. [2020]. They raised this crucial question: "'How reliable are our current measures of progress in machine learning?"' - in the context of reliability of the accuracy of closed-set DNN models evaluated on standard benchmarks such as CIFAR-10, ImageNet, etc. These studies show that dataset replication to semantically the same dataset (CIFAR-10, ImageNet) leads to a drop in model accuracy. The results of these studies suggest that current DNN accuracy numbers on standard benchmarks are susceptible to subtle changes in the experiment design related to how the dataset is replicated. Further concerns about the reliability of the performance measures come from Hendrycks et al. [2019b], who show that DNNs are susceptible to natural adversarial examples - hence the experiment only replicates within the standard benchmark.

The issue of irreproducibility of results and analysis of potential sources of variability have been investigated in many different areas of machine learning: NLP Belz et al. [2021], deep reinforcement learning Henderson et al. [2018], forecasting with ML methods Makridakis et al. [2018], recommender systems Ferrari Dacrema et al. [2021] or image recognition with deep learning Bouthillier et al. [2019]. A systematic taxonomy of sources of variation that lead to irreproducibility of ML results is given by Gundersen et al. [2022].

Table 1: Best and the second best OOD detectors for the comprehensive benchmark Yang et al. [2022], along with the AUC result and difference in AUC between the second and first method. The SoTA OOD detectors were proposed in: MDS(Lee et al. [2018]), KNN(Sun et al. [2022]), Gram(Sastry and Oore [2020]), MLS(Hendrycks et al. [2022]), VIM(Wang et al. [2022]). MSP(Hendrycks and Gimpel [2017]).

| **Benchmark**<br>ID vs OOD | **OOD Detector**<br>Best(w/ AUC) / 2nd Best |
|---|---|
| MNIST vs Near-OOD | MDS(98.0) / KNN(-1.5) |
| MNIST vs Far-OOD | Gram(99.8) / MLS(-0.9) |
| CIFAR-10 vs Near-OOD | KNN(90.5) / VIM(-2.5) |
| CIFAR-10 vs Far-OOD | KNN(92.8) / VIM(-0.1) |
| CIFAR-100 vs Near-OOD | MLS(81.0) / MSP(-0.9) |
| CIFAR-100 vs Far-OOD | VIM(82.4) / KNN(-0.2) |
| ImageNet vs Near-OOD | KNN(80.8) / VIM(-0.9) |
| ImageNet vs Far-OOD | VIM(98.4) / KNN(-0.4) |

In this work, we want to analyze the reliability of progress measures in the OOD detection field. To our best knowledge, the problem of replicating results of experimental studies that justify new SoTA OOD detection methods was not raised in the literature.

The problem of inconclusive rankings of OOD detectors. Recent comprehensive studies show that no OOD detector consistently outperforms other methods Tajwar et al. [2021], Yang et al. [2022]. In Table 1, we illustrate this on commonly used ID datasets with a comprehensive collection of OOD datasets grouped as Near- or Far-OOD (semantically similar or dissimilar to the ID data). We observe that the improvement between methods is usually marginal. However, in this work, we show that the instability ranges (in AUC) due to experimental factors are much larger; hence in line with our findings, we should take the OOD rankings with caution.

## 3 EXPERIMENTAL RESULTS

### 3.1 OVERVIEW

Typical DNN models are designed to achieve the best possible closed-set classification accuracy. There are many decisions to be made when designing them, such as setting

the appropriate hyperparameters. In practice, many of these decisions have minimal impact on the final closed-set results. On the other hand, our experiments suggest that they can differ significantly in how the model builds its decision boundaries to separate classes in the feature space. Therefore, even small (in a closed-set perspective) changes in hyperparameters can greatly affect OOD performance.

We analyze the following sources of instability that can affect OOD performance: **minor model architecture implementation differences** (for example, there are many implementations for ResNet for CIFAR-10), **the influence of initial seed** (we suggest that initial weights values can much influence how the model's final decision boundaries are built, and although it does not strongly affect the accuracy of the closed-set it is of enormous importance in the open-set classification problem), **when exactly trained model is stopped** (usually the optimizer during last epochs slightly improves - if at all - the quality of closed classification), **train/test split of in-distribution data**, **which OOD examples to choose for evaluation** (some datasets like SVHN contain many more examples that test in-distribution examples, we only need to select a subset from them), and **choosing the right data augmentation strategy** (although it is evident that this also strongly changes the closed-set accuracy, the changes for OOD detection are even more significant).

We quantify the effect of these experiment variations on the instability of OOD performance measures. In Sections 3.3 through 3.8, we focus on image data with representations generated by CNNs; in Section 3.9, we expand the evaluation onto text data with BERT (transformer-based) representations.

### 3.2 EXPERIMENT ORGANIZATION

All experiments followed a similar procedure. First, we trained from scratch the CNN model for classical closed-set classification of images and tuned up the BERT-based model for texts. Next, we calculated the OOD scores and, on the basis of these, we evaluated each OOD detection method.

Our experiments in the image domain were based on ResNetHe et al. [2016](mostly ResNet-152) architecture for CIFAR-10 as in-distribution images and the MobileNet-v2Sandler et al. [2018] architecture for CIFAR-100. We used fixed seed (mainly as 0), SGD optimizer, cross-entropy loss, scheduler for learning rate, early stopping and RandomCrop, and RandomHorizontalFlip with Normalize as an augmentation strategy. As out-of-distribution images, we used SVHN and opposite CIFARs. The experiments in the text domain were based on BERT representations (Devlin et al. [2019]) fed into the fully connected layer for classification; see Section 3.9 for details.

We maintained a 1:1 ratio of known to unknown data in the testing phase. We tested 7 OOD scoring methods based on different principles: MSP (Hendrycks and Gimpel [2017]), MaxLogits (denoted in tables as ML, Hendrycks et al. [2019a]), FreeEnergy (denoted as FE, Liu et al. [2020]), KNN (Sun et al. [2022]), LOF (Breunig et al. [2000]) with Euclidean and Cosine distance, and Mahalanobis (denoted in tables as Mah) (Maciejewski et al. [2022]). The former three methods work in logits space, the latter - in the feature space. We used standard evaluation metrics: TNR at TPR 95%, AUC (or AUCROC), detection accuracy (DTACC), and AUPR. However, we only report AUC in the paper as it is the most representative; the other metrics are presented in the supplementary materials.

We have not reported all hyperparameters for the sake of clarity. However, we have publicly available codes where all details are presented. Furthermore, we keep the same parameters for all experiments except for the source of the instability under investigation.

### 3.3 CNN ARCHITECTURE FEATURES

The CNN architecture is usually designed for high-resolution images, such as ImageNet. The models need to be modified to perform experiments on smaller datasets. For example, to train the ResNet model on CIFAR-10 from scratch, specific modifications (such as adjusting the size of the convolutional kernels, the number of pooling operations, or even the number of convolutional layers) are applied in the network architecture. There are no clear guidelines in this regard. We want to highlight that simply using the name of a model such as "ResNet" can be insufficient to explain the details of implementation and the subtle differences in architecture. Those differences can impact OoD detection performance, which is not considered in many research studies. In our experiments, we have trained three different models based on the publicly available version of ResNet-101 or ResNet-110. We denoted them as type-0[1], type-1[2], and type-2[3]. The results are presented in Table 2.

It can be seen that all three models have similar closed-set accuracy, around 93%. However, the OOD detection metrics are not stable. See MaxLogit (denoted as ML in the table) for SVHN as OOD data - the range of AUC is from 70.13 to 88.99. The LOF with Euclidean distance ($\text{LOF}_E$) for CIFAR-100 as OOD data has a range from 78.72 to 86.82. Note also that the rankings of OOD detectors are different for each type. For example, for CIFAR-10 vs. SVHN and type-0, the top-3 methods are (KNN, Mahalanobis, MSP), while for type-2 - (LOF with cosine distance ($\text{LOF}_C$), Ma-

[1] https://github.com/kuangliu/pytorch-cifar

[2] https://github.com/y0ast/pytorch-snippets/tree/main/minimal_cifar

[3] https://github.com/akamaster/pytorch_resnet_cifar10

halanobis, LOF with euclidean distance (LOF$_E$)). Similarly, for CIFAR-100 vs. CIFAR-10, OOD detector rankings also change with the architecture type.

### 3.4 CNN TRAINING INITIAL SEEDS

We trained models with ten different initial seeds. There are no other differences between the models. Results are presented in Table 3.

We see that the closed-set accuracy of the models is very stable. The standard deviation for MobileNet trained on CIFAR-100 is only 0.27, and for ResNet trained on CIFAR-10, it is 0.31. However, we postulate that the initial seed has a significant effect on the representations generated by the CNN, while the decision boundaries are more stable. In particular, the feature-based OOD detectors (such as KNN, LOF, or Mah) are extremely variable, while the logit-based methods (ML, MSP, FE) are less sensitive. This effect is most prominent for ResNet with CIFAR-10 vs. SVHN: the standard deviation of AUC for the KNN OOD detector is equal to 14.89, with a delta of 50.02 (maximum minus minimum value). Hence, this method can be the best or the worst, and it only depends on the initial seed. We observe a similar phenomenon for other benchmarks and methods: see MobileNet model in problem CIFAR-100 vs. SVHN and LOF with Euclidean distance (LOF$_E$) with delta 11.50 and ranking from 1st to last, or Mahalanobis (ResNet in problem CIFAR-10 vs. CIFAR-100) where the standard deviation of AUC is 6.84 with a delta of 21.53.

### 3.5 CNN TRAINING EPOCHS

Next, we trained the models and observed their state during training. We stored the state of the models after each epoch together with the OOD detection evaluation. The closed-set accuracy of the models did not change significantly after reaching a specific value. We adopted an early stopping approach, i.e., the models stopped learning after 10 epochs without reducing the training loss.

The obtained AUC and ACC are presented in Figures 1. Again, we can see that the closed-set accuracy is very stable. Similarly, the AUC is stable for MobileNet and Resnet when CIFAR-100 is OOD data. However, interestingly, in the problem with SVHN as OOD data, both models show high fluctuations in the AUC metric. Consequently, the final ranking of the OOD detectors changes. It only depends on the choice of the final epoch.

### 3.6 OOD EXAMPLES SELECTION

In performed experiments, we maintain a 1:1 ratio of known to unknown samples in our experiments. The in-distribution test set of CIFAR-10 and CIFAR-100 contains 10,000 sam-

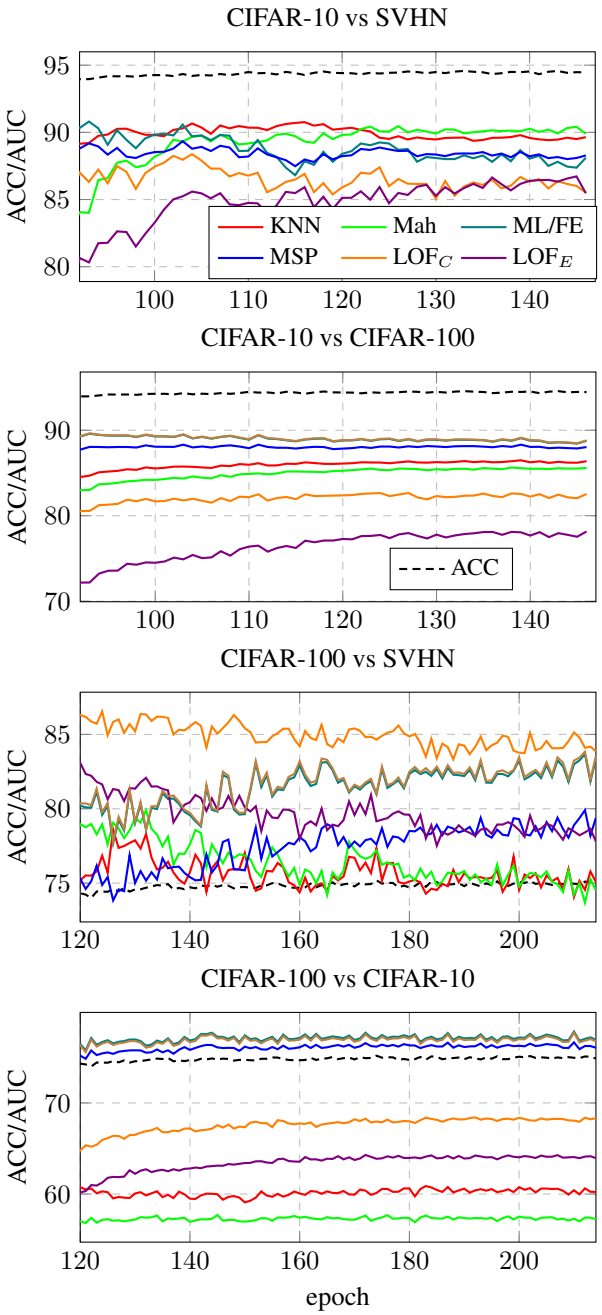


Figure 1: OOD detection instability as a function of the number of learning epochs for four different OOD tasks. The black dashed line shows the close-set (ACC) accuracy, while the solid lines show the AUC for the analyzed OOD methods. The MaxLogit (ML) and FreeEnergy (FE) plots are indistinguishable, so they are represented by a single line. There is almost no change in the closed set ACCs, while for SVHN as OOD, the AUC results for all methods show high variability, which may cause rank changes for some OOD methods. For CIFAR-100/CIFAR-10 as OOD set, the AUC has less variability than in the previous case and this does not affect the ranking of the OOD methods.

Table 2: Results for the closed set classification on CIFAR-10 and OOD detection for three modifications of the ResNet architecture and two OOD data sets (SVHN and CIFAR-100). Seven OOD methods were tested: KNN, Mahalanobis (Mah), MaxLogit (ML), MSP, LOF with cosine distance ($LOF_C$), LOF with euclidean distance ($LOF_E$), and FreeEnergy (FE) (see section 3.2 for references). Bold shows the best results for each of OOD tasks. It can be seen that changing the details of the architecture leads to different winning OOD methods and can cause large changes in the OOD quality metrics (i.e., AUC from 70.13% to 88.99% for MaxLogit (ML) for the CIFAR-10 vs SVHN task).

| Type | Closed-set ACC | OOD AUC | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | | KNN | Mah | ML | MSP | $LOF_C$ | $LOF_E$ | FE |
| CIFAR-10 vs SVHN | | | | | | | | |
| type-0 | 94.50 | **92.54** | 91.29 | 88.99 | 89.44 | 87.49 | 87.67 | 88.99 |
| type-1 | 93.64 | **93.61** | 90.07 | 80.05 | 85.32 | 91.36 | 91.78 | 79.90 |
| type-2 | 92.91 | 88.94 | 89.69 | 70.13 | 76.60 | **89.96** | 88.99 | 70.07 |
| CIFAR-10 vs CIFAR-100 | | | | | | | | |
| type-0 | 94.50 | 86.37 | 85.11 | 88.20 | 87.83 | 78.72 | 76.66 | **88.23** |
| type-1 | 93.64 | **88.58** | 87.91 | 86.78 | 86.52 | 86.82 | 86.05 | 86.81 |
| type-2 | 92.91 | **86.90** | 85.24 | 84.12 | 84.79 | 83.36 | 82.13 | 84.14 |

ples. In contrast, OOD subsets such as the SVHN train and the opposite CIFAR train include 72,000 and 50,000, respectively. This means that we can select different subsets for OOD. We measure how the selection of these samples affects the OOD performance. The results are shown in Table 4.

We observe that the effect of the sub-sampling of the OOD data from the same distribution is minimal. In the supplementary materials, we have also shown the table with the rankings, which are also stable and mainly independent of the selected subset. The result of the experiments confirms that there are easier or harder examples in the OOD data samples in the context of the OOD detection problem.

### 3.7 TRAIN-TEST SPLIT

We performed experiments with different train/test splits. In practice, we combined the original in-distribution subsets into one. Then we split them again with a different seed, but with the same ratio. Although this factor is not changed in most benchmarks, we decided to do this study to complement the study shown in the previous subsection (influence of the subset of OOD data). Moreover, in real-world OOD detection problems, the impact of splitting data into train/validation/test subsets needs to be more carefully noticed.

The results are shown in Table 5. Our study confirms earlier observations. Different training data imply building slightly different decision boundaries. The models perform well on the classical closed-set classification task (the standard deviation of accuracy is small, less than 0.5% for both models). However, using the same models for OOD detection problems shows much less stable evaluation metrics. For instance, the delta of AUC for CIFAR-100 vs. SVHN is greater than 10 for all OOD detectors; all four methods can be the best in ranking. For CIFAR-10 vs. SVHN again, all four different OOD detectors can be the first in the order. For CIFAR-10 vs. CIFAR-100, the effect is less observable. However, note that the standard deviation for AUC for features-based OOD detectors is around 4.5.

### 3.8 AUGMENTATION STRATEGIES

Next, we trained models using different augmentation strategies. We tested the following approaches using the Albumentations library Buslaev et al. [2020]: None, Affine, ColorJitter, CoarseDropout, CropAndPad, MixUp. We additionally used Normalize and HorizontalFlip(except None) for all of them.

The results are shown in Table 6. In contrast to the previous experiments, the changing data augmentation strategy affects both closed-set accuracy and OOD detection. This factor has a significant impact on how decision boundaries are constructed. We can see a wide range of AUCs, often above 20 percent. In addition, the ranking list is confused depending on the setting of this factor - the effect is especially noticeable for CIFAR-10 vs. SVHN. The higher closed-set accuracy does not guarantee better results in the OOD detection task, although the correlation is evident.

### 3.9 TEXT BASED OOD

We extended the empirical evaluation to text classification based on BERT representations. The [CLS] token was used

Table 3: Instability of OOD detection decisions as an effect of different random seeds used during training. We trained each of the analyzed architectures (ResNet and MobileNet) ten times on the same close set tasks (CIFAR-10 and CIFAR-100, respectively), but with different seeds of random number generators. We analyzed seven OOD methods (the names are explained in the caption of Table 2. In the table, we show the accuracy of the closed set (ACC) as the mean and standard deviation, the AUC metric in the same convention, and with its deltas (maximum minus minimum value). We also give the ranks of the OOD methods, where 0 means the best method and 6 means the worst, and the range of rank values obtained. We can see little variation in the accuracy of the closed set (std c.a. 0.3 p.p.), but large variation in OOD (up to 50 p.p. of AUC spread). As a result, the rank of the OOD method could be selected in almost any order just by peeking at the seed used during training.

| MobileNet with closed set ACC = 74.75±0.31 | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Method | CIFAR-100 vs SVHN | | | | CIFAR-100 vs CIFAR-10 | | | |
| | AUC | | Rank | | AUC | | Rank | |
| | mean±std | delta | mean±std | range | mean±std | delta | mean±std | range |
| KNN | 65.20±7.63 | 23.03 | 5.60±0.92 | 3-6 | 62.04±1.53 | 4.27 | 4.90±0.54 | 4-6 |
| Mah | 70.52±7.08 | 20.44 | 4.40±1.20 | 2-5 | 58.19±2.25 | 8.37 | 5.70±0.90 | 3-6 |
| ML | 82.47±4.60 | 17.37 | 1.70±1.19 | 1-5 | 77.73±0.54 | 1.92 | 0.00±0.00 | 0-0 |
| MSP | 78.11±4.21 | 15.06 | 3.60±1.02 | 2-6 | 76.53±0.35 | 1.03 | 2.00±0.00 | 2-2 |
| $LOF_C$ | 82.72±3.47 | 10.23 | 1.50±1.36 | 0-4 | 65.89±2.01 | 6.40 | 3.10±0.30 | 3-4 |
| $LOF_E$ | 76.23±3.88 | 11.50 | 3.50±1.20 | 1-6 | 63.46±1.39 | 5.23 | 4.30±0.46 | 4-5 |
| FE | 82.92±4.75 | 17.98 | 0.70±1.19 | 0-4 | 77.63±0.58 | 2.09 | 1.00±0.00 | 1-1 |
| ResNet with closed set ACC = 92.73±0.27 | | | | | | | | |
| Method | CIFAR-10 vs SVHN | | | | CIFAR-10 vs CIFAR-100 | | | |
| | AUC | | Rank | | AUC | | Rank | |
| | mean±std | delta | mean±std | range | mean±std | delta | mean±std | range |
| KNN | 76.84±14.89 | 50.02 | 3.50±2.01 | 0-6 | 64.27±7.68 | 24.13 | 3.90±0.83 | 3-6 |
| Mah | 78.30±6.29 | 23.01 | 4.60±0.80 | 3-6 | 63.94±6.84 | 21.53 | 4.00±0.77 | 3-5 |
| ML | 84.35±1.51 | 4.76 | 1.70±1.19 | 0-4 | 87.74±0.42 | 1.28 | 1.00±0.00 | 1-1 |
| MSP | 85.77±1.09 | 3.28 | 0.70±0.90 | 0-2 | 85.87±0.42 | 1.34 | 2.00±0.00 | 2-2 |
| $LOF_C$ | 70.80±6.30 | 22.39 | 5.70±0.46 | 5-6 | 57.76±3.38 | 9.07 | 5.70±0.90 | 3-6 |
| $LOF_E$ | 82.96±7.03 | 27.81 | 2.10±1.22 | 0-3 | 60.15±3.82 | 13.09 | 4.40±0.92 | 3-5 |
| FE | 84.27±1.52 | 4.77 | 2.70±1.19 | 1-5 | 87.77±0.42 | 1.29 | 0.00±0.00 | 0-0 |

Table 4: AUC scatter due to random selection of OOD images. We randomly selected 10,000 images (to balance ID and OOD data) 100 times from the OOD data set (out of 72,000 for SVHN and 50,000 among CIFAR-10 and CIFAR-100) and performed OOD detection using one of seven OOD methods (the names of the methods are explained in the caption of Table 2) in four OOD tasks. The results show that random selection of OOD examples can change the AUC by about 1 percentage point.

| Method | CIFAR-10 vs SVHN | | CIFAR-10 vs CIFAR-100 | | CIFAR-100 vs SVHN | | CIFAR-100 vs CIFAR-10 | |
|---|---|---|---|---|---|---|---|---|
| | mean±std | range | mean±std | range | mean±std | range | mean±std | range |
| KNN | 75.56±0.18 | 0.90 | 60.67±0.22 | 1.14 | 87.80±0.17 | 0.88 | 59.80±0.27 | 1.50 |
| Mah | 75.52±0.20 | 0.87 | 57.55±0.24 | 1.18 | 85.26±0.18 | 1.02 | 58.67±0.27 | 1.43 |
| ML | 81.72±0.15 | 0.80 | 76.88±0.18 | 0.85 | 83.88±0.12 | 0.63 | 87.73±0.13 | 0.69 |
| MSP | 78.25±0.18 | 0.87 | 76.09±0.16 | 0.96 | 86.69±0.11 | 0.53 | 86.20±0.13 | 0.66 |
| $LOF_C$ | 84.31±0.15 | 0.74 | 68.41±0.20 | 0.94 | 72.08±0.21 | 1.39 | 55.56±0.23 | 1.28 |
| $LOF_E$ | 78.87±0.18 | 0.97 | 64.31±0.22 | 1.12 | 91.40±0.12 | 0.63 | 62.73±0.22 | 1.09 |
| FE | 81.84±0.14 | 0.80 | 76.70±0.18 | 0.89 | 83.75±0.12 | 0.62 | 87.76±0.13 | 0.69 |

Table 5: Instability of OOD results for closed-set train-test split. We used the same architectures (ResNet for the CIFAR-10 data and MobileNet for CIFAR-100) and trained them 10 times, using different (random) train-test splits. Then, each trained model was applied to OOD detection. Again, we test seven different methods (their acronyms are explained in the caption of Table 2. As one can see, the closed-se accuracy (ACC) variance is small, but the OOD results,i.e., AUC and method ranks, differ greatly.

| MobileNet with closed set ACC = 74.98±0.50 | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Method | CIFAR-100 vs SVHN | | | | CIFAR-100 vs CIFAR-10 | | | |
| | AUC | | Rank | | AUC | | Rank | |
| | mean±std | delta | mean±std | range | mean±std | delta | mean±std | range |
| KNN | 71.33±4.27 | 12.83 | 5.30±1.19 | 3-6 | 62.10±1.52 | 5.08 | 4.20±0.75 | 3-5 |
| Mah | 76.94±6.36 | 22.20 | 4.00±1.84 | 0-6 | 55.78±2.49 | 8.67 | 6.00±0.00 | 6-6 |
| ML | 85.09±7.12 | 22.67 | 1.90±1.70 | 0-5 | 82.14±6.81 | 16.06 | 0.30±0.46 | 0-1 |
| MSP | 82.55±7.78 | 22.80 | 3.50±1.57 | 2-6 | 81.43±7.79 | 18.12 | 1.40±0.92 | 0-2 |
| $LOF_C$ | 84.35±3.49 | 10.49 | 2.10±1.30 | 0-4 | 65.17±3.81 | 11.40 | 3.20±0.40 | 3-4 |
| $LOF_E$ | 79.90±5.83 | 19.32 | 2.80±1.33 | 1-4 | 62.40±3.16 | 8.74 | 4.60±0.49 | 4-5 |
| FE | 85.31±7.13 | 22.62 | 1.40±1.80 | 0-5 | 82.03±6.77 | 15.99 | 1.30±0.46 | 1-2 |
| ResNet with closed set ACC = 94.41±0.24 | | | | | | | | |
| Method | CIFAR-10 vs SVHN | | | | CIFAR-10 vs CIFAR-100 | | | |
| | AUC | | Rank | | AUC | | Rank | |
| | mean±std | delta | mean±std | range | mean±std | delta | mean±std | range |
| KNN | 86.66±3.13 | 11.68 | 2.00±1.73 | 0-5 | 82.05±4.06 | 13.11 | 3.80±0.40 | 3-4 |
| Mah | 86.79±1.96 | 5.29 | 2.10±1.64 | 0-4 | 82.97±2.93 | 8.76 | 3.20±0.40 | 3-4 |
| ML | 86.27±2.74 | 8.76 | 2.40±0.80 | 1-4 | 88.53±0.50 | 1.54 | 1.00±0.00 | 1-1 |
| MSP | 87.02±2.04 | 7.68 | 1.80±1.72 | 0-5 | 87.94±0.33 | 1.13 | 2.00±0.00 | 2-2 |
| $LOF_C$ | 82.28±4.57 | 15.96 | 4.20±1.89 | 0-6 | 77.04±5.21 | 16.03 | 5.00±0.00 | 5-5 |
| $LOF_E$ | 75.94±3.85 | 13.93 | 5.90±0.30 | 5-6 | 67.27±5.92 | 19.08 | 6.00±0.00 | 6-6 |
| FE | 86.25±2.75 | 8.77 | 2.60±1.20 | 1-4 | 88.57±0.50 | 1.51 | 0.00±0.00 | 0-0 |

Table 6: Instability of OOD detection as a function of different augmentation methods used during training of the close-set model. We analyze six different augmentation methods, four OOD tasks, and present results (AUC) for seven OOD methods. We also report results for a close set (ACCs in column 2). The results show a large impact of augmentation techniques on the OOD results (the AUC varies by up to 20 percentage points). With SVHN as the OOD, almost any OOD method can be considered to be the best by choosing the appropriate augmentation method.

| Augmentation | ACC | KNN | Mah | ML | MSP | LOF$_C$ | LOF$_E$ | FE |
|---|---|---|---|---|---|---|---|---|
| MobileNet CIFAR-100 vs SVHN | | | | | | | | |
| None | 53.73 | 71.34 | 80.05 | 74.84 | 70.48 | **81.30** | 79.22 | 75.35 |
| Affine | 74.41 | 79.09 | 83.00 | 75.54 | 75.73 | 83.36 | **86.45** | 75.07 |
| CoarseDropout | 67.18 | 61.15 | 67.55 | 84.08 | 78.20 | 81.49 | 73.05 | **84.86** |
| ColorJitter | 65.98 | 61.56 | 67.05 | 81.70 | 75.55 | **84.99** | 80.10 | 82.49 |
| CropAndPad | 72.59 | 66.16 | 77.57 | 83.42 | 80.65 | **85.95** | 81.81 | 83.53 |
| MixUp | 68.65 | 67.33 | **86.59** | 75.90 | 76.11 | 86.13 | 85.97 | 72.92 |
| MobileNet CIFAR-100 vs CIFAR-10 | | | | | | | | |
| None | 53.73 | 56.20 | 54.21 | **66.89** | 66.51 | 56.96 | 48.81 | 66.67 |
| Affine | 74.41 | 61.98 | 57.03 | **78.96** | 77.55 | 67.19 | 61.96 | 78.84 |
| CoarseDropout | 67.18 | 61.09 | 57.40 | **72.55** | 72.14 | 61.96 | 56.69 | 72.37 |
| ColorJitter | 65.98 | 59.56 | 58.33 | **72.47** | 71.74 | 66.16 | 59.14 | 72.32 |
| CropAndPad | 72.59 | 62.97 | 56.92 | **76.32** | 75.47 | 65.07 | 60.10 | 76.19 |
| MixUp | 68.65 | 61.98 | 57.25 | **73.98** | 73.98 | 56.25 | 58.23 | 72.12 |
| ResNet CIFAR-10 vs SVHN | | | | | | | | |
| None | 83.64 | 78.88 | 74.15 | 80.14 | 79.25 | **82.07** | 75.67 | 80.10 |
| Affine | 94.76 | **92.65** | 91.01 | 91.89 | 90.49 | 91.68 | 88.08 | 91.97 |
| CoarseDropout | 89.27 | 61.16 | 60.26 | 87.68 | 85.59 | 69.31 | 78.85 | **87.74** |
| ColorJitter | 87.98 | 88.23 | 75.89 | 84.07 | 83.93 | **90.45** | 88.15 | 83.98 |
| CropAndPad | 94.01 | 91.52 | 88.39 | 93.32 | 90.98 | 90.90 | 86.83 | **93.45** |
| MixUp | 89.41 | 84.39 | **89.25** | 51.27 | 88.62 | 78.95 | 86.26 | 29.08 |
| ResNet CIFAR-10 vs CIFAR-100 | | | | | | | | |
| None | 83.64 | 67.50 | 66.41 | 80.98 | 78.52 | 61.68 | 56.69 | **81.10** |
| Affine | 94.76 | 88.23 | 87.60 | 89.77 | 88.92 | 86.86 | 84.94 | **89.81** |
| CoarseDropout | 89.27 | 59.36 | 55.06 | 85.75 | 83.45 | 56.16 | 55.06 | **85.81** |
| ColorJitter | 87.98 | 81.08 | 78.12 | 84.40 | 82.45 | 81.37 | 77.75 | **84.48** |
| CropAndPad | 94.01 | 87.17 | 85.03 | 88.76 | 87.93 | 83.71 | 79.37 | **88.81** |
| MixUp | 89.41 | 66.80 | 71.69 | 76.33 | **82.18** | 64.14 | 59.94 | 70.73 |

Table 7: Instability of OOD detection decisions as an effect of different random seeds used during training for text classification based on BERT (transformer based) representations. We used the AGNEWS dataset; the ID data are three selected classes (World, Sports, Business), and OOD data are texts from the Sci Tech class. We trained BERRT 8 times, but with different seeds of random number generators.

| BERT with closed set ACC = 97.49±0.11 | | | | |
|---|---|---|---|---|
| Method | AUC | | Rank | |
| | mean±std | delta | mean±std | range |
| KNN | 79.21±5.21 | 14.76 | 1.50±1.58 | 0-4 |
| Mah | 78.09±2.57 | 8.27 | 2.75±2.17 | 0-6 |
| ML | 72.56±7.54 | 23.16 | 4.25±1.71 | 1-6 |
| MSP | 71.06±11.79 | 35.68 | 3.75±2.11 | 0-6 |
| $LOF_C$ | 78.14±3.27 | 11.16 | 2.00±1.22 | 0-4 |
| $LOF_E$ | 77.79±3.22 | 11.04 | 2.75±0.97 | 1-4 |
| FE | 72.54±7.54 | 23.18 | 4.00±2.12 | 0-6 |

as a feature vector and a fully connected layer was added to BERT for classification. The whole network was tuned up on the data. Experiments were carried out on the AGNEWS dataset[4] with four subject classes of text documents; the ID data are three selected classes (World, Sports, Business), and OOD data are texts from the Sci Tech class. The instability of OOD detection due to the 8 different seeds used during training is shown in Table 7. The conclusions are the same as those reported in Section 3.4. We see a large variation in OOD (up to 35 p.p. spread in the AUC). As a result, the rank of the OOD method could be chosen in almost any order simply by looking at the seed used during training.

## 4 CONCLUSION

We showed that current benchmark studies for evaluating the performance of OOD detectors are susceptible to experimental details that significantly change the results. For instance, subtle changes in DNN model training (such as seed or number of training epochs) that do not affect the closed-set accuracy may dramatically change the performance numbers of OOD detectors. The highest instability (with AUC range ca. 20 up to 50 p.p.) is related to the training seed or architecture details; see Table 3 and Table 7 for CNN-based and transformer-based (BERT) representations, respectively. Hence, an OOD detector benchmarked on two DNN models trained with different seeds may evaluate as very successful or useless. These conclusions hold for OOD detection in the image and text domains and for both CNN and BERT features.

This instability issue can be explained by the different nature of the discriminative model (DNN as a closed-set classifier) and generative model (used by many OOD detectors). By retraining a DNN, we obtain a stable discriminative model (stable closed-set accuracy) but unstable generative models used by OOD detectors (such as Mahalanobis, LOF, or KNN). The instability of generative models of known ID classes is not surprising, considering the fact that they are built from scarce high-dimensional data. This effect is most clearly demonstrated for the ResNet model (with 2K-dimensional representations, the highest dimensionality considered in this work); see Table 3. In this study (ResNet with CIFAR-10 as in-distribution), generative models are built with 5K training samples per CIFAR-10 class in 2K-dimensional space. All OOD detectors working on generative models in the feature space, such as KNN, Mahalanobis, and LOF, show significantly higher instability (with delta ca. 20-50 and 10-20 for SVHN and CIFAR-100 as OOD data, respectively) than the logit-based OOD detectors, such as ML, MSP or FE (with delta values significantly smaller). Deeper insight into the nature and way to control the instability of the logit and feature-space OOD detectors is a matter for future work.

This work concludes that currently reported OOD performance measures should be considered unreliable since they depend on the subtle variations in the experiment. The main incentive of this work was to signal this issue and to identify the primary sources of variability in OOD detection experiments. The research community active in OOD detection needs to address this problem. The first step towards improvement is to realize the issues with the current benchmarks/procedures used to measure progress in OOD detection. However, providing a mature solution to control the variability in OOD experiments is a matter of future work.

Based on this work, we can formulate the following conclusions/recommendations for improving the quality of OOD experiments.

The first and most important conclusion is the postulate that all relevant experimental factors should be rigorously reported along with the results; otherwise, which is the common practice, comparing OOD detectors may lead to misleading and unreliable conclusions.

Secondly, multiple training runs should be recommended to assess the variance, with OOD detection performance metrics reported as confidence intervals rather than point results. This procedure is an undisputable experimental standard in empirical studies, e.g., natural sciences. The difficulty and some impracticality of this postulate in our field lie in the high cost of retraining representation generators for large-scale models. This issue deserves more careful attention in the machine learning community, and one of the goals of our work is to trigger this discussion.

[4] http://groups.di.unipi.it/~gulli/AG_corpus_of_news_articles.html

## References


Anja Belz, Shubham Agarwal, Anastasia Shimorina, and Ehud Reiter. A systematic review of reproducibility research in natural language processing. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 381–393, 2021.

Xavier Bouthillier, César Laurent, and Pascal Vincent. Unreproducible research is reproducible. In *International Conference on Machine Learning*, pages 725–734. PMLR, 2019.

Markus M Breunig, Hans-Peter Kriegel, Raymond T Ng, and Jörg Sander. Lof: identifying density-based local outliers. In *Proceedings of the 2000 ACM SIGMOD international conference on Management of data*, pages 93–104, 2000.

Alexander Buslaev, Vladimir I Iglovikov, Eugene Khvedchenya, Alex Parinov, Mikhail Druzhinin, and Alexandr A Kalinin. Albumentations: fast and flexible image augmentations. *Information*, 11(2):125, 2020.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. In Jill Burstein, Christy Doran, and Thamar Solorio, editors, *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 4171–4186. Association for Computational Linguistics, 2019. doi: 10.18653/v1/n19-1423.

Logan Engstrom, Andrew Ilyas, Shibani Santurkar, Dimitris Tsipras, Jacob Steinhardt, and Aleksander Madry. Identifying statistical bias in dataset replication. In *International Conference on Machine Learning*, pages 2922–2932. PMLR, 2020.

Maurizio Ferrari Dacrema, Simone Boglio, Paolo Cremonesi, and Dietmar Jannach. A troubling analysis of reproducibility and progress in recommender systems research. *ACM Transactions on Information Systems (TOIS)*, 39(2):1–49, 2021.

Odd Erik Gundersen, Kevin Coakley, and Christine Kirkpatrick. Sources of irreproducibility in machine learning: A review. *arXiv preprint arXiv:2204.07610*, 2022.

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*, pages 770–778, 2016.

Peter Henderson, Riashat Islam, Philip Bachman, Joelle Pineau, Doina Precup, and David Meger. Deep reinforcement learning that matters. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32, 2018.

Dan Hendrycks and Kevin Gimpel. A baseline for detecting misclassified and out-of-distribution examples in neural networks. In *International Conference on Learning Representations*, 2017.

Dan Hendrycks, Steven Basart, Mantas Mazeika, Mohammadreza Mostajabi, Jacob Steinhardt, and Dawn Song. Scaling out-of-distribution detection for real-world settings. *arXiv preprint arXiv:1911.11132*, 2019a.

Dan Hendrycks, Kevin Zhao, Steven Basart, Jacob Steinhardt, and Dawn Song. Natural adversarial examples. *arXiv preprint arXiv:1907.07174*, 2019b.

Dan Hendrycks, Nicholas Carlini, John Schulman, and Jacob Steinhardt. Unsolved problems in ml safety. *arXiv preprint arXiv:2109.13916*, 2021.

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 *International Conference on Machine Learning*, pages 8759–8773. PMLR, 2022.

Kimin Lee, Kibok Lee, Honglak Lee, and Jinwoo Shin. A simple unified framework for detecting out-of-distribution samples and adversarial attacks. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, NIPS'18, page 7167–7177, Red Hook, NY, USA, 2018. Curran Associates Inc.

Shiyu Liang, Yixuan Li, and Rayadurgam Srikant. Enhancing the reliability of out-of-distribution image detection in neural networks. *arXiv preprint arXiv:1706.02690*, 2017.

Weitang Liu, Xiaoyun Wang, John Owens, and Yixuan Li. Energy-based out-of-distribution detection. *Advances in neural information processing systems*, 33:21464–21475, 2020.

Xingjun Ma, Bo Li, Yisen Wang, Sarah M Erfani, Sudanthi Wijewickrema, Grant Schoenebeck, Dawn Song, Michael E Houle, and James Bailey. Characterizing adversarial subspaces using local intrinsic dimensionality. *arXiv preprint arXiv:1801.02613*, 2018.

Henryk Maciejewski, Tomasz Walkowiak, and Kamil Szyc. Out-of-distribution detection in high-dimensional data using mahalanobis distance-critical analysis. In *Computational Science–ICCS 2022: 22nd International Conference, London, UK, June 21–23, 2022, Proceedings, Part I*, pages 262–275. Springer, 2022.

Spyros Makridakis, Evangelos Spiliotis, and Vassilios Assimakopoulos. Statistical and machine learning forecasting methods: Concerns and ways forward. *PloS one*, 13(3): e0194889, 2018.

Rowan McAllister, Yarin Gal, Alex Kendall, Mark Van Der Wilk, Amar Shah, Roberto Cipolla, and Adrian Weller. Concrete problems for autonomous vehicle safety: Advantages of bayesian deep learning. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence*, IJCAI'17, page 4745–4753. AAAI Press, 2017. ISBN 9780999241103.

Giovanna Nicora, Miguel Rios, Ameen Abu-Hanna, and Riccardo Bellazzi. Evaluating pointwise reliability of machine learning prediction. *Journal of Biomedical Informatics*, 127:103996, 2022.

Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. Do cifar-10 classifiers generalize to cifar-10? *arXiv preprint arXiv:1806.00451*, 2018.

Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. Do imagenet classifiers generalize to imagenet? In *International conference on machine learning*, pages 5389–5400. PMLR, 2019.

Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. Mobilenetv2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4510–4520, 2018.

Chandramouli Shama Sastry and Sageev Oore. Detecting out-of-distribution examples with gram matrices. In *International Conference on Machine Learning*, pages 8491–8501. PMLR, 2020.

Yiyou Sun, Yifei Ming, Xiaojin Zhu, and Yixuan Li. Out-of-distribution detection with deep nearest neighbors. In *International Conference on Machine Learning*, pages 20827–20840. PMLR, 2022.

Fahim Tajwar, Ananya Kumar, Sang Michael Xie, and Percy Liang. No true state-of-the-art? ood detection methods are inconsistent across datasets. *arXiv preprint arXiv:2109.05554*, 2021.

Haoqi Wang, Zhizhong Li, Litong Feng, and Wayne Zhang. Vim: Out-of-distribution with virtual-logit matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4921–4930, 2022.

Jingkang Yang, Pengyun Wang, Dejian Zou, Zitang Zhou, Kunyuan Ding, Wenxuan Peng, Haoqi Wang, Guangyao Chen, Bo Li, Yiyou Sun, et al. Openood: Benchmarking generalized out-of-distribution detection. *arXiv preprint arXiv:2210.07242*, 2022.