Task Agnostic and Post-hoc Unseen Distribution Detection

Radhika Dua[1] Seongjun Yang[1] Yixuan Li[2] Edward Choi[1]

[1]KAIST [2]University of Wisconsin-Madison

{radhikadua, seongjunyang, edwardchoi}@kaist.ac.kr

sharonli@cs.wisc.edu

Organization. In the supplementary material, we provide:

- a detailed description of TAP-MOS, a task-agnostic extension of MOS used as a baseline in most of our experiments (Section 1).
- analysis of well-known OOD detection methods (Section 2).
- details on synthetic dataset generation (Section 3).
- discussion on the appropriateness of Mahalanobis distance, choice of GMM clustering in TAPUDD and tuning of hyperparameters (Section 4).
- extended description of experimental settings (Section 5).
- additional results on a synthetic dataset for binary classification, results on conventional OOD datasets for multi-class classification tasks, and results on anomaly detection task (Section 6).
- quantitative results with other metrics, including AUPR, FPR95 (Section 7).

1. Task Agnostic and Post-hoc MOS (*TAP-MOS*)

We present an extension of MOS [9] which was proposed for OOD detection in large-scale classification problems. Since we aim to present a baseline that does not rely on the label space, we develop a clustering-based OOD detection method, Task Agnostic and Post-hoc Minimum Others Score (*TAP-MOS*), in the features space. The training datasets' features extracted from a model trained for a specific task are given as input to the *TAP-MOS* module. *TAP-MOS* module partition the features of in-distribution data into K clusters using Gaussian Mixture Model (GMM) with "*full*" covariance and train a cluster classification model. Motivated by the success of MOS, we perform group based learning and form K groups, $\mathcal{G}_1$, $\mathcal{G}_2$, ..., $\mathcal{G}_\text{K}$, where each group $\mathcal{G}_\text{k}$ comprises of samples of cluster k. A new category "*others*" is then introduced in each group $\mathcal{G}_\text{k}$. The class labels in each group are re-assigned during the training of cluster classification model. "*Others*" class is defined as the ground-truth class for the groups that do not include cluster c. Following MOS [9], we calculate the group-wise softmax for each group $\mathcal{G}_\text{k}$ as:

$$p_c^k(\mathbf{x}) = \frac{e^{f_c^k(\mathbf{x})}}{\sum_{c' \in \mathcal{G}_\text{i}} e^{f_{c'}^k(\mathbf{x})}}, c \in \mathcal{G}_\text{k}, \tag{1}$$

where $f_c^k(\mathbf{x})$ and $p_c^k(\mathbf{x})$ represent the output logit and softmax probability of a class c in group $\mathcal{G}_\text{k}$, respectively.

The training objective for cluster classification is the sum of cross-entropy losses across all the groups:

$$\mathcal{L} = -\frac{1}{N}\sum_{n=1}^{N}\sum_{k=1}^{K}\sum_{c \in \mathcal{G}_\text{k}} y_c^k \log\left(p_c^k(\mathbf{x})\right), \tag{2}$$

where y_c^k and $p_c^k(\mathbf{x})$ denote the re-assigned class labels and the softmax probability of category c in $\mathcal{G}_k$, and N denotes the total number of training samples.

For OOD detection, we utilize the *Minimum Others Score* (MOS) which uses the lowest "*others*" score among all groups to distinguish between ID and OOD samples and is defined as:

$$\mathcal{S}_{\textit{TAP-MOS}} = -\min_{1 \le k \le K} p_{others}^k(\mathbf{x}). \tag{3}$$

To align with the conventional notion of having high score for ID samples and low score for OOD samples, the negative sign is applied. We hypothesize that *TAP-MOS* will fail when samples are near or away from the periphery of all clusters because it performs One-vs-All classification and detect a sample as OOD only if it is detected as "*others*" class by all groups. We validate our hypothesis by conducting experiments on the synthetic datasets in Section 4.1 of the main paper.

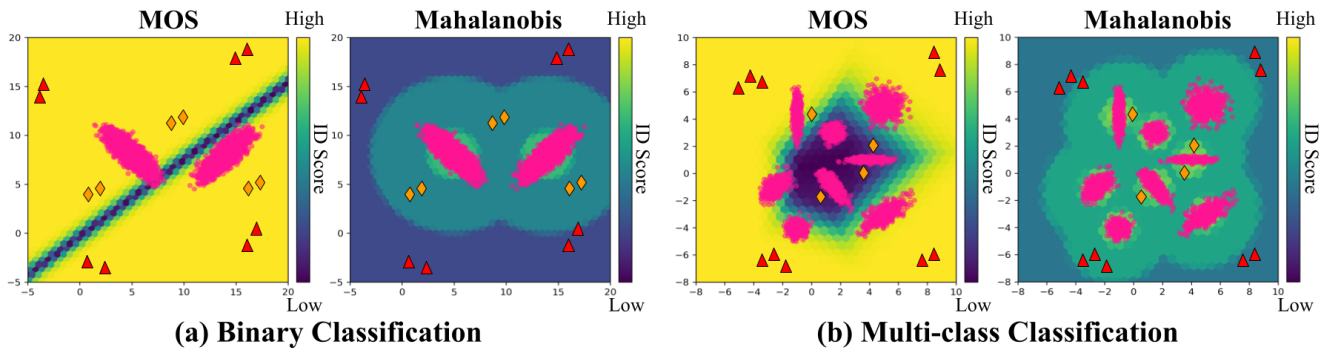


Figure 1. ID score landscape of the existing representative post-hoc OOD detection methods (Mahalanobis, and MOS) on synthetic 2D binary and multi-class classification datasets. A sample is regarded as OOD when it has a **low ID score**. The **Pink Points** represent the in-distribution data; **Red Triangles** and **Orange Diamonds** represents OOD samples. Results demonstrate that MOS fails to detect samples near or away from the periphery of all classes (*e.g.*, **Red Triangles**) and Mahalanobis fails to detect samples near the ID classes (*e.g.*, **Orange Diamonds**) as OOD.

2. Analysis of Well-known OOD Detection Methods

MOS. MOS [9] performs One-vs-All classification and detects a sample as OOD only if it is detected as the "*others*" class by all groups. We hypothesize that MOS will fail to detect samples near or away from the periphery of all classes as OOD. This is because the groups in the corner will detect samples on the same side of the decision boundary of the one vs. all groups classification as ID. We now conduct an experiment on the 2D synthetic datasets for binary and multi-class classification tasks to determine if our hypothesis holds true. More details on synthetic datasets are provided in Section 3. Fig. 3a and Fig. 3b presents the ID score landscape of MOS in binary class and multi-class classification tasks respectively. The **Pink Points** represent the in-distribution data; **Red Triangles** and **Orange Diamonds** represents OOD samples. Results demonstrate that MOS works very well when the OOD samples are in between multiple classes (can be seen in blue color in 2-D plane). However, it fails to detect samples near or away from the corner classes (denoted by **Red Triangles**).

Mahalanobis. Mahalanobis OOD detector [11] approximates each class as multi-variate gaussian distribution with tied covariance. However, in reality, the features can be correlated differently in different classes. In particular, features of a few classes can be correlated positively, and features of some other classes might be correlated negatively. We hypothesize that Mahalanobis might fail to detect OOD samples near the ID classes in such cases. We conduct an experiment in the same manner as above to test our hypothesis. Fig. 3a and Fig. 3b presents the ID score landscape of Mahalanobis in binary class and multi-class classification tasks respectively. Results demonstrate that Mahalanobis works very well when the OOD samples are located far away from the ID classes but it fails to detect samples located near the ID classes (denoted by **Orange Diamonds**).

SSD. SSD [18] is an outlier detector based on unlabeled in-distribution data which utilizes self-supervised representation learning followed by Mahalanobis distance based OOD detection. In self-supervised representation learning, SSD uses $NT-Xent$ loss function from SimCLR [2] which utilizes multiple data augmentation techniques. SSD has achieved remarkable OOD detection performance and even outperforms several supervised OOD detection methods. However, we hypothesize that when OOD samples are

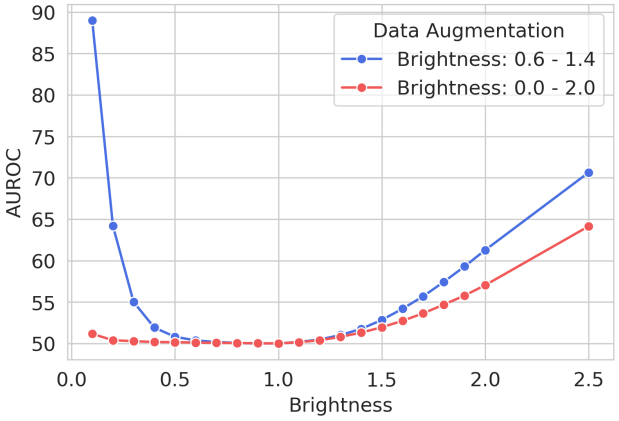


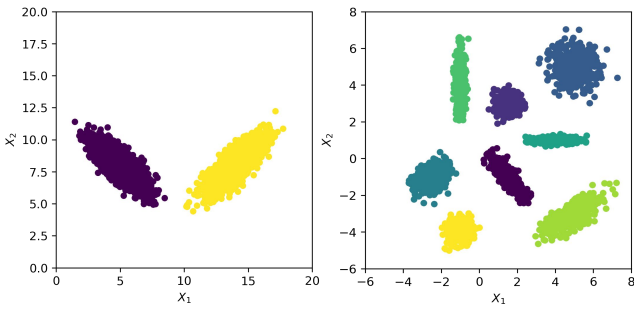
Figure 2. Impact of varying intensity of brightness in data augmentation on the OOD detection performance of SSD.

from the same distribution as the augmented data, the model might fail to detect them as OOD. This can be problematic in several real-world tasks. For instance, in 3D vision, it is desirable to detect the shapes rotated by more than some threshold. However, if we used the rotation technique in data augmentation, SSD might fail to detect the samples as OOD. This can also lead to severe consequences in safety-critical applications. We conduct an experiment to determine if our hypothesis holds true. We used the CIFAR-10

dataset for self-supervised training in SSD and used data augmentation similar to SSD and varied the range of intensity of brightness used in data augmentation. Then, we evaluate the SSD model in the NAS setup. More specifically, we shift the brightness of CIFAR-10 samples with varying levels of intensities and evaluate the performance of SSD when trained with augmentation of brightness. Fig. 2 presents the impact of using different intensities of brightness in data augmentation on the performance of SSD for NAS detection. We observe that when the brightness intensity from 0.6 to 1.4 is used in data augmentation, the SSD model fails to detect samples from these brightness intensities as OOD. Further, when the brightness intensity from 0.0 to 2.0 is used in data augmentation, the SSD model even fails to detect extraordinarily dark and light images. This demonstrates that SSD fails to detect OOD samples from the same distribution as the augmented data. Similar to SSD, OOD detection methods that utilizes data augmentation for self-supervised learning might fail in scenarios where the model encounters OOD samples from distribution same as the distribution of augmented dataset. Therefore, we do not compare our approach against such OOD detection methods in all the experiments.

We observe that MOS, Mahalanobis, and SSD do not perform well in certain scenarios. Moreover, MOS and Mahalanobis OOD detection methods require the class label information of the training datasets. Therefore, they cannot be directly used for OOD detection in tasks other than classification. This motivates the necessity of an unsupervised OOD detection method that is not only task-agnostic and architecture-agnostic but also addresses the scenarios where MOS, Mahalanobis, and SSD do not perform well.

3. Synthetic Dataset Generation



(a) Binary Classification dataset (b) Multi-class Classification

Figure 3. 2-D synthetic datasets for (a) binary classification and (b) multi-class classification tasks.

We generate synthetic datasets in $\mathbb{R}^2$ for binary and multi-class classification tasks, as shown in Fig. 3. The in-distribution (ID) data $\mathbf{x} \in \mathcal{X} = \mathbb{R}^2$ is sampled from a Gaussian mixture model. All the samples except the ID samples in the 2-D plane represent the OOD samples. Dataset

Dataset	Cluster	Mean (μ_x, μ_y)	Covariance
Binary	Cluster 1	(5.0, 8.0)	([1.0, -0.8], [-0.8, 1.0])
	Cluster 2	(14.0, 8.0)	([1.0, 0.8], [0.8, 1.0])
Multi-class	Cluster 1	(1.5, -1.0)	([0.2, -0.3], [-0.2, 0.2])
	Cluster 2	(1.5, 3.0)	([0.1, 0.0], [0.0, 0.1])
	Cluster 3	(5.0, 5.0)	([0.4, 0.0], [0.0, 0.4])
	Cluster 4	(-2.5, -1.0)	([0.1, 0.2], [0.2, 0.1])
	Cluster 5	(4.0, 1.0)	([0.4, 0.0], [0.0, 0.01])
	Cluster 6	(-1.0, 4.3)	([0.02, 0.0], [0.0, 0.7])
	Cluster 7	(5.0, -3.0)	([0.5, 0.4], [0.3, 0.1])
	Cluster 8	(-1.0, -4.0)	([0.1, 0.0], [0.0, 0.1])

Table 1. Mean and covariance per cluster used for generating dataset for binary and multi-class classification tasks.

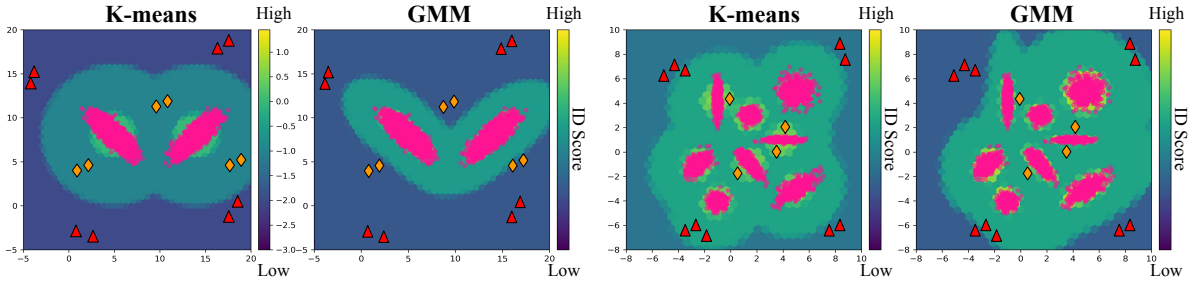
for binary classification and multi-class classification tasks comprises of 2 and 8 clusters, respectively. For each cluster with mean (μ_x, μ_y) and covariance, 3000 and 500 data points are sampled in binary and multi-class classification tasks, respectively. More details on the mean and covariance of each cluster is provided in Table 1.

4. Discussion

Appropriateness of Mahalanobis distance (MD). Given density estimation in high-dimensional space is a known intractable problem, we view MD as a reasonable approximation that leads to empirical efficacy. Moreover, we believe that MD in TAPUDD is safe since the ensembling module does a ***reasonably simple approximation*** by aggregating the MD obtained from GMM with a different number of clusters. Evaluating the compatibility of our test-time framework on methods trained with added regularization to explicitly make Mahalanobis distance more appropriate for OOD detection [13] can be an interesting future direction to explore.

Reason for using GMM. We aim to use Mahalanobis distance to measure the distance between a test sample and local training set clusters in the latent space, hence GMM with full covariance is a natural fit. We compare GMM with K-means in Fig. 4 and observe that GMM is flexible in learning the cluster shape in contrast to K-means, which learned spherical cluster shapes. Consequently, K-means performs poorly when detecting OOD samples near the cluster. Other popular clustering methods such as agglomerative clustering or DBSCAN are less compatible with Mahalanobis distance and require careful hyperparameter adjustment, such as the linking strategies for agglomerative clustering or the epsilon value for DBSCAN.

Tuning of Hyperparameters. Although it is possible to tune hyperparameters K and n_e, our experiments indicate there is very little need to tune them. We observe that our approach can effectively detect OOD samples across different tasks and datasets as long as K consists of a sufficient number of diverse clusters (approximately 12) and n_e is equal to more than half of the number of participants in K.



(a) Binary Classification **(b) Multi-class Classification**

Figure 4. ID score landscape of TAP-Mahalanobis on synthetic 2D binary and multi-class classification datasets on using K-means and GMM with full covariance. A sample is deemed as OOD when it has a **low ID score**. The **Pink Points** represent the in-distribution data; **Red Triangles** and **Orange Diamonds** represent OOD samples. Results demonstrate that on using K-means, TAP-Mahalanobis fails to detect OOD samples near the clusters (*e.g.*, **Orange Diamonds**). However, on using GMM with full covariance, Mahalanobis effectively detects all OOD samples.

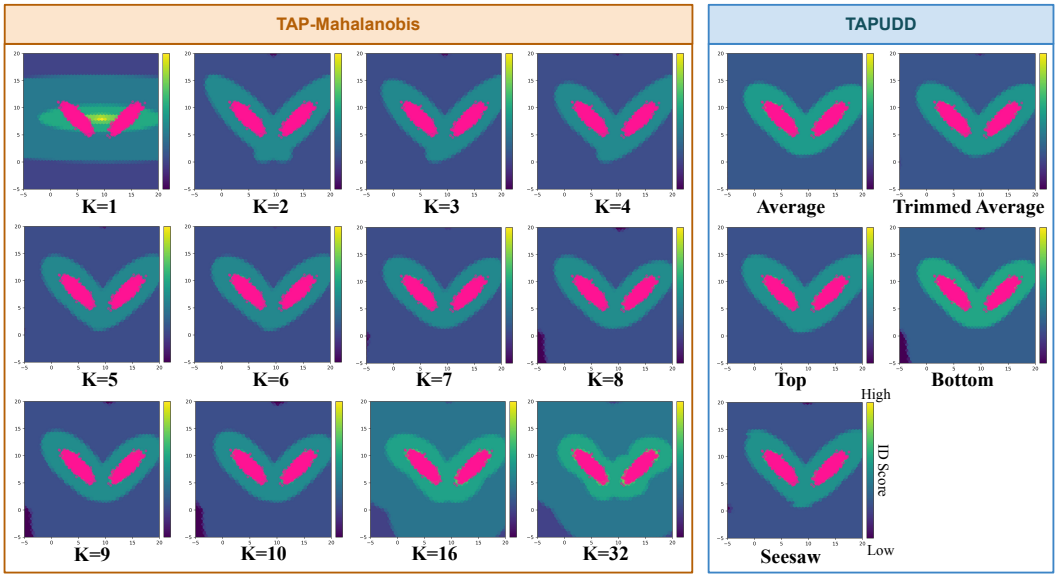


Figure 5. ID score landscape of TAP-Mahalanobis for different values of K (*i.e.*, number of clusters); and TAPUDD for different ensemble variations on synthetic 2D binary classification dataset. A sample is regarded as OOD when it has a **low ID score**. The **Pink Points** represent the in-distribution data. Results demonstrate that TAP-Mahalanobis does not perform well for some values of K whereas TAPUDD with all ensembling strategies perform better or on-par with TAP-Mahalanobis.

We used $K = [1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 16, 32]$ and $n_e = 8$ and observed that the same hyperparameters can be used to obtain good OOD detection performance in different tasks and datasets.

5. Experimental Details

For binary classification and regression tasks, we use the RSNA Bone Age dataset [5], a real-world dataset that contains 12611 left-hand X-ray images of the patient, along with their gender and age (0 to 20 years). We randomly split the dataset in 8:1:1 ratio to form train, val, and test split with 9811, 1400, and 1400 samples, respectively. Following [17], to reflect diverse X-ray imaging set-ups in the hospital, we vary the brightness factor of the test set between 0 and 6.5 and form 20 different NAS datasets. In-distribution data comprises images with a brightness factor of 1.0 (unmodified images).

Binary Classification. We use a ResNet18 [6] model, pretrained on ImageNet [3], and add two fully-connected layers containing 128 and 2 hidden units with a relu activation. We train the network to classify gender given the x-ray image. Each model is trained for 100 epochs using SGD optimizer with a learning rate of 0.001 and momentum of 0.9, using a batch size of 64.

Regression. We use a ResNet18 [6] model that is pretrained on ImageNet [3] and train it to predict the age given the x-ray image. After the average pooling layer, we add two fully-connected layers with 128 and 1 units with a relu

OOD Dataset	Baselines						Ours (Task-Agnostic)	
	MSP [8]	ODIN [12]	Energy [14]	MB [11]	KL [7]	Gram [1]	TAP-MB (K = 8)	TAPUDD (Average)
LSUN (R)	91.0	94.1	92.8	99.7	70.3	99.9	96.3	96.4
LSUN (C)	91.9	91.2	93.9	96.7	81.9	97.8	94.5	94.2
TinyImgNet (R)	91.0	94.0	92.4	99.5	73.8	99.7	93.8	94.3
TinyImgNet (C)	91.4	93.1	93.0	98.6	74.1	99.2	94.3	94.5
SVHN	89.9	96.7	91.2	99.1	85.5	99.5	92.8	93.4
CIFAR100	86.4	85.8	87.1	88.2	69.2	79.0	88.2	88.9

Table 2. Comparison of OOD Detection Performance of Resnet34 model trained on CIFAR10 on diverse OOD datasets measured by AUROC. The hyperparameters of ODIN and the hyperparameters and parameters of Mahalanobis are tuned using a random sample of the OOD dataset. MB and TAP-MB refers to Mahalanobis and TAP-Mahalanobis, respectively.

Method	**Network**	**Airplane**	**Automobile**	**Bird**	**Cat**	**Deer**	**Dog**	**Frog**	**Horse**	**Ship**	**Truck**	**Average**
VAE	-	70.4	38.6	67.9	53.5	74.8	52.3	68.7	49.3	69.6	38.6	58.4
OCSVM	-	63.0	44.0	64.9	48.7	73.5	50.0	72.5	53.3	64.9	50.8	58.6
AnoGAN	DCGAN	67.1	54.7	52.9	54.5	65.1	60.3	58.5	62.5	75.8	66.5	61.8
PixelCNN	PixelCNN	53.1	99.5	47.6	51.7	73.9	54.2	59.2	78.5	34.0	66.2	61.8
DSVDD	LeNet	61.7	65.9	50.8	59.1	60.9	65.7	67.7	67.3	75.9	73.1	64.8
OCGAN	OCGAN	75.7	53.1	64.0	62.0	72.3	62.0	72.3	57.5	82.0	55.4	65.6
Reconstruction Error	Resnet18	71.5	39.2	68.7	55.9	72.6	54.4	63.3	49.1	71.3	37.6	58.4
MB (K = 1)	Resnet18	69.2	66.9	66.3	52.3	74.8	50.7	77.8	52.9	64.0	52.1	62.7
TAP-MB (K = 1)	Resnet18	69.2	66.8	66.2	52.3	74.8	50.7	77.8	52.9	64.0	52.1	62.7
TAP-MB (K = 2)	Resnet18	73.0	67.3	66.5	50.7	74.4	49.9	74.7	52.4	66.2	52.6	62.8
TAP-MB (K = 3)	Resnet18	73.4	68.0	66.9	54.9	74.4	50.2	76.1	51.7	61.3	52.3	**62.9**
TAP-MB (K = 4)	Resnet18	68.1	64.1	63.6	56.6	73.5	48.7	75.3	49.5	54.8	48.6	60.3

Table 3. Comparison of TAP-Mahalanobis with other detectors for anomaly detection task on CIFAR-10 dataset. TAP-MB and MB denotes TAP-Mahalanobis and Mahalanobis, respectively.

activation. Each model is trained for 100 epochs using SGD optimizer with a learning rate of $1e-05$, weight decay of 0.0001, and momentum of 0.9, using a batch size of 64. We also apply gradient clipping with a clip value of 5.0.

The results are measured by computing mean and standard deviation across 10 trials upon randomly chosen seeds. We perform all experiments on NVIDIA GeForce RTX A6000 GPUs.

6. Additional Results

6.1. Evaluation on Synthetic Datasets

TAPUDD outperforms TAP-Mahalanobis. We present a comparison of TAPUDD against TAP-Mahalanobis on 2-D synthetic dataset for binary classification task, in continuation to the discussion in Section 4.1. Fig. 5 presents the ID score landscape of TAP-Mahalanobis for different values of K and TAPUDD with different ensemble variations for binary classification in a 2-D synthetic dataset. The **Pink Points** represent the in-distribution data. We observe that for certain values of K, TAP-Mahalanobis fails to detect some OOD samples. However, all ensemble variations of TAPUDD effectively detect OOD samples and performs better, or on par, with TAP-Mahalanobis. Thus, TAPUDD eliminates the necessity of choosing the optimal value of K.

6.2. OOD Detection in Multi-class Classification

As stated in Section 4 of the main paper, we also evaluate our approach for OOD detection in multi-class classification task on benchmark datasets to further bolster the effectiveness of our proposed approach. We use the pre-trained ResNet34 [6] model trained on CIFAR-10 dataset (opensourced in [11]). We consider the test set as the in-distribution samples and evaluate our approach on diverse OOD datasets used in literature (TinyImagenet, LSUN [19], SVHN [16] and CIFAR100). Table 2 presents the OOD detection performance of our approach and baselines based on AUROC score. We observe that our task-agnostic and post-hoc approach performs better or comparable to the baselines.

6.3. Anomaly Detection

We evaluate our approach for anomaly detection task in which one of the CIFAR-10 classes is considered as in-distribution and the samples from rest of the classes are

considered as anomalous. We train a Resnet18 based auto-encoder model using MSE loss function which aims to minimize the reconstruction error between the output and input image. Since, in this work, we provide a task-agnostic and post-hoc approach to detect samples from unseen distribution, we consider that we are given a model trained on one of the classes CIFAR10 and our objective is to detect anomalous samples (*i.e.*, samples from other classes of CIFAR10). We first compare our approach with two baselines that does not modify the base architecture. Reconstruction-error based baseline which rely on reconstruction error to determine if a sample is anomalous or not. Mahalanobis distance based detector with number of classes as 1 to detect anomalous samples. Further, we also compare our approach with various well-known baselines for anomaly detection, including VAE, OCSVM, AnoGAN, PixelCNN, DSVDD, OCGAN. Although it is unfair to compare with these baselines as they have different base models, we compare against these baselines since they are used widely in literature. We do not compare with other baselines including CSI, SSD as they might fail in certain scenarios (described in Section 2 of the Appendix). Table 3 presents a comparison of TAP-Mahalanobis with other detectors for anomaly detection task on CIFAR-10 dataset. We observe that our task-agnostic and post-hoc approach is better than reconstruction-error based baseline. Our approach is also better than or on-par with other baselines used in the literature. This demonstrates the effectiveness of our approach on anomaly detection task.

7. Quantitative Results with Different Performance Metrics

Method	**iNaturalist**	**SUN**	**Places**	**Textures**	**Average**
Expected	Low	Low	Low	**High**	–
MSP [8]	97.26	94.41	94.12	95.65	95.36
ODIN [12]	97.80	96.23	95.33	96.11	96.37
Mahalanobis [11]	87.35	90.32	90.25	92.52	90.11
Energy [14]	97.62	96.55	95.47	96.04	96.42
KL Matching [7]	97.98	94.11	93.62	97.96	95.92
MOS [9]	99.62	98.17	97.36	96.68	97.96
TAPUDD (Average)	**91.87**	**91.02**	90.08	**99.68**	**93.16**

Table 4. OOD detection performance comparison between TAPUDD method and baselines measured by AUPR. Ideally, all methods should follow the expected results obtained from our analysis (described in first row in green color) conducted in Section 4.4 of the main paper. However, as highlighted in green color, only Mahalanobis and our proposed approach follow the expected results. This highlights the failure of existing baselines, including MSP, ODIN, Energy, KL Matching, and MOS. Further, amongst all methods following the expected results (highlighted in green color), *our approach is highly sensitive to OOD samples and significantly outperforms the baselines*.

We report additional metrics to evaluate the unseen distribution detection performance of baselines and our approach in binary classification, regression, and large-scale classification tasks. In Table 5 and Table 6, we compare the NAS detection performance of baselines and our approach in binary classification task based on AUPR and FPR95, respectively. We also report the NAS detection performance of baselines and our method in regression task based on AUPR and FPR95 in Table 7 and Table 8, respectively. Results demonstrate that our proposed approaches, TAPUDD and TAP-Mahalanobis are more sensitive to NAS samples compared to competitive baselines. Further, we report AUPR to evaluate the OOD detection performance of different methods in large-scale classification task Table 4. As expected from the analysis conducted in Section 4.4 of the main paper, the results indicate that our approach detects samples from the Textures dataset as more OOD compared to samples from iNaturalist, SUN, and Places (similar to the way humans perceive).

References

[1] Sastry Shama Chandramouli and Oore Sageev. Detecting out-of-distribution examples with gram matrices. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2020.

[2] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey E. Hinton. A simple framework for contrastive learning of visual representations. *ArXiv*, abs/2002.05709, 2020.

[3] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, K. Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 248–255, 2009.

[4] Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2016.

[5] Safwan S. Halabi, Luciano M. Prevedello, Jayashree Kalpathy-Cramer, Artem B. Mamonov, Alexander Bilbily, Mark Cicero, Ian Pan, Lucas Araújo Pereira, Rafael Teixeira Sousa, Nitamar Abdala, Felipe Campos Kitamura, Hans H. Thodberg, Leon Chen, George Shih, Katherine Andriole, Marc D. Kohli, Bradley J. Erickson, and Adam E. Flanders. The rsna pediatric bone age machine learning challenge. *Radiology*, 290(3):498–503, 2019.

[6] Kaiming He, X. Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016.

[7] Dan Hendrycks, Steven Basart, Mantas Mazeika, Mohammadreza Mostajabi, Jacob Steinhardt, and Dawn Xiaodong Song. A benchmark for anomaly segmentation. *ArXiv*, abs/1911.11132, 2019.

[8] Dan Hendrycks and Kevin Gimpel. A baseline for detecting misclassified and out-of-distribution examples in neural net-

Brightness	Baselines							Ours (Task-Agnostic)		
	MSP [8]	ODIN [12]	Energy [14]	MB [11]	KL [7]	MOS [9] (K = 2)	Gram [1]	TAP-MOS (K = 2)	TAP-MB (K = 2)	TAPUDD (Average)
0.0	95.6±3.5	95.6±3.5	95.2±3.2	100.0±0.0	62.6±29.8	96.0±3.3	99.4±1.0	83.9±13.0	100.0±0.0	100.0±0.0
0.2	69.5±3.2	69.4±3.3	69.2±3.6	88.4±4.1	45.5±2.0	69.1±3.3	74.3±1.7	67.3±6.4	88.9±3.7	88.8±4.8
0.4	59.5±2.4	58.3±2.5	58.1±2.9	70.3±6.2	47.3±0.9	58.3±2.2	71.0±0.8	57.1±4.3	70.8±5.0	71.8±6.0
0.6	55.6±1.8	53.5±1.7	53.4±1.7	59.5±3.4	48.9±0.8	53.5±1.7	70.2±0.8	52.6±2.7	59.6±2.9	60.5±3.5
0.8	53.4±1.5	50.7±0.8	50.6±0.8	51.7±1.0	50.1±0.8	50.6±0.8	69.8±0.8	50.5±1.3	51.7±0.9	52.1±1.0
1.0	52.8±1.4	50.0±0.0	50.0±0.0	50.0±0.0	50.0±0.0	50.0±0.0	69.7±0.8	50.0±0.0	50.0±0.0	50.0±0.0
1.2	54.1±1.4	51.8±0.6	51.8±0.6	54.2±1.1	49.3±0.5	51.8±0.5	69.9±0.8	51.5±1.2	54.5±1.1	54.5±1.1
1.4	57.4±1.3	55.6±1.0	55.6±1.0	60.9±1.9	48.1±0.8	55.8±0.9	70.7±0.6	54.3±2.2	61.4±2.2	61.6±2.3
1.6	61.2±1.7	60.1±1.8	60.2±1.9	68.2±3.2	46.4±1.1	60.3±1.3	71.9±0.6	57.8±3.6	69.1±3.7	69.2±3.7
1.8	65.1±2.4	64.5±2.6	64.7±2.8	74.8±4.4	45.7±1.5	64.8±1.8	73.3±0.6	61.5±5.7	75.9±4.6	75.9±4.9
2.0	68.4±3.3	68.1±3.5	68.2±3.9	80.5±5.2	46.0±2.1	68.4±2.5	74.5±0.8	64.8±8.0	81.8±4.9	81.5±5.3
2.5	74.8±5.1	74.7±5.2	74.8±5.7	90.8±4.0	45.5±2.5	75.2±3.7	77.3±1.6	70.0±8.3	91.5±4.2	91.0±4.6
3.0	80.5±4.5	80.5±4.5	80.4±5.2	95.5±2.3	43.9±2.1	81.0±3.2	80.1±2.1	73.0±7.5	95.8±2.9	95.4±3.2
3.5	82.4±4.3	82.4±4.3	82.3±4.9	97.1±1.9	44.1±2.9	82.8±3.3	81.7±3.0	73.9±8.7	97.1±2.6	96.8±2.7
4.0	84.0±4.2	84.0±4.2	83.8±4.8	97.8±1.5	43.2±3.0	84.2±3.6	83.5±3.8	74.3±9.7	97.8±2.3	97.6±2.3
4.5	85.2±3.7	85.2±3.7	85.1±4.2	98.4±0.9	42.1±3.0	85.5±3.2	84.8±3.7	74.5±10.1	98.3±1.9	98.2±1.7
5.0	86.1±3.3	86.1±3.4	86.0±3.9	98.8±0.6	41.3±2.8	86.4±2.7	86.0±3.6	74.7±10.3	98.7±1.4	98.7±1.1
5.5	86.8±3.4	86.8±3.4	86.6±3.9	99.1±0.4	40.5±2.4	87.1±2.6	87.0±3.4	74.9±10.5	99.0±1.1	99.0±0.8
6.0	87.3±3.4	87.3±3.4	87.2±3.9	99.2±0.3	40.3±2.4	87.6±2.6	88.1±3.4	75.0±10.6	99.1±0.8	99.2±0.6
6.5	87.6±3.3	87.6±3.3	87.5±3.8	99.3±0.3	40.5±2.5	87.9±2.5	88.8±3.5	75.0±10.8	99.3±0.7	99.3±0.5
Average	72.4	71.6	71.5	81.7	46.1	71.8	78.6	65.8	82.0	**82.1**

Table 5. NAS detection performance in binary classification task (gender prediction) for NAS shift of brightness in RSNA boneage dataset measured by AUPR. Highlighted row presents the performance on in-distribution dataset. MB and TAP-MB refers to Mahalanobis and TAP-Mahalanobis, respectively.

Brightness	Baselines							Ours (Task-Agnostic)		
	MSP [8]	ODIN [12]	Energy [14]	MB [11]	KL [7]	MOS [9] (K = 2)	Gram [1]	TAP-MOS (K = 2)	TAP-MB (K = 2)	TAPUDD (Average)
0.0	60.0±49.0	60.0±49.0	60.0±49.0	0.0±0.0	60.0±49.0	60.0±49.0	10.0±30.0	80.0±40.0	0.0±0.0	0.0±0.0
0.2	90.2±1.5	90.2±1.5	90.9±2.2	68.3±13.3	91.4±1.4	90.1±1.4	86.3±6.1	92.3±1.2	62.7±15.2	59.8±15.2
0.4	93.0±1.0	93.0±1.0	93.4±1.4	90.3±3.8	93.4±0.9	93.1±1.2	92.4±1.6	93.6±1.0	88.9±4.4	87.6±4.7
0.6	94.1±0.5	94.1±0.5	94.1±0.8	93.9±1.3	93.9±1.2	94.2±0.7	93.8±0.7	94.6±0.4	93.7±1.4	93.0±1.6
0.8	94.4±0.4	94.4±0.4	94.4±0.6	95.3±0.6	94.2±1.5	94.6±0.6	94.6±0.4	95.1±0.5	95.4±0.7	95.1±0.5
1.0	95.0±0.0	95.0±0.0	95.0±0.0	95.0±0.0	95.0±0.0	95.0±0.0	95.0±0.0	95.0±0.0	95.0±0.0	95.0±0.0
1.2	94.8±0.6	94.8±0.6	94.6±0.6	94.0±0.6	94.0±1.7	94.8±0.7	94.7±0.8	94.9±0.6	93.6±0.6	93.5±0.7
1.4	93.5±0.9	93.5±0.9	93.1±0.9	91.0±1.4	93.1±1.6	93.3±1.0	92.9±1.4	94.0±0.7	90.3±2.1	89.7±2.4
1.6	92.0±1.0	92.0±1.0	91.6±1.0	85.4±1.7	91.9±1.7	91.7±1.4	90.2±2.0	93.6±1.1	84.1±3.3	83.2±3.5
1.8	90.2±1.2	90.2±1.2	89.8±1.2	78.6±3.0	91.1±1.3	90.3±1.3	87.6±2.7	93.2±2.1	76.9±5.1	75.5±5.2
2.0	88.4±1.8	88.4±1.8	87.9±1.8	71.5±4.6	89.7±1.7	88.3±1.4	83.9±3.8	92.8±2.9	69.3±6.9	67.4±7.9
2.5	85.7±3.5	85.7±3.5	85.2±3.6	49.6±10.2	88.0±2.5	85.3±3.4	78.1±7.0	91.4±3.9	47.3±13.9	45.7±14.8
3.0	84.2±3.8	84.1±3.8	83.5±4.3	30.1±12.7	86.8±2.8	83.9±3.7	73.8±8.2	90.8±4.5	28.7±17.2	28.0±17.8
3.5	82.8±5.5	82.8±5.5	82.4±6.3	19.9±14.1	86.0±3.9	82.8±5.1	70.3±9.1	90.7±5.6	19.1±17.6	19.3±17.8
4.0	82.3±6.6	82.3±6.6	81.7±7.3	13.4±11.9	85.7±4.8	82.1±6.0	67.8±10.1	90.4±7.1	13.3±15.4	14.4±15.3
4.5	81.7±7.5	81.7±7.5	81.1±7.9	8.0±8.1	85.1±5.1	81.6±6.3	65.6±10.0	91.0±7.1	9.2±12.3	10.0±11.5
5.0	81.0±7.4	81.0±7.4	80.6±7.9	5.0±5.2	84.8±5.0	80.9±6.6	63.5±9.6	91.2±7.3	6.0±9.2	7.1±8.4
5.5	80.7±7.9	80.7±7.9	80.5±8.3	3.2±3.8	85.0±5.1	80.9±6.6	61.5±10.2	91.5±7.6	4.4±7.1	5.0±6.0
6.0	80.2±8.5	80.2±8.5	80.0±8.4	2.3±3.0	84.7±5.6	80.5±7.2	59.7±10.7	91.5±8.2	3.1±5.2	3.7±4.2
6.5	79.8±8.6	79.8±8.6	79.7±9.2	1.9±2.7	84.4±5.7	80.3±7.5	58.7±11.4	91.6±8.2	2.4±4.0	2.7±3.0
Average	86.2	86.2	86.0	49.8	87.9	86.2	76.0	92.0	49.2	**48.8**

Table 6. NAS detection performance in binary classification task (gender prediction) for NAS shift of brightness in RSNA boneage dataset measured by FPR95. Highlighted row presents the performance on in-distribution dataset. MB and TAP-MB refers to Mahalanobis and TAP-Mahalanobis, respectively.

Brightness	Baselines			Ours (Task-Agnostic)		
	DE [10]	MC Dropout [4]	SWAG* [15]	TAP-MOS (K = 8)	TAP-Mahala (K = 8)	TAPUDD (Average)
0.0	100.0±NA	34.9±NA	100.0±NA	74.4±20.8	99.8±0.4	100.0±0.0
0.2	53.9±NA	48.4±NA	51.4±NA	69.2±17.2	89.6±12.8	89.6±6.1
0.4	50.0±NA	51.0±NA	49.4±NA	69.2±16.9	75.6±15.8	68.1±4.8
0.6	50.1±NA	50.4±NA	49.2±NA	64.3±12.3	58.0±8.4	56.3±2.8
0.8	50.2±NA	50.1±NA	49.8±NA	57.3±6.4	51.5±2.3	50.2±0.9
1.0	50.0±NA	49.7±NA	50.0±NA	50.0±0.0	50.0±0.0	50.0±0.0
1.2	50.4±NA	48.7±NA	50.8±NA	49.3±3.6	50.2±0.5	56.2±1.1
1.4	53.6±NA	47.9±NA	54.8±NA	51.0±6.8	51.1±1.2	65.2±2.5
1.6	55.5±NA	46.7±NA	62.1±NA	55.2±10.1	52.0±2.1	75.1±3.0
1.8	60.3±NA	45.4±NA	74.0±NA	61.5±13.5	53.0±3.1	83.2±4.2
2.0	69.9±NA	43.9±NA	83.2±NA	67.2±15.9	55.0±4.6	89.3±4.2
2.5	94.4±NA	40.1±NA	92.3±NA	76.8±16.0	61.0±9.3	96.3±2.0
3.0	98.4±NA	37.4±NA	92.7±NA	83.4±13.2	64.8±11.7	98.5±0.7
3.5	99.3±NA	35.6±NA	94.8±NA	88.5±9.7	68.5±13.2	99.1±0.4
4.0	99.8±NA	34.3±NA	97.2±NA	90.8±6.6	71.3±13.3	99.4±0.3
4.5	100.0±NA	33.4±NA	98.0±NA	91.5±4.6	73.8±12.3	99.6±0.3
5.0	100.0±NA	32.4±NA	98.6±NA	91.4±4.2	77.1±10.7	99.6±0.3
5.5	100.0±NA	32.0±NA	98.9±NA	90.7±4.9	80.2±9.0	99.6±0.3
6.0	100.0±NA	31.7±NA	99.0±NA	89.7±5.8	82.6±8.3	99.6±0.4
6.5	100.0±NA	31.5±NA	99.2±NA	88.6±6.7	84.3±7.9	99.6±0.5
Average	76.8	41.3	77.3	73.0	67.5	**83.7**

Table 7. NAS detection performance in regression task (age prediction) for NAS shift of brightness in RSNA boneage dataset measured by AUPR. Highlighted row presents the performance on in-distribution dataset. DE and TAP-MB denotes Deep Ensemble and TAP-Mahalanobis, respectively. SWAG* = SWAG + Deep Ensemble.

Brightness	Baselines			Ours (Task-Agnostic)		
	DE [10]	MC Dropout [4]	SWAG* [15]	TAP-MOS (K = 8)	TAP-Mahala (K = 8)	TAPUDD (Average)
0.0	0.0±NA	100.0±NA	0.0±NA	80.0±42.2	0.0±0.0	0.0±0.0
0.2	91.9±NA	99.2±NA	94.0±NA	74.9±21.6	51.2±27.4	58.9±18.7
0.4	94.6±NA	96.1±NA	94.7±NA	71.4±23.4	69.1±18.4	90.5±2.5
0.6	94.7±NA	95.6±NA	95.0±NA	84.0±9.6	88.9±5.4	94.4±1.0
0.8	95.0±NA	95.6±NA	95.1±NA	91.8±2.8	94.1±1.5	95.4±0.4
1.0	95.0±NA	94.5±NA	95.0±NA	95.0±0.0	95.0±0.0	95.0±0.0
1.2	94.7±NA	95.5±NA	94.1±NA	95.2±1.2	94.5±1.4	92.6±2.1
1.4	89.2±NA	95.6±NA	93.7±NA	92.3±5.6	93.8±2.2	86.0±5.8
1.6	78.8±NA	97.5±NA	88.6±NA	87.4±9.8	92.9±3.1	74.9±9.7
1.8	69.7±NA	98.7±NA	87.5±NA	79.8±16.8	90.3±5.3	62.1±14
2.0	53.3±NA	99.1±NA	81.7±NA	73.5±24.1	88.2±7.0	49.2±18.8
2.5	14.9±NA	100±NA	60.3±NA	63.2±31.1	81.4±12.6	23.6±16.1
3.0	6.9±NA	100±NA	53.6±NA	54.5±29.9	76.7±15.4	8.8±6.8
3.5	2.8±NA	100.0±NA	35.2±NA	43.5±25.6	73.0±17.9	3.9±3.0
4.0	0.8±NA	100.0±NA	20.2±NA	36.6±20.2	69.6±19.0	2.4±2.1
4.5	0.0±NA	100.0±NA	13.4±NA	33.7±17.2	67.3±17.6	1.3±1.6
5.0	0.0±NA	100.0±NA	8.2±NA	32.8±17.5	65.4±16.2	1.2±1.9
5.5	0.1±NA	100.0±NA	5.7±NA	34.6±20.2	61.6±16.0	1.6±3.3
6.0	0.0±NA	100.0±NA	4.5±NA	37.7±23.0	57.7±16.6	2.1±4.5
6.5	0.0±NA	100.0±NA	3.1±NA	41.8±25.9	54.1±17.6	2.5±5.5
Average	44.1	98.4	56.2	65.2	73.2	**42.3**

Table 8. NAS detection performance in regression task (age prediction) for NAS shift of brightness in RSNA boneage dataset measured by FPR95. Highlighted row presents the performance on in-distribution dataset. DE and TAP-MB denotes Deep Ensemble and TAP-Mahalanobis, respectively. SWAG* = SWAG + Deep Ensemble.

works. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2017.

[9] Rui Huang and Yixuan Li. Mos: Towards scaling out-of-distribution detection for large semantic space. *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8706–8715, 2021.

[10] Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.

[11] 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 (NeurIPS)*, 2018.

[12] Shiyu Liang, Yixuan Li, and Rayadurgam Srikant. Enhancing the reliability of out-of-distribution image detection in neural networks. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2018.

[13] Juncheng Liu, Zhouhui Lian, Yi Wang, and Jianguo Xiao. Incremental kernel null space discriminant analysis for novelty detection. In *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 792–800, 2017.

[14] Weitang Liu, Xiaoyun Wang, John Owens, and Yixuan Li. Energy-based out-of-distribution detection. *Advances in Neural Information Processing Systems (NeurIPS)*, 33, 2020.

[15] Wesley J Maddox, Pavel Izmailov, Timur Garipov, Dmitry P Vetrov, and Andrew Gordon Wilson. A simple baseline for bayesian uncertainty in deep learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 13153–13164, 2019.

[16] Yuval Netzer, Tao Wang, Adam Coates, A. Bissacco, Bo Wu, and A. Ng. Reading digits in natural images with unsupervised feature learning. 2011.

[17] Jeonghoon Park, Jimin Hong, Radhika Dua, Daehoon Gwak, Yixuan Li, Jaegul Choo, and E. Choi. Natural attribute-based shift detection. *ArXiv*, abs/2110.09276, 2021.

[18] Vikash Sehwag, Mung Chiang, and Prateek Mittal. {SSD}: A unified framework for self-supervised outlier detection. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2021.

[19] Fisher Yu, Yinda Zhang, Shuran Song, Ari Seff, and Jianxiong Xiao. Lsun: Construction of a large-scale image dataset using deep learning with humans in the loop. *ArXiv*, abs/1506.03365, 2015.