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

Abstract

Despite the recent advances in out-of-distribution(OOD) detection, anomaly detection, and uncertainty estimation tasks, there do not exist a task-agnostic and post-hoc approach. To address this limitation, we design a novel clustering-based ensembling method, called **T***ask* **A***gnostic and* **P***ost-hoc* **U***nseen* **D***istribution* **D***etection (TAPUDD) that utilizes the features extracted from the model trained on a specific task. Explicitly, it comprises of* TAP-Mahalanobis*, which clusters the training datasets' features and determines the minimum Mahalanobis distance of the test sample from all clusters. Further, we propose the* Ensembling module *that aggregates the computation of iterative* TAP-Mahalanobis *for a different number of clusters to provide reliable and efficient cluster computation. Through extensive experiments on synthetic and real-world datasets, we observe that our task-agnostic approach can detect unseen samples effectively across diverse tasks and performs better or on-par with the existing task-specific baselines. We also demonstrate that our method is more viable even for large-scale classification tasks.*

1. Introduction

Deep neural networks have achieved phenomenal performance in diverse domains such as computer vision and healthcare [3, 35, 12]. However, they struggle to handle samples from an unseen distribution, leading to unreliable predictions and fatal errors in safety-critical applications. In an ideal situation, a robust model should be capable of making predictions on samples from the learned distributions, and at the same time, flag unknown inputs from unfamiliar distributions so that humans can make a responsible decision. For instance, in safety-critical tasks such as cancer detection, the machine learning assistant must issue a warning and hand over the control to the doctors when it detects an unusual sample that it has never seen during training. Thus, in practice, it is important for a model to know when *not* to predict. This task of detecting samples from an unseen distribution is referred to as out-of-distribution (OOD) detection [20, 29, 27, 5, 30, 24, 19, 44, 51, 34].

Most of these OOD detection methods mainly focusing on classification tasks have shown great success. However, they are not directly applicable to other tasks like regression. Although a few bayesian and non-bayesian techniques [13, 26, 33, 16] estimate uncertainty in regression tasks, they are not post-hoc as it often requires a modification to the training pipeline, or multiple trained copies of the model, or training a model with an optimal dropout rate. This raises an under-explored question:

> *Can we design a task-agnostic, and post-hoc approach for unseen distribution detection ?*

Motivated by this, we propose a novel clustering-based ensembling framework, "**T**ask **A**gnostic and **P**ost-hoc **U**nseen **D**istribution **D**etection (TAPUDD)", which comprises of two modules, *TAP-Mahalanobis* and *Ensembling*. *TAP-Mahalanobis* partitions the training datasets' features into clusters and then determines the minimum Mahalanobis distance of a test sample from all the clusters. The *Ensembling* module aggregates the outputs obtained from *TAP-Mahalanobis* iteratively for a different number of clusters. It enhances reliability and eliminates the need to determine an optimal number of clusters. Since TAPUDD is a post-hoc approach and doesn't require training the model, it is more efficient and easy to deploy in real-world.

To demonstrate the efficacy of our approach, we conduct experiments on 2-D synthetic datasets for binary and multi-class classification tasks and observe that our method effectively detects the outliers in both tasks. Further, we extensively evaluate our approach on real-world datasets for diverse tasks. In particular, we evaluate our approach for binary classification (gender prediction) and regression (age prediction) task on RSNA boneage dataset to detect the samples shifted by brightness. We observe that our method successfully identifies the shifted samples. We also evaluate our approach on large-scale classification tasks and obtained logical performance on diverse OOD datasets. To sum up, our contributions include:

- We propose a novel task-agnostic and post-hoc approach, **TAPUDD**, to detect unseen samples across diverse tasks like classification, regression, etc.
- For the first time, we empirically show that a single approach can be used for multiple tasks with stable performance. We conduct exhaustive experiments on synthetic and real-world datasets for regression, binary classification, and multi-class classification tasks to demonstrate the effectiveness of our method.
- We conduct ablation studies to illustrate the effect of number of clusters in *TAP-Mahalanobis* and ensembling strategies in TAPUDD. We observe that TAPUDD performs better or on-par with *TAP-Mahalanobis* and eliminates the necessity to determine the optimal number of clusters.

2. Related Work

To enhance the reliability of deep neural networks, there exist several efforts along the following research directions:

Out-of-distribution Detection. Recent works have introduced reconstruction-error based [43, 53, 9, 41, 40, 7], density-based [42, 6, 34, 11, 15, 37, 45], and self-supervised [14, 21, 1, 44] OOD detection methods. Other efforts include post-hoc methods [27, 30, 20, 29, 5, 19, 36] that do not require modification to the training procedure. However, there is no approach that is post-hoc and does not require the class label information of the training data.

Uncertainty Estimation. Research in this direction primarily estimates the uncertainty to enhance the robustness of networks in regression tasks. Well-known methods to estimate uncertainty include bayesian [32, 38, 2, 16, 31, 18, 28, 25, 50, 46, 33] and non-bayesian [13, 26] approaches, which have shown remarkable success. However, they require significant modification to the training pipeline, multiple trained copies of the model, and are not post-hoc.

Anomaly Detection. This task aims to detect anomalous samples shifted from the defined normality. Prior work [40, 49, 14, 43, 4, 53, 9] proposed methods to solve anomaly detection. However, more recently, [44, 48, 1, 22] proposed a unified method to solve both OOD detection and anomaly detection. Nonetheless, these methods require end-to-end training and are not post-hoc.

There exist no unified approach to enhance the reliability of neural networks across distinct tasks like classification, regression, etc. In contrast to all the aforementioned efforts, our work presents a post-hoc, and task-agnostic approach to detect unknown samples across varied tasks.

3. Background and Methodology

3.1. Problem Formulation

We assume that the in-distribution data $\mathcal{D}_{\text{IN}} = \{X, Y\}$ is composed of N i.i.d. data points with inputs $X = \{\mathbf{x}_1, ..., \mathbf{x}_N\}$ and labels $Y = \{y_1, ..., y_N\}$. Specifically, $\mathbf{x}_i \in \mathbb{R}^d$ represents a d-dimensional input vector, and y_i is its corresponding label. For classification problems, the label y_i is one of the C classes. For regression problems, the label is real-valued, that is $y_i \in \mathbb{R}$. For autoencoding (or self-supervised or unsupervised learning tasks), the label y_i is absent. Let $f : X \rightarrow \mathcal{Z}$, where $\mathcal{Z} \in \mathbb{R}^m$, denote the feature extractor often parameterized by a deep neural network which maps a sample from the d-dimensional input space (X) to the m-dimensional feature space ($\mathcal{Z}$).

Our goal is to obtain the feature of a sample from a trained DNN using the feature extractor (f), and equip it with an OOD detector which can detect samples from different distribution than the training distribution (OOD samples) or samples shifted from the training distribution based on some attribute [39]. We wish to design a task-agnostic, and post-hoc OOD detector which only requires the features of training data obtained from the trained model. Since such OOD detector does not require the raw training data, it minimizes privacy risk, a significant advantage for tasks where data-related privacy is a serious concern.

3.2. Background

Mahalanobis distance-based OOD detection method. Mahalanobis OOD detector [27] approximates each class as multi-variate gaussian distribution and use the minimum Mahalanobis distance of a test sample from all classes for OOD detection. Given density estimation in high-dimensional space is a known intractable problem, Mahalanobis approach is viewed as a reasonable approximation that has been widely adopted. In particular, it fits the gaussian distribution to the features of the training data $X = \{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_N, y_N)\}$ and compute per-class mean $\mu_c = \frac{1}{N_c} \sum_{i:y_i=c} f(\mathbf{x}_i)$ and a shared covariance matrix $\Sigma = \frac{1}{N} \sum_{c=1}^{C} \sum_{i:y_i=c} (f(\mathbf{x}_i) - \mu_c)(f(\mathbf{x}_i) - \mu_c)^T$, where $f(\mathbf{x}_i)$ denotes the penultimate layer features of an input sample $\mathbf{x}_i$, C denotes the total number of classes in the training dataset, and N_c denotes the total number of samples with class label c. Given a test sample $\mathbf{x}_{test}$, the mahalanobis score is defined as:

$$\mathcal{S}_{\text{Mahalanobis}} = -\min_c (f(\mathbf{x}_{test}) - \mu_c)^T \Sigma^{-1} (f(\mathbf{x}_{test}) - \mu_c),$$

where $f(\mathbf{x}_{test})$ denotes the penultimate layer features of a test sample $\mathbf{x}_{test}$.

3.3. TAPUDD: Task Agnostic and Post-hoc Unseen Distribution Detection

We propose a novel, **T**ask **A**gnostic and **P**ost-hoc **U**nseen **D**istribution **D**etection (TAPUDD) method, as shown in Fig. 1. The method comprises of two main modules *TAP-Mahalanobis* and *Ensembling*.

TAP-Mahalanobis. Given training samples $X = \{\mathbf{x}_1, ..., \mathbf{x}_N\}$, we extract the features of the in-distribution

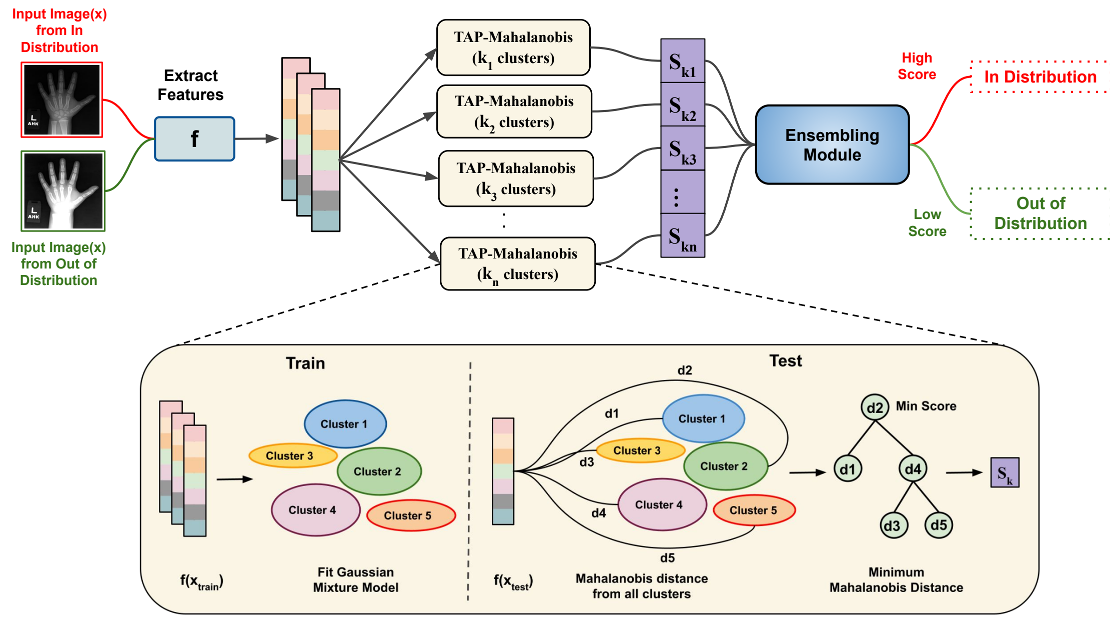

Figure 1. **TAPUDD.** Our method first extracts the features of an input image $\mathbf{x}$ from the feature extractor f of a model trained on a specific task. *TAP-Mahalanobis* module then uses the extracted features $f(\mathbf{x}_{train})$ to fit the gaussian mixture model and computes the minimum mahalanobis distance S_k for the given feature vector $f(\mathbf{x}_{test})$. Further, the *Ensembling* module aggregates the mahalanobis distance (S_{k1} to S_{kn}) obtained from iterative computation of *TAP-Mahalanobis* for different number of clusters (k_1 to k_n) to enhance the reliability.

data from a model trained for a specific task using a feature extractor f. We then pass these features to the *TAP-Mahalanobis* module. It first partition the features of the in-distribution data into K clusters using Gaussian Mixture Model (GMM) with "*full*" covariance. Then, we model the features in each cluster independently as multivariate gaussian and compute the empirical cluster mean and covariance of training samples $X = \{\mathbf{x}_1, ..., \mathbf{x}_N\}$ and their corresponding cluster labels $C = \{c_1, ..., c_N\}$ as:

$$\mu_c = \frac{1}{N_c}\sum_{i:c_i=c} f(\mathbf{x}_i), \Sigma_c = \frac{1}{N_c}\sum_{i:c_i=c}(f(\mathbf{x}_i)-\mu_c)(f(\mathbf{x}_i)-\mu_c)^T,$$

where $f(\mathbf{x}_i)$ denotes the penultimate layer features of an input sample $\mathbf{x}_i$ from a cluster c_i.

Then, given a test sample, $\mathbf{x}_{test}$, we obtain the negative of the minimum of the Mahalanobis distance from the center of the clusters as follows:

$$\mathcal{S}_{\text{TAP-Mahalanobis}} = -\min_c (f(\mathbf{x}_{test}) - \mu_c)^T \Sigma_c^{-1} (f(\mathbf{x}_{test}) - \mu_c),$$

where $f(\mathbf{x}_{test})$ denotes the penultimate layer features of a test sample $\mathbf{x}_{test}$. We then use the score $\mathcal{S}_{\text{TAP-Mahalanobis}}$ to distinguish between ID and OOD samples. To align with the conventional notion of having high score for ID samples and low score for OOD samples, negative sign is applied.

However, it is not straightforward to determine the value of the number of clusters K for which the OOD detection performance of *TAP-Mahalanobis* is optimal for different tasks and datasets. To illustrate the effect of using a different number of clusters K in *TAP-Mahalanobis* on the OOD detection performance, we conduct an ablation in Sec. 4.5. To this end, we present an *Ensembling* module.

Ensembling. This module not only eliminates the need to determine the optimal value of K but also provides more reliable results. We obtain *TAP-Mahalanobis* scores for different values of $K \in [k_1, k_2, k_3, ..., k_n]$ and aggregate them to obtain an ensembled score, $\mathcal{S}_{\text{Ensemble}}$. This ensures that a sample is detected as OOD only if a majority of the participants in ensembling agrees with each other.

Remark. We empirically compare GMM with K-means and observe that GMM is more 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. Please refer to Sec. 4 of the Appendix for a detailed discussion on GMM clustering.

3.4. Ensembling Strategies

Given the dependency of TAPUDD on the aggregation of *TAP-Mahalanobis* scores obtained for different values of K (*i.e.*, number of clusters) , a natural question arises: *how do different ensembling strategies affect the performance of unseen distribution detection*? To answer this, we systematically consider the following five ensembling strategies:

Average. We consider $\mathcal{S}_{\text{TAP-Mahalanobis}}$ obtained from *TAP-Mahalanobis* module for different K with equal importance and average them to obtain an ensembled score, $\mathcal{S}_{\text{Ensemble}}$.

Trimmed Average. For certain values of K, the *TAP-*

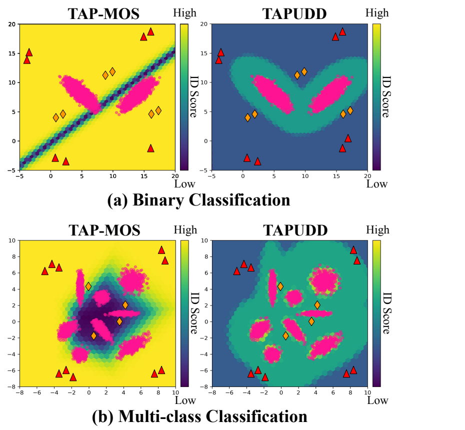

Figure 2. ID score landscape of TAP-MOS and TAPUDD on synthetic 2-D binary and multi-class classification datasets. A sample is deemed as OOD when it has a **low ID score**. The **Pink Points** represent the in-distribution data; **Red Triangles** and **Orange Diamonds** represent the far and near OOD samples, respectively. TAP-MOS fails to detect certain OOD samples, whereas TAPUDD effectively detects all OOD samples.

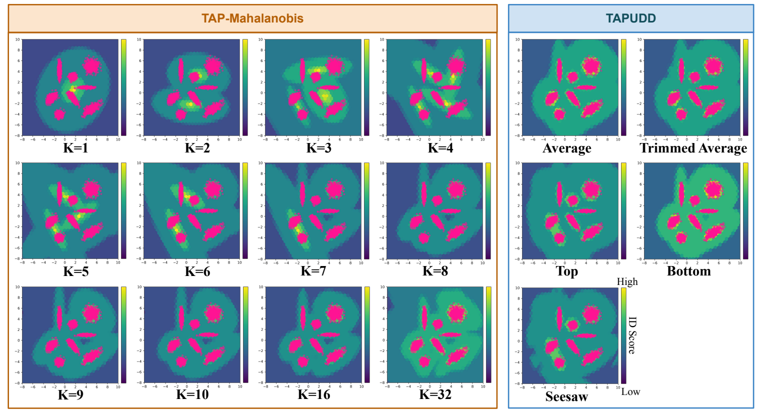

Figure 3. ID score landscape of *TAP-Mahalanobis* for different values of K (*i.e.*, number of clusters); and TAPUDD for different ensemble variations on synthetic 2D multi-class classification dataset. A sample is deemed as OOD when it has a **low ID score**. The **Pink Points** represent the in-distribution 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*.

Mahalanobis module can provide extremely high or extremely low $\mathcal{S}_{\text{TAP-Mahalanobis}}$. Therefore, to reduce bias caused by extreme *TAP-Mahalanobis* scores, we eliminate "m" *TAP-Mahalanobis* scores from top and bottom and then take the average of the remaining $n_e > K/2$ *TAP-Mahalanobis* scores to obtain a final ensembled score.

Seesaw. In general, the voting ensemble in regression tasks includes the average of all participants. However, some participants might completely disagree with the other participants, and including all of them in ensembling might provide inaccurate results. To this end, we present "*seesaw*" ensembling strategy wherein we sort the *TAP-Mahalanobis* scores obtained for different values of K and average the $n_e > K/2$ participants that agree with each other. In other words, if a majority of participants agree to a high *TAP-Mahalanobis* score, we pick the top n_e participants; otherwise, we select the bottom n_e participants.

Top. We sort the *TAP-Mahalanobis* scores obtained for different values of K and then obtain the ensembled score $\mathcal{S}_{\text{Ensemble}}$ by averaging the top $n_e > K/2$ scores.

Bottom. We sort the *TAP-Mahalanobis* scores obtained for different values of K and average the bottom $n_e > K/2$ scores to obtain an ensembled score $\mathcal{S}_{\text{Ensemble}}$.

4. Experiments and Results

In this section, we validate our approach, TAPUDD, by conducting experiments on 2-D synthetic datasets (Sec. 4.1). To further bolster the effectiveness of our method, we present empirical evidence to validate TAPUDD on several real-world tasks, including binary classification (Sec. 4.2), regression (Sec. 4.3), and large-scale classification (Sec. 4.4). For binary classification and regression task, we evaluate our approach on Natural Attribute-based Shift (NAS) detection dataset. In NAS detection[39], a sample is shifted from the training distribution based on attributes like brightness, age, etc. Throughout all experiments, we use TAP-MOS, a task-agnostic and post-hoc extension of MOS[24], as an additional baseline. Unlike the original MOS where class hierarchy was used to do clustering, we perform GMM clustering. More details on TAP-MOS are provided in Sec. 1 of the Appendix.

4.1. Evaluation on Synthetic Datasets

Experimental Details. We generate synthetic datasets in $\mathbb{R}^2$ for binary classification and multi-class classification tasks. The in-distribution (ID) data $\mathbf{x} \in \mathcal{X} = \mathbb{R}^2$ is sampled from a Gaussian mixture model (refer to Sec. 3 of the Appendix for more details). All the samples except the ID samples in the 2-D plane represent the OOD samples. We consider the 2-D sample as the penultimate layer features on which we can directly apply OOD detection methods like TAPUDD, *TAP-Mahalanobis*, and TAP-MOS.

TAPUDD outperforms TAP-MOS. We compare the OOD detection performance of TAP-MOS and TAPUDD. Fig. 2a and Fig. 2b presents the ID score landscape of TAP-MOS and TAPUDD for binary classification and multi-class classification, respectively. The **Pink Points** represent the in-distribution data; **Red Triangles** and **Orange Diamonds** represent far and near OOD samples, respectively. Here for TAP-MOS, we present results using number of clusters as 2 and 8 in binary classification and multi-class classification, respectively. For TAPUDD, we present the results of the "*average*" ensembling strategy computed across the num-

Brightness	Baselines							Ours (Task-Agnostic)		
	MSP [20]	ODIN [29]	Energy [30]	MB [27]	KL [19]	MOS [24] (K = 2)	Gram [5]	TAP-MOS (K = 2)	TAP-MB (K = 2)	TAPUDD (Average)
0.0	91.4±6.6	91.4±6.6	90.8±6.0	100.0±0.0	44.0±44.1	92.3±6.3	98.9±2.0	71.7±21.7	100.0±0.0	100.0±0.0
0.2	67.1±2.6	67.1±2.6	66.7±2.8	86.4±4.4	46.0±3.8	66.9±2.4	62.4±3.7	63.0±6.3	87.0±4.5	86.7±5.6
0.4	57.8±1.8	57.9±1.8	57.6±2.2	68.7±5.5	47.4±1.4	57.8±1.6	54.0±1.1	55.4±3.7	69.0±4.6	69.5±5.8
0.6	53.5±1.2	53.5±1.3	53.4±1.4	58.8±3.0	48.6±1.0	53.4±1.2	51.6±0.4	52.0±1.9	58.9±2.5	59.6±3.1
0.8	50.9±0.6	50.9±0.6	50.9±0.6	51.5±1.1	49.8±1.0	50.9±0.6	50.4±0.3	50.4±0.8	51.5±0.9	51.9±1.1
1.0	50.0±0.0	50.0±0.0	50.0±0.0	50.0±0.0	50.0±0.0	50.0±0.0	50.0±0.0	50.0±0.0	50.0±0.0	50.0±0.0
1.2	51.8±0.4	51.8±0.4	51.9±0.4	54.5±1.0	49.4±0.5	51.9±0.4	50.8±0.5	51.3±0.9	54.8±1.2	54.6±1.1
1.4	55.9±0.7	56.0±0.7	56.0±0.6	61.2±1.5	48.3±1.4	56.1±0.6	53.2±1.1	53.6±2.0	61.8±2.3	61.5±2.1
1.6	60.4±1.0	60.5±1.0	60.5±1.0	68.4±2.4	46.8±2.1	60.6±0.8	56.7±1.1	56.0±3.5	69.1±3.6	68.9±3.5
1.8	64.5±1.5	64.6±1.6	64.6±1.7	74.6±3.5	46.4±3.2	64.8±1.1	60.2±1.3	58.4±5.8	75.7±4.2	75.3±4.4
2.0	67.5±2.5	67.5±2.5	67.6±2.7	80.2±4.1	47.4±4.5	67.8±1.8	63.1±1.9	60.4±8.4	81.2±4.4	80.8±4.8
2.5	72.9±4.2	72.9±4.2	73.0±4.7	89.9±3.7	46.6±5.5	73.3±3.2	69.1±3.2	64.2±9.8	90.5±4.6	90.1±4.8
3.0	77.5±4.0	77.5±4.0	77.4±4.5	94.7±2.4	43.9±5.2	77.8±3	74.3±3.7	66.2±10.2	94.9±3.5	94.6±3.7
3.5	79.0±4.3	79.0±4.3	79.0±4.8	96.3±2.1	44.1±6.2	79.3±3.5	76.8±4.8	66.3±11.7	96.4±3.2	96.2±3.2
4.0	80.4±4.4	80.4±4.4	80.4±4.9	97.2±1.6	42.4±6.1	80.6±3.9	79.5±5.5	66.3±13.1	97.2±2.8	97.1±2.6
4.5	81.6±4.0	81.7±4.0	81.6±4.5	98.0±1.0	40.4±6.3	81.9±3.5	81.4±5.2	66.1±13.8	97.8±2.3	97.9±1.9
5.0	82.6±3.6	82.6±3.6	82.6±4.1	98.4±0.7	38.7±6.2	82.9±3.0	83.0±4.8	66.1±14.2	98.3±1.6	98.4±1.3
5.5	83.3±3.7	83.3±3.7	83.2±4.2	98.7±0.5	37.2±5.6	83.6±2.9	84.4±4.4	66.1±14.7	98.7±1.2	98.8±0.9
6.0	83.8±3.7	83.8±3.7	83.7±4.3	98.9±0.5	36.6±6.0	84.1±2.9	85.7±4.2	66.0±15.0	98.9±0.9	99.0±0.7
6.5	84.0±3.8	84.0±3.8	83.9±4.3	99.0±0.5	36.9±6.5	84.3±2.9	86.6±4.3	65.9±15.4	99.1±0.7	99.2±0.5
Average	69.8	69.8	69.7	81.3	44.6	70.0	68.6	60.8	**81.5**	**81.5**

Table 1. NAS detection performance in binary classification task for NAS shift of brightness in RSNA boneage dataset measured by AUROC. Highlighted row presents the performance on ID data. MB and TAP-MB refers to Mahalanobis and *TAP-Mahalanobis*, respectively. Our task-agnostic approach significantly outperforms all baselines and is comparable to MB. Note that **MB is task-specific** and cannot be used in tasks other than classification.

ber of clusters in $[1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 16, 32]$. We observe that our proposed approach TAPUDD detects all OOD samples effectively. We also notice that TAP-MOS fails to detect OOD samples near or far from the periphery of all the clusters. Thus, to understand the scenarios where TAPMOS fails to identify OOD, we conduct a detailed analysis of the MOS approach in Sec. 2 of the Appendix and observe that it also fails for the similar corner cases.

TAPUDD outperforms TAP-Mahalanobis. We present a comparison to demonstrate the effectiveness of TAPUDD against *TAP-Mahalanobis* in Fig. 3. We present the ID score landscape of *TAP-Mahalanobis* for different values of K and TAPUDD with different ensemble variations for multi-class 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 in 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. We also provide results on a 2-D synthetic dataset for binary classification in Sec. 6.1 of the Appendix.

4.2. NAS Detection in Binary Classification

Experimental Details. We use the RSNA Bone Age dataset [17], composed of left-hand X-ray images of the patient and their gender and age (0 to 20 years). We alter the brightness of the X-ray images by a factor between 0 and 6.5 and form 20 different NAS datasets to reflect the X-ray imaging set-ups in different hospitals following [39]. In-distribution data consists of images with a brightness factor 1.0. We trained a ResNet18 model using the cross-entropy loss and assessed it on the ID test set composed of images with a brightness factor of 1.0. Further, we evaluate the NAS detection performance of our method and compare it with representative task-specific OOD detection methods on NAS datasets. Extensive details on the experimental set-up are described in Sec. 5 of the Appendix. For NAS detection, we measure the area under the receiver operating characteristic curve (AUROC), a commonly used metric for OOD detection. Additionally, we report the area under the precision-recall curve (AUPR) and the false positive rate of OOD examples when the true positive rate of in-distribution examples is at 95% (FPR95) in Sec. 7 of the Appendix.

Results. The in-distribution classification accuracy averaged across 10 seeds of the gender classifier trained using cross-entropy loss is 92.22. We compare the NAS detection performance of our proposed approach with competitive post-hoc OOD detection methods in literature in Table 1. As expected, the NAS detection performance of our approach and all baselines except KL Matching increase as the shift in the brightness attribute increases. We also observe that our proposed approaches, *TAPUDD* and *TAP-Mahalanobis* are more sensitive to NAS samples compared to competitive baselines, including Maximum Softmax Probability [20], ODIN [29], Mahalanobis distance [27], energy score [30],

Brightness	Baselines			Ours (Task-Agnostic)		
	DE [26]	MC Dropout [13]	SWAG* [33]	TAP-MOS (K = 8)	TAP-MB (K = 8)	TAPUDD (Average)
0.0	100.0±NA	6.9±NA	99.9±NA	57.8±31.5	100.0±0.1	100.0±0.0
0.2	57.0±NA	45.5±NA	51.4±NA	68.7±18.0	87.9±6.1	88.8±6.7
0.4	51.3±NA	50.8±NA	49.8±NA	70.7±16.4	64.5±6.9	66.6±5.0
0.6	50.7±NA	49.7±NA	49.5±NA	65.3±11.5	54.6±4.4	55.1±2.5
0.8	50.5±NA	49.9±NA	49.7±NA	57.7±6.0	48.9±1.7	49.2±1.0
1.0	50.0±NA	49.8±NA	50.0±NA	50.0±0.0	50.0±0.0	50.0±0.0
1.2	50.3±NA	48.5±NA	50.8±NA	48.7±4.0	57.6±1.8	57.8±1.9
1.4	54.5±NA	46.7±NA	55.8±NA	50.8±8.0	68.4±3.4	68.4±3.4
1.6	58.6±NA	44.5±NA	63.5±NA	55.4±11.3	78.7±3.6	78.6±3.7
1.8	64.9±NA	41.6±NA	71.6±NA	62.1±14.5	86.4±3.5	86.3±3.6
2.0	75.8±NA	38.4±NA	79.3±NA	67.3±16.8	91.9±3.0	91.7±3.2
2.5	95.6±NA	31.1±NA	89.8±NA	76.2±16.4	97.5±1.5	97.4±1.4
3.0	98.4±NA	25.8±NA	90.7±NA	82.8±13.5	99.0±0.6	99.0±0.5
3.5	99.3±NA	21.7±NA	93.7±NA	88.1±10.2	99.4±0.3	99.4±0.3
4.0	99.8±NA	18.0±NA	96.4±NA	90.7±6.8	99.6±0.3	99.6±0.2
4.5	100.0±NA	14.9±NA	97.4±NA	91.7±4.4	99.7±0.2	99.7±0.1
5.0	100.0±NA	11.7±NA	98.1±NA	91.7±3.8	99.8±0.1	99.7±0.1
5.5	100.0±NA	9.7±NA	98.5±NA	91.0±4.5	99.8±0.1	99.8±0.2
6.0	100.0±NA	7.9±NA	98.7±NA	89.7±5.7	99.8±0.1	99.8±0.2
6.5	100.0±NA	7.0±NA	98.9±NA	88.3±7.0	99.8±0.2	99.8±0.3
Average	77.8	31.0	76.7	72.2	**84.2**	**84.3**

Table 2. NAS detection performance in regression task (age prediction) for NAS shift of brightness in RSNA boneage dataset measured by AUROC. Highlighted row presents the performance on the ID dataset. DE, MC Dropout, TAP-MB, and NA denotes Deep Ensemble, Monte Carlo Dropout, *TAP-Mahalanobis*, and Not Applicable respectively. SWAG* = SWAG + Deep Ensemble.

Gram matrices [5], MOS [24], and KL matching [19]. All these task-specific baselines require the label information of the training dataset for OOD detection and cannot be used directly in tasks other than classification. On the other hand, our proposed task-agnostic approach does not require the access to class label information and it can be used across different tasks like regression.

4.3. NAS Detection in Regression

Experimental Details. We use the RSNA Bone Age dataset (described in Sec. 4.2) and solve the age prediction task. In this task, the objective is to automatically predict the patient's age given a hand X-ray image as an input. As described in Sec. 4.2, we vary the brightness of images by a factor between 0 and 6.5 and form 20 different NAS datasets. In-distribution data comprises images of brightness factor 1.0 (unmodified images). We train a ResNet18 with MSE loss and evaluate it on the test set composed of images with a brightness factor 1.0. Further, we evaluate the NAS detection performance of our proposed method and compare its performance with representative bayesian and non-bayesian uncertainty estimation methods on NAS datasets with attribute shift of brightness. Additionally, we compare the NAS detection performance of our approach with a well-known bayesian approach for uncertainty estimation, SWAG [33]. Extensive details on the experimental set-up are described in Sec. 5 of the Appendix. For NAS detection, we measure AUROC and additionally report AUPR and FPR95 in Sec. 7 of the Appendix.

Results. The in-distribution Mean Absolute Error (MAE) in year averaged across 10 seeds of the Resnet18 model trained using MSE loss is 0.801. We compare the NAS detection performance of our proposed approach with well-known uncertainty estimation methods, namely Deep Ensemble (DE) [26], Monte Carlo Dropout (MC Dropout) [13], and SWAG [33]. Although Deep Ensemble, MC Dropout, and SWAG are not applicable to a pre-trained model, we compare against these baselines as a benchmark, as it has shown strong OOD detection performance across regression examples. For DE, we retrain 10 models of the same architecture (Resnet18) using MSE loss from different initializations. Since SWAG is not directly applicable for OOD detection, we apply SWAG* which is a combination of deep ensembling on top of SWAG. From Table 2, as expected, we observe that the NAS detection performance of our approach and all baselines increase as the shift in the brightness attribute increases. We also observe that our proposed approaches, TAPUDD and *TAP-Mahalanobis*, are more sensitive to NAS samples and effectively detect them compared to the baselines and TAP-MOS. Additionally, it can be seen that TAP-MOS fails to detect extremely dark samples (Brightness Intensity 0.1). This might be because these NAS samples could locate near or away from the periphery of all clusters where MOS does not work well (as discussed in Sec. 4.1). This demonstrates that our proposed approach is better than the existing approaches.

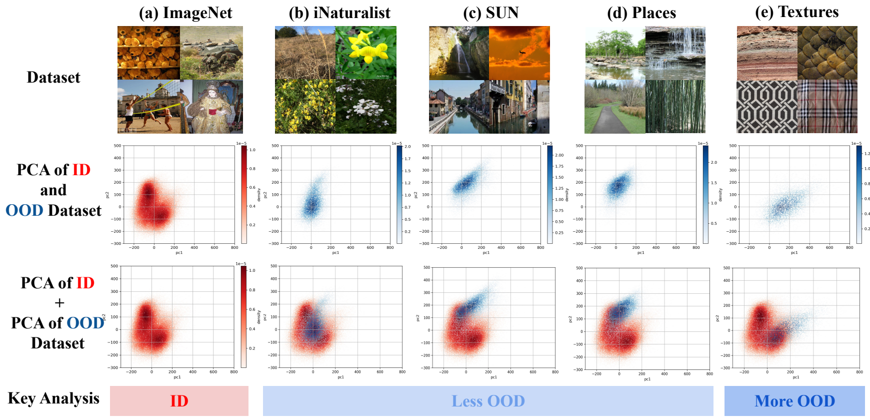

Figure 4. ***(first row)*** Examples of ID images sampled from Imagenet and OOD images sampled from iNaturalist, SUN, Places, and Textures datasets; ***(second row)*** Point-density based PCA visualization to demonstrate the location and density of ID and OOD datasets; ***(third row)*** Point-density based PCA visualization of ID dataset overlapped by PCA of OOD datasets to illustrate the location and density of OOD datasets relative to the ID dataset. ***(fourth row)*** From ***first*** and ***third row***, the key analysis is that **Textures is more OOD from Imagenet than the other three OOD datasets**.

Method	iNaturalist		SUN		Places		Textures		Average	
	AUROC ↑	FPR95 ↓	AUROC ↑	FPR95 ↓	AUROC ↑	FPR95 ↓	AUROC ↑	FPR95 ↓	AUROC ↑	FPR95 ↓
Expected Results	Low	High	Low	High	Low	High	**High**	**Low**	–	–
MSP [20]	87.70	63.52	78.22	80.01	76.67	81.31	74.46	82.70	79.26	76.88
ODIN [29]	89.49	62.61	83.83	71.89	80.60	76.51	76.29	81.31	82.55	73.08
Mahalanobis [27]	59.60	95.62	67.96	91.58	66.48	92.05	74.96	51.54	67.25	82.70
Energy [30]	88.64	64.35	85.25	65.30	81.31	72.77	75.78	80.60	82.75	70.76
KL Matching [19]	93.06	27.24	78.74	67.56	76.53	72.67	87.07	49.47	83.85	54.23
MOS [24]	98.15	9.23	92.01	40.38	89.05	49.49	81.27	60.30	90.12	39.85
TAPUDD (Average)	**70.00**	**84.46**	**70.47**	**79.52**	**66.97**	**84.72**	**97.59**	**10.85**	**76.26**	**64.88**

Table 3. OOD detection performance in the large-scale classification task. ↑ indicates larger values are better, while ↓ indicates smaller values are better. Ideally, all methods should follow the expected results obtained from our analysis (described in first row in green color). 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***.

4.4. OOD Detection for Large-scale Classification

Experimental Details. We use ImageNet-1k[10] as the in-distribution dataset and evaluate our approach on diverse OOD datasets presented in MOS[24], including iNaturalist[23], Places[52], SUN[47], and Textures[8]. For all baselines, we follow the experimental setup used in MOS[24]. We obtained the base model for our approach by following the experimental setting of finetuning the fully connected layer of pretrained ResNetv2-101 model by flattened softmax used in MOS. We measure AUROC and FPR95 and report AUPR in Sec. 7 of the Appendix.

Analysis. In binary classification (Sec. 4.2) and regression (Sec. 4.3) tasks, we evaluated our approach on NAS samples and knew that the NAS detection performance should increase with the increase in attribute shift. However, in this experiment, since the shift does not increase in a continuous manner, it is non-trivial to determine for which OOD datasets scores should be high. To this end, we perform an analysis wherein from the ***first row*** of Fig. 4, we observe that the Textures dataset is visually much more different to human eyes from Imagenet compared to the other three OOD datasets. Further, in the ***second*** and ***third rows*** of Fig. 4, we apply principal component analysis (PCA) on the feature representations obtained from the penultimate layer of the model to visualize the location of in-distribution samples (Imagenet) and diverse OOD datasets. From this, we observe that most of the samples from the Textures dataset are located away from the highly-dense regions of Imagenet compared to the other three datasets, thus indicating that Textures is more OOD. Hence, ***we expect a reliable OOD detection model to detect Textures as more OOD than the other three OOD datasets***.

Results. To determine the OOD detection methods that follow the expected analysis, we evaluate the performance of our approach and well-known post-hoc OOD detection methods, including Maximum Softmax Probability [20], ODIN [29], Mahalanobis distance [27], energy score [30], MOS [24], and KL matching [19]. It is also worth noting that all the baselines require class label information for OOD detection and cannot be used for tasks other than classification. From Table 3, we observe that only our approach (TAPUDD) and Mahalanobis follow the expected results, as highlighted in green color. This demonstrates that all baselines except Mahalanobis are less reliable. Further, we compare the performance of methods following the expected re-

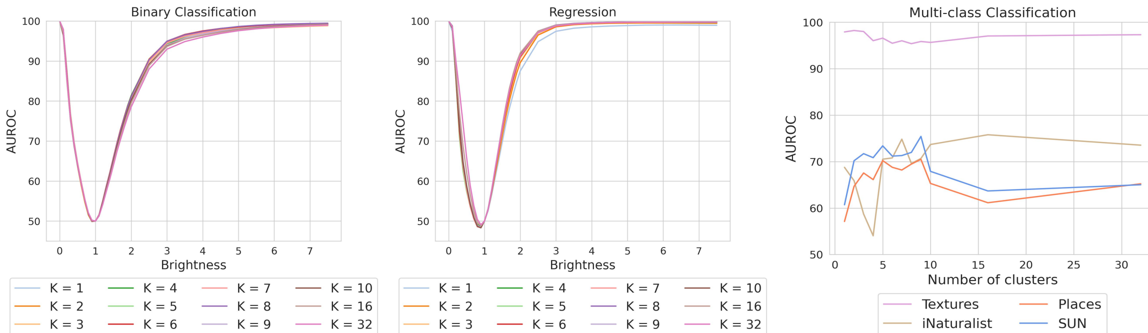

Figure 5. Effect on OOD detection performance by using different number of clusters in *TAP-Mahalanobis*.

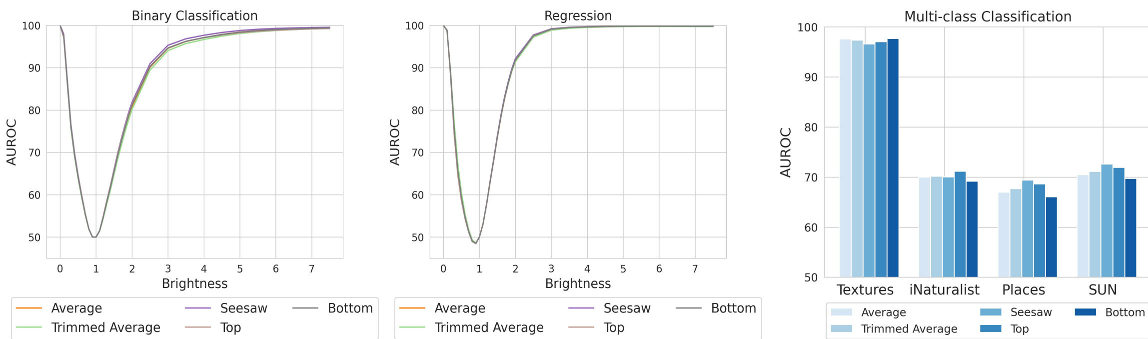

Figure 6. Comparison of OOD detection performance of TAPUDD under different ensembling strategies.

sults (rows highlighted in green color) and observe that ***our approach (TAPUDD) is more sensitive to OOD samples and outperforms the baselines in all OOD datasets***.

4.5. Ablations

Effect of Number of Clusters in TAP-Mahalanobis. We investigate the effect of the number of clusters on the OOD detection performance of *TAP-Mahalanobis* in Fig. 5. We observe that the performance of *TAP-Mahalanobis* varies with the value of K (*i.e.*, number of clusters) across different datasets and tasks. This implies that we cannot use a particular value of K for all tasks and datasets.

TAPUDD with Different Ensembling Strategies. We contrast the OOD detection performance of TAPUDD under different ensembling strategies for the three tasks in Fig. 6. We observe that TAPUDD shows competitive performance with diverse ensembling strategies for all the tasks and dataset shifts. Also, we observe that "seesaw" is slightly more sensitive towards NAS samples in binary classification and regression, and towards OOD samples from Places and SUN in large-class classification.

Further, Fig. 5 and Fig. 6 illustrate that using ensembling strategies provides us with an OOD detector which is almost as good as the best performer of *TAP-Mahalanobis*.

We also evaluate our approach on other conventional OOD datasets (ID: CIFAR-10; OOD: CIFAR-100, SVHN, etc) and for anomaly detection task in Sec. 6.2 and 6.3 of the Appendix, respectively. Additionally, we provide a discussion on tuning of hyperparameters in Sec. 4 of Appendix.

5. Conclusion

In this work, we propose a task-agnostic and post-hoc approach, TAPUDD, to detect samples from the unseen distribution. TAPUDD is a clustering-based ensembling approach composed of *TAP-Mahalanobis* and *Ensembling* modules. *TAP-Mahalanobis* module groups the semantically similar training samples into clusters and determines the minimum Mahalanobis distance of the test sample from the clusters. To enhance reliability and to eliminate the necessity to determine the optimal number of clusters for *TAP-Mahalanobis*, the *Ensembling* module aggregates the distances obtained from the *TAP-Mahalanobis* module for different values of clusters. We validate the effectiveness of our approach by conducting extensive experiments on diverse datasets and tasks. As future work, it would be interesting to extensively evaluate TAPUDD to detect samples from unseen distribution in natural language processing, 3D vision, and healthcare.

6. Acknowledgments

We would like to thank Isha Dua, Divyam Madaan and the anonymous reviewers for helpful feedback. This work was supported by Institute of Information & Communications Technology Planning & Evaluation (IITP) grants (No.2019-0-00075 Artificial Intelligence Graduate School Program(KAIST) and No.2022-0-009840101003), and National Research Foundation of Korea (NRF) grant (NRF-2020H1D3A2A03100945) funded by the Korea government (MSIT).

References

[1] Liron Bergman and Yedid Hoshen. Classification-based anomaly detection for general data. *ArXiv*, abs/2005.02359, 2020.

[2] Charles Blundell, Julien Cornebise, Koray Kavukcuoglu, and Daan Wierstra. Weight uncertainty in neural networks. *ArXiv*, abs/1505.05424, 2015.

[3] Mariusz Bojarski, Davide Del Testa, Daniel Dworakowski, Bernhard Firner, Beat Flepp, Prasoon Goyal, Lawrence D. Jackel, Mathew Monfort, Urs Muller, Jiakai Zhang, Xin Zhang, Jake Zhao, and Karol Zieba. End to end learning for self-driving cars, 2016.

[4] Raghavendra Chalapathy, Aditya Krishna Menon, and Sanjay Chawla. Anomaly detection using one-class neural networks. *ArXiv*, abs/1802.06360, 2018.

[5] 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.

[6] Hyunsun Choi, Eric Jang, and Alexander A Alemi. Waic, but why? generative ensembles for robust anomaly detection. *arXiv preprint arXiv:1810.01392*, 2018.

[7] Sung-Ik Choi and Sae-Young Chung. Novelty detection via blurring. *ArXiv*, abs/1911.11943, 2020.

[8] Mircea Cimpoi, Subhransu Maji, Iasonas Kokkinos, Sammy Mohamed, and Andrea Vedaldi. Describing textures in the wild. *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3606–3613, 2014.

[9] Lucas Deecke, Robert A. Vandermeulen, Lukas Ruff, Stephan Mandt, and M. Kloft. Image anomaly detection with generative adversarial networks. In *Proceedings of the European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECML-PKDD)*, 2018.

[10] 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.

[11] Yilun Du and Igor Mordatch. Implicit generation and modeling with energy based models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

[12] Andre Esteva, Brett Kuprel, Roberto Novoa, Justin Ko, Swetter Susan, Helen Balu, M, and Sebastian Thrun. Dermatologist-level classification of skin cancer with deep neural networks. 2016.

[13] 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.

[14] Izhak Golan and Ran El-Yaniv. Deep anomaly detection using geometric transformations. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2018.

[15] Will Grathwohl, Kuan-Chieh Wang, Jörn-Henrik Jacobsen, David Kristjanson Duvenaud, Mohammad Norouzi, and Kevin Swersky. Your classifier is secretly an energy based model and you should treat it like one. *ArXiv*, abs/1912.03263, 2020.

[16] Alex Graves. Practical variational inference for neural networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2011.

[17] 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.

[18] Leonard Hasenclever, Stefan Webb, Thibaut Lienart, Sebastian J. Vollmer, Balaji Lakshminarayanan, Charles Blundell, and Yee Whye Teh. Distributed bayesian learning with stochastic natural gradient expectation propagation and the posterior server. *ArXiv*, abs/1512.09327, 2017.

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

[20] Dan Hendrycks and Kevin Gimpel. A baseline for detecting misclassified and out-of-distribution examples in neural networks. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2017.

[21] Dan Hendrycks, Mantas Mazeika, Saurav Kadavath, and Dawn Xiaodong Song. Using self-supervised learning can improve model robustness and uncertainty. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

[22] José Miguel Hernández-Lobato and Ryan P. Adams. Probabilistic backpropagation for scalable learning of bayesian neural networks. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2015.

[23] Grant Van Horn, Oisin Mac Aodha, Yang Song, Yin Cui, Chen Sun, Alexander Shepard, Hartwig Adam, Pietro Perona, and Serge J. Belongie. The inaturalist species classification and detection dataset. *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8769–8778, 2018.

[24] 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.

[25] Anoop Korattikara, Vivek Rathod, Kevin Murphy, and Max Welling. Bayesian dark knowledge, 2015.

[26] 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.

[27] 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.

[28] Yingzhen Li, José Miguel Hernández-Lobato, and Richard E. Turner. Stochastic expectation propagation.

In *Advances in Neural Information Processing Systems (NeurIPS)*, 2015.

[29] 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.

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

[31] Christos Louizos and Max Welling. Structured and efficient variational deep learning with matrix gaussian posteriors. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2016.

[32] David J. C. Mackay. Bayesian methods for adaptive models. 1992.

[33] 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.

[34] Ahsan Mahmood, Junier Oliva, and Martin Andreas Styner. Multiscale score matching for out-of-distribution detection. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2020.

[35] Tomás Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2013.

[36] Eric T. Nalisnick, Akihiro Matsukawa, Yee Whye Teh, Dilan Görür, and Balaji Lakshminarayanan. Do deep generative models know what they don't know? *ArXiv*, abs/1810.09136, 2019.

[37] Eric T. Nalisnick, Akihiro Matsukawa, Yee Whye Teh, and Balaji Lakshminarayanan. Detecting out-of-distribution inputs to deep generative models using a test for typicality. *ArXiv*, abs/1906.02994, 2019.

[38] Radford M. Neal. Bayesian learning for neural networks. 1995.

[39] 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.

[40] Pramuditha Perera, Ramesh Nallapati, and Bing Xiang. Ocgan: One-class novelty detection using gans with constrained latent representations. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (cvpr)*, pages 2893–2901, 2019.

[41] Stanislav Pidhorskyi, Ranya Almohsen, Donald A. Adjeroh, and Gianfranco Doretto. Generative probabilistic novelty detection with adversarial autoencoders. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2018.

[42] Jie Ren, Peter J Liu, Emily Fertig, Jasper Snoek, Ryan Poplin, Mark A DePristo, Joshua V Dillon, and Balaji Lakshminarayanan. Likelihood ratios for out-of-distribution detection. *arXiv preprint arXiv:1906.02845*, 2019.

[43] Thomas Schlegl, Philipp Seeböck, Sebastian M. Waldstein, Ursula Margarethe Schmidt-Erfurth, and Georg Langs. Unsupervised anomaly detection with generative adversarial networks to guide marker discovery. In *International Conference on Information Processing in Medical Imaging*, 2017.

[44] 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.

[45] Joan Serrà, David Álvarez, V. Gómez, Olga Slizovskaia, José F. Núñez, and Jordi Luque. Input complexity and out-of-distribution detection with likelihood-based generative models. *ArXiv*, abs/1909.11480, 2020.

[46] Jost Tobias Springenberg, Aaron Klein, Stefan Falkner, and Frank Hutter. Bayesian optimization with robust bayesian neural networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2016.

[47] Yi Sun, Yuheng Chen, Xiaogang Wang, and Xiaoou Tang. Deep learning face representation by joint identification-verification. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 1988–1996, 2014.

[48] Jihoon Tack, Sangwoo Mo, Jongheon Jeong, and Jinwoo Shin. Csi: Novelty detection via contrastive learning on distributionally shifted instances. *ArXiv*, abs/2007.08176, 2020.

[49] Siqi Wang, Yijie Zeng, Xinwang Liu, En Zhu, Jianping Yin, Chuanfu Xu, and M. Kloft. Effective end-to-end unsupervised outlier detection via inlier priority of discriminative network. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

[50] Max Welling and Yee Whye Teh. Bayesian learning via stochastic gradient langevin dynamics. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2011.

[51] Zhisheng Xiao, Qing Yan, and Yali Amit. Do we really need to learn representations from in-domain data for outlier detection? *ArXiv*, abs/2105.09270, 2021.

[52] Bolei Zhou, Àgata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. Places: A 10 million image database for scene recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40:1452–1464, 2018.

[53] Bo Zong, Qi Song, Martin Renqiang Min, Wei Cheng, Cristian Lumezanu, Dae ki Cho, and Haifeng Chen. Deep autoencoding gaussian mixture model for unsupervised anomaly detection. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2018.