Journal of Machine Learning Research 24 (2023) 1-35 Submitted 6/23; Revised 10/23; Published 11/23

Principled Out-of-Distribution Detection via Multiple Testing

Akshayaa Magesh AMAGESH2@ILLINOIS.EDU
Department of Electrical and Computer Engineering
University of Illinois Urbana-Champaign
Champaign, IL 61820, USA

Venugopal V. Veeravalli VVV@ILLINOIS.EDU
Department of Electrical and Computer Engineering
University of Illinois Urbana-Champaign
Champaign, IL 61820, USA

Anirban Roy ANIRBAN.ROY@SRI.COM
Computer Science Laboratory
SRI International
Menlo Park, CA 94061

Susmit Jha SUSMIT.JHA@SRI.COM
Computer Science Laboratory
SRI International
Menlo Park, CA 94061

Editor: Daniel Roy

Abstract

We study the problem of out-of-distribution (OOD) detection, that is, detecting whether a machine learning (ML) model's output can be trusted at inference time. While a number of tests for OOD detection have been proposed in prior work, a formal framework for studying this problem is lacking. We propose a definition for the notion of OOD that includes both the input distribution and the ML model, which provides insights for the construction of powerful tests for OOD detection. We also propose a multiple hypothesis testing inspired procedure to systematically combine any number of different statistics from the ML model using conformal p-values. We further provide strong guarantees on the probability of incorrectly classifying an in-distribution sample as OOD. In our experiments, we find that threshold-based tests proposed in prior work perform well in specific settings, but not uniformly well across different OOD instances. In contrast, our proposed method that combines multiple statistics performs uniformly well across different datasets and neural networks architectures.

Keywords: OOD characterization, Conformal p-values, Conditional False Alarm Guarantees, Benjamini-Hochberg procedure.

1. Introduction

Given the ubiquitous use of ML models in safety-critical applications such as self-driving and medicine, there is a need to develop methods to detect whether an ML model's output at inference time can be trusted. This problem is commonly referred to as the out-of-distribution (OOD) detection problem. If an output is deemed untrustworthy by an OOD detector, one

can abstain from making decisions based on the output, and default to a safe action. There has been a flurry of works on this problem in recent years. A particular area of focus has been on OOD detection for deep learning models (see Lee et al., 2018; Liang et al., 2018; Sastry and Oore, 2020; Huang et al., 2021; Liu et al., 2020; Kaur et al., 2022). While neural networks generalize quite well to inputs from the same distribution as the training distribution, recent works have shown that they tend to make incorrect predictions with high confidence, even for unrecognizable or irrelevant inputs (see e.g., Szegedy et al., 2013; Nguyen et al., 2015; Hendrycks and Gimpel, 2017).

In many of the prior works on OOD detection, OOD inputs are considered to be inputs that are not generated from the input training distribution (see, e.g., Liang et al., 2018; Sastry and Oore, 2020), which better describes the classical problem of outlier detection. However, in contrast to outlier detection, the goal in OOD detection is to flag untrustworthy outputs from a *given* ML model. Thus, it is essential for the definition of an OOD sample to involve the ML model. One of the contributions of this paper is a formal definition for the notion of OOD that involves both the input distribution and the ML model.

In a line of work in OOD detection, it is assumed that the detector has access (exposure) to OOD examples, which can be used to train an auxiliary classifier, or to tune hyperparameters for the detection model (see Lee et al., 2018; Hendrycks et al., 2019; Liang et al., 2018, 2022). Other works rely on identifying certain patterns observed in the training data distribution, and use these patterns to train the original ML model to help detect OOD examples. For instance, in the work by Kaur et al. (2022), a neural network is trained to leverage in-distribution equivariance properties for OOD detection. There is another line of work in which tests are designed based on statistics from generative models trained for OOD detection. For instance, in the work by Bergamin et al. (2022), statistics from a deep generative model are combined through p-values using the Fisher test. In this paper, we focus exclusively on developing methods that do not use any OOD samples, and can be applied to *any* pre-trained ML model.

Prior work has primarily been focused on identifying promising test statistics and corresponding thresholds, sometimes motivated by empirical observations of the values taken by these statistics for certain in-distribution and OOD inputs. For instance, in the work by Lee et al. (2018), a confidence score is constructed through a weighted sum of Mahalanobis distances across layers, using the class conditional Gaussian distributions of the features of the neural network under Gaussian discriminant analysis. Liang et al. (2018) proposed a statistic based on input perturbations and temperature-scaled softmax scores. Liu et al. (2020) proposed a *free energy* score based on the denominator of the temperature-scaled softmax score. Sastry and Oore (2020) derived scores from Gram matrices, through the sum of deviations of the Gram matrix values from their respective range observed over the training data. In the work by Angelopoulos et al. (2021), the broad goal is to find all candidate functions from a given collection in an offline manner through multiple testing, such that any one of these candidate functions controls some risk at inference time. This approach is applied by Angelopoulos et al. (2021) to the problem of OOD detection to select suitable thresholds for a given test statistic to control the false alarm rate. Huang et al. (2021) used vector norms of gradients from a pre-trained network to form a test statistic. Haroush et al. (2021) studied OOD detection in Convolutional Neural Networks (CNNs), where spatial and channel reduction techniques are employed to produce statistics per layer, and these layer

statistics are combined to form a final score using a method motivated by the tests proposed by Simes (1986) and Fisher (1992). Thus, their proposed algorithm computes a single score using all the intermediate features of the CNN and its corresponding empirical p-value. They provide marginal false alarm guarantees averaged over all possible validation datasets used to compute the empirical p-value. Additionally, the proposed method by Haroush et al. (2021) can be applied only to CNNs, and not any general ML model. To summarize, from prior work, it is unclear which among these scores/statistics is the best for OOD detection, or if there exists such a test statistic that is useful for all possible out-distributions. The latter question was raised by Zhang et al. (2021), where they posit that one can construct an out-distribution for any single score or statistic that results in poor detection performance.

The false alarm probability or type-I error of a test refers to the probability of a single in-distribution sample being misclassified as OOD, and the detection power refers to the probability of correctly identifying an OOD distribution sample. Note that the detection power is also referred to as detection accuracy in prior OOD works. In much of the prior work on OOD detection, the false alarm probability is estimated using empirical evaluations on certain in-distribution datasets. What is lacking in such works is a rigorous theoretical analysis of the probability of false alarm, which can be used to meet pre-specified false alarm constraints. Such false alarm guarantees are crucial for the responsible deployment of OOD methods in practice. Note that it is not possible to give any theoretical guarantees on the detection power of an OOD detection test without prior information about the class of all possible out-distributions, which is typically not available in practice. Therefore, in prior work on OOD detection, the detection powers of candidate OOD methods that meet the same pre-specified false alarm levels are compared empirically.

In this work, we propose a method inspired by *multiple hypothesis testing* (Holm, 1979; Benjamini and Hochberg, 1995; Benjamini and Yekutieli, 2001) to systematically combine multiple test statistics for OOD detection. Our method works for combining any number of statistics with an arbitrary dependence structure, for instance the Mahalanobis distances (Lee et al., 2018) and the Gram matrix deviations across layers (Sastry and Oore, 2020) of a neural network. We should emphasize there is no obvious way to directly combine such disparate statistics with provable guarantees for OOD detection. Detection procedures for multiple hypothesis testing are usually based on combining p-values across hypotheses (Holm, 1979; Benjamini and Hochberg, 1995; Benjamini and Yekutieli, 2001). However, in the problem of OOD detection, the probability measures under both the in-distribution (null) and out-of-distribution (alternate) settings are unknown, and thus the actual p-values cannot be computed. In conformal inference methods (Vovk et al., 1999; Balasubramanian et al., 2014) the p-values are replaced with *conformal p-values*, which are estimates computed from the empirical CDF of the test statistics. These conformal p-values are data-dependent, as they are calculated from in-distribution samples. In the procedure proposed in this paper, we use conformal p-values and provide rigorous theoretical guarantees on the probability of false alarm, conditioned on the dataset used for computing the conformal p-values.

Contributions

1. We formally characterize the notion of OOD, using which we provide insights on why it is necessary for OOD tests to involve more than just the new unseen input and the final output of the ML model for OOD detection.

2. We propose a new approach for OOD detection inspired by multiple testing. Our proposed test allows us to combine, in a systematic way, any number of different test statistics produced from the ML model with arbitrary dependence structures.

3. We provide strong theoretical guarantees on the probability of false alarm, conditioned on the dataset used for computing the conformal p-values. This is stronger than false alarm guarantees in prior work (e.g., Balasubramanian et al., 2014; Kaur et al., 2022), where the guarantees are given in terms of an expectation over all possible datasets.

4. We perform extensive experiments across different datasets to demonstrate the efficacy of our method. We perform ablation studies to show that combining various statistics using our method produces uniformly good results across various types of OOD examples and deep neural network (DNN) architectures.

2. Problem Statement and OOD Modeling

Consider a learning problem with $(X, Y) \sim \mathrm{P}_{X,Y}$, where (X, Y) is the input-output pair and $\mathrm{P}_{X,Y}$ is the distribution of the dataset available at training time. Let the dataset available at training time be denoted by $\mathcal{T} = \{(X_1, Y_1), (X_2, Y_2), ..., (X_n, Y_n)\}$, where n is the size of the dataset. Let the ML model be denoted by $f(\mathbf{W}, .)$, where $\mathbf{W}$ is the random variable denoting the parameters of the ML model. For instance, $\mathbf{W}$ depicts the weights and biases in a neural network. Let $(X_{\text{test}}, Y_{\text{test}})$ be a random variable generated from an unknown distribution, and $(x_{\text{test}}, y_{\text{test}})$ be an instance of this random variable seen by the ML model at inference time. Given $\mathcal{T}$ and the ML model, the goal is to detect if this new unseen sample might produce an *incorrect* output with high confidence. This might happen because either the input does not conform to the training data distribution, or if the ML model is unable to capture the true relationship between the input X_{test} and the true label Y_{test}. Whether a new unseen sample is OOD or not depends on both the ML model and the distribution $\mathrm{P}_{X,Y}$.

A precise mathematical definition of the OOD detection problem that captures both the input distribution and the ML model appears to be lacking in prior work. The most common definition is based on testing between the following hypotheses (see, e.g., Liang et al., 2018):

$$\begin{aligned} \mathrm{H}_0 &: X_{\text{test}} \sim \mathrm{P}_X \\ \mathrm{H}_1 &: X_{\text{test}} \not\sim \mathrm{P}_X, \end{aligned} \tag{1}$$

where H_0 corresponds to 'in-distribution' and H_1 corresponds to 'out-of-distribution'. However, such a definition does not involve the ML model, and better describes the problem of outlier detection, which is fundamentally different from the problem of OOD detection.

Let $\hat{Y} = f(\mathbf{W}, X)$, and consider the distribution $\mathrm{P}_{X,\hat{Y}} = \mathrm{P}_X \times \mathrm{P}_{\hat{Y}|X}$ as the joint distribution of the input and the output of the ML model. Using this joint distribution as

the *'in-distribution'*, consider the following testing problem:

$$\begin{aligned} \mathrm{H}_0 &: (X_{\text{test}}, Y_{\text{test}}) \sim \mathrm{P}_{X,\hat{Y}} \\ \mathrm{H}_1 &: (X_{\text{test}}, Y_{\text{test}}) \not\sim \mathrm{P}_{X,\hat{Y}}. \end{aligned} \tag{2}$$

Note that this is a definition of OOD detection that involves both the input distribution and the ML model (through $\mathrm{P}_{\hat{Y}|X} = \mathrm{P}_{f(\mathbf{W},X=x)|X=x}$). It also captures both the cases where the input is not drawn from P_X, and when the ML model is unable to capture the relationship between the unseen input and its label.

The hypothesis test in (2) involves the true label Y_{test} and the distribution $\mathrm{P}_{X,\hat{Y}}$. Since these quantities are unknown, the model prediction $\hat{Y}_{\text{test}}$ and the empirical distribution $\hat{\mathrm{P}}_{X,\hat{Y}}$ of $(X, \hat{Y})$ based on the training data, respectively, may be used instead. When the ML model performs well during training, the training loss is approximately 0, i.e,

$$\text{Training loss} = \frac{1}{n}\sum_{i=1}^{n} L(Y_i, \hat{Y}_i) \approx 0, \tag{3}$$

where $L(Y, \hat{Y})$ is a non-negative loss function with $L(Y, \hat{Y}) = 0$ implying $Y = \hat{Y}$. Thus, when the training loss is approximately 0, $Y = \hat{Y}$ holds for almost all the training data points. This implies that the empirical versions $\hat{\mathrm{P}}_{X,\hat{Y}}$ and $\hat{\mathrm{P}}_{X,Y}$ of the distributions $\mathrm{P}_{X,\hat{Y}}$ and $\mathrm{P}_{X,Y}$, respectively, are approximately equal to each other. Using the empirical versions of the distributions in (2), we arrive at a formulation that tests whether the new unseen sample conforms to the distribution $\hat{\mathrm{P}}_{X,Y}$ or not, which again does not involve the ML model. Thus, in order to incorporate the ML model $f(\mathbf{W}, .)$ in an operational OOD detection framework, we conclude that it is necessary to use other functions of the input derived from the ML model[1] in addition to just the final output $\hat{Y}$ in constructing test statistics for effective OOD detection. Such a strategy is commonly employed, without theoretical justification, in many OOD detection works, for instance, through the use of intermediate features of a neural network to calculate the Mahalanobis score (Lee et al., 2018) and Gram matrix score (Sastry and Oore, 2020), and gradient information to calculate the GradNorm score (Huang et al., 2021). The discussion above provides a qualitative theoretical justification for these strategies developed in prior works.

3. Proposed Framework and Algorithm

In this section, we describe our proposed framework formally, and present our algorithm to combine any number of different functions of the input with an arbitrary dependence structure.

In our formulation of OOD detection in (2), we posit that, in addition to the input and the output from the ML model, it is necessary to use other functions[2] of the input, which are dependent on the ML model. We refer to these functions as *score functions*, denoted by

1. In this paper, we use the term *statistic* or *score* interchangeably to denote these functions of the input derived from the ML model.
2. Without loss of generality, we may assume that these functions are scalar-valued.

$s^1(.), \ldots, s^K(.)$. The outputs of the score functions are scalar-valued *scores* $T^1, T^2, \ldots, T^K$:

$$\begin{aligned} T^1 &= s^1(X) \\ &\vdots \\ T^K &= s^K(X). \end{aligned} \tag{4}$$

The scores are chosen based on some prior information, say using empirical observations, that the distributions of the scores for in-distribution samples are concentrated around smaller values, whereas they are likely to be concentrated around larger values for OOD samples (Liang et al., 2018, Sec 5.2). For a new input X_{test}, let $(T^1_{\text{test}}, T^2_{\text{test}}, \ldots, T^K_{\text{test}})$ be the corresponding scores. Note that one of scores T^k_{test} could be based on the final output from the learning model $\hat{Y}_{\text{test}}$.

3.1 Motivation for multiple testing framework

In order to construct an OOD detection test for the new sample X_{test} using the scores, the scores would need to be combined in some manner. Since we do not know the dependence structure between the scores, combining them in an ad hoc manner, such as summing them up, cannot be justified and may result in tests with low power (probability of detection) for many OOD distributions.

For instance, consider a simple bivariate Gaussian setting as follows:

$$\begin{aligned} &\mathrm{H}_0 : (T^1, T^2) \sim \mathcal{N}(0, I) \\ &\mathrm{H}_1 : (T^1, T^2) \not\sim \mathcal{N}(0, I). \end{aligned} \tag{5}$$

Ad hoc combining: Consider the test that combines the test statistics in an ad hoc manner by summing them. Let the statistic $T = T^1 + T^2$, and let Q be the p-value when the observed value of the statistic is t. Recall that the p-value is given by:

$$Q = \mathrm{P}_{\mathrm{H}_0} \{T \geq t\}. \tag{6}$$

For given $\alpha > 0$, let $\mathcal{T}_1$ denote the test which rejects H_0 if $Q < \alpha$. For test $\mathcal{T}_1$, the probability of false alarm, i.e.,

$$\mathrm{P}_{\mathrm{H}_0}(\text{reject } \mathrm{H}_0), \tag{7}$$

can be controlled at α, by exploiting the fact that p-values have a uniform distribution under the null hypothesis. However, the detection power of the test under different possible distributions under the alternate hypothesis might be poor. For instance, if $(T^1, T^2) \sim \mathcal{N}((1, -1), I)$ under the alternate hypothesis, the statistic $T = T^1 + T^2$ has the same distribution under the null and alternate hypotheses. Thus the detection power of test $\mathcal{T}_1$ is α as well. It is possible to find many such joint distributions for the alternate hypothesis, under which the detection power of test $\mathcal{T}_1$ is poor, i.e., it is close to the probability of false alarm.

Combining Inspired by Multiple Testing: Consider the following split of the above testing problem into two binary hypothesis testing problem corresponding to the statistics T^1 and T^2:

$$\begin{aligned} &\mathrm{H}_{0,1} : T^1 \sim \mathcal{N}(0, 1) \qquad \mathrm{H}_{1,1} : T^1 \not\sim \mathcal{N}(0, 1) \\ &\mathrm{H}_{0,2} : T^2 \sim \mathcal{N}(0, 1) \qquad \mathrm{H}_{1,2} : T^2 \not\sim \mathcal{N}(0, 1). \end{aligned} \tag{8}$$

Let Q^1 and Q^2 be the p-values corresponding to the two individual tests in (8), and $Q^{(1)} \leq Q^{(2)}$ be the ordered p-values. Let

$$m = \max\{i : Q^{(i)} \leq i\alpha/2\}. \tag{9}$$

Then, let test $\mathcal{T}_2$ be defined such that it rejects H_0 if $m \geq 1$. Similar to test $\mathcal{T}_1$, the probability of false alarm of test $\mathcal{T}_2$ can be controlled at level α. On the other hand, we see that the detection power of test $\mathcal{T}_2$ when $(T^1, T^2) \sim \mathrm{P}_\mu$, where $\mathrm{P}_\mu = \mathcal{N}((\mu_1, \mu_2), I)$, satisfies the following condition:

$$\mathrm{P}_\mu(\text{reject } \mathrm{H}_0) \geq 1 - \min\{1 - \Psi(\Psi^{-1}(\alpha/2) - \mu_1), 1 - \Psi(\Psi^{-1}(\alpha/2) - \mu_2)\}, \tag{10}$$

where $\Psi(.)$ is the complementary cumulative distribution function of a $\mathcal{N}(0, 1)$ random variable. The detection power satisfies a minimum quality of performance under all distributions for the alternate hypothesis. Note that it is also possible for some distributions for the alternate hypothesis, for instance if $(T^1, T^2) \sim \mathcal{N}((1, 1), I)$, that test $\mathcal{T}_1$ has better detection power than test $\mathcal{T}_2$.

Thus, if we do not have any prior information on the behaviour of the statistics under the alternate hypotheses, combining multiple test statistics in an ad hoc manner (such as summing them) might not be desirable. Further, there is no obvious way to combine two completely different set of statistics, say the Mahalanobis scores from different layers of a DNN and the energy score.

3.2 Proposed OOD Detection Test

Motivated by the above discussion, we propose the following multiple testing framework for OOD detection:

$$\begin{aligned} &\mathrm{H}_{0,1} : T^1_{\text{test}} \sim \mathrm{P}^1 \qquad \mathrm{H}_{1,1} : T^1_{\text{test}} \nsim \mathrm{P}^1 \\ &\vdots \\ &\mathrm{H}_{0,K} : T^K_{\text{test}} \sim \mathrm{P}^K \qquad \mathrm{H}_{1,K} : T^K_{\text{test}} \nsim \mathrm{P}^K, \end{aligned} \tag{11}$$

where $\mathrm{P}^1, \ldots, \mathrm{P}^K$ are the distributions of the corresponding scores when X_{test} is an in-distribution sample as defined in (2). It is clear to see that if the new input X_{test} is an in-distribution sample, then all $\mathrm{H}_{0,i}$ are true in (11), and if X_{test} is an OOD sample, then one or more of $(\mathrm{H}_{0,1}, \ldots, \mathrm{H}_{0,K})$ are likely to be false. Thus, we propose a test that declares the instance as OOD, if any of the $\mathrm{H}_{0,i}$ are rejected.

We propose an algorithm for OOD detection inspired by the Benjamini-Hochberg (BH) procedure given by Benjamini and Yekutieli (2001) (preliminaries are provided in the Appendix). Most multiple testing techniques, including the BH procedure, involve computing the p-values of the individual tests. The p-value of a realization t^i of the test statistic T^i, $i \in [K]$, is given by

$$q^i = \mathrm{P}_{\mathrm{H}_{0,\mathrm{i}}}\left\{T^i \geq t^i\right\} = 1 - F_{\mathrm{H}_{0,i}}(t^i), \tag{12}$$

where $F_{\mathrm{H}_{0,i}}(.)$ is the CDF of T^i. The p-value for T^i_{test} is a random variable

$$Q^i = 1 - F_{\mathrm{H}_{0,i}}(T^i_{\text{test}}). \tag{13}$$

The distribution of this p-value under null hypothesis is uniform over $[0, 1]$. Its distribution under the alternate hypothesis concentrates around 0, and is difficult to characterize in general. Also, while a p-value close to 0 is evidence against the null hypothesis, a large p-value does not provide evidence in favor of the null hypothesis.

If we do not know the distributions under the null hypotheses to calculate the exact p-values, conformal inference methods suggest evaluating the empirical CDF of T^i under the null hypothesis using a hold-out set (denoted by $\mathcal{T}_{\text{cal}}$) known as the calibration set, to construct a *conformal p-value* $\hat{Q}^i$. A conformal p-value satisfies the following property:

$$\mathrm{P}_{\mathrm{H}_{0,\mathrm{i}}}\left\{\hat{Q}^i \leq t\right\} \leq t, \tag{14}$$

when X_{test} is independent from $\mathcal{T}_{\text{cal}}$ and T^i has a continuous distribution. The classical conformal p-value (see Vovk et al., 1999) is given by:

$$\hat{Q}^i = \frac{1 + |\{j \in \mathcal{T}_{\text{cal}} : T^i_j \geq T^i_{\text{test}}\}|}{1 + |\mathcal{T}_{\text{cal}}|}. \tag{15}$$

The estimate $\hat{Q}^i$ is said to be a marginally valid conformal p-value, as it depends on $\mathcal{T}_{\text{cal}}$. In other words, (14) can be rewritten as follows:

$$\mathrm{E}\left[\mathrm{P}_{\mathrm{H}_{0,\mathrm{i}}}\left\{\hat{Q}^i \leq t|\mathcal{T}_{\text{cal}}\right\}\right] \leq t, \tag{16}$$

where the expectation is over all possible calibration datasets. The property in (14) is however not valid conditionally, i.e., $\mathrm{P}_{\mathrm{H}_0}\left\{\hat{Q}^i \leq t|\mathcal{T}_{\text{cal}}\right\}$ need not be upper-bounded by t. This is important to note, as false alarm guarantees given for out-of-distribution detection methods using conformal inference (see, e.g., Balasubramanian et al., 2014; Kaur et al., 2022) are based on (16). Such guarantees are not strong, as they only guarantee that the probability of false alarm, averaged over all possible calibration data sets, is controlled. While the problem of conditional coverage has been discussed in the context of sequential testing for distribution shifts (e.g., Podkopaev and Ramdas, 2021) and conformal inference (e.g., Vovk, 2012), it has not been discussed widely under the setting of single sample OOD detection.

The related problem of outlier testing using conformal p-values is studied by Bates et al. (2023). However, the result from Bates et al. (2023), stating that conformal p-values satisfy the PRDS (Positive Regression Dependent on a Subset) property, which is required for the False Discovery Rate (FDR) control in the BH procedure, is valid only under the setting where the individual test statistics (and hence the original p-values) are independent. The PRDS property does not hold for the conformal p-values $\hat{Q}^i$ in Algorithm 1, since the corresponding p-values Q^i (see (13)) are highly dependent through the common input X_{test}. In addition, the conditional false alarm guarantees provided by Bates et al. (2023) utilize calibration conditionally valid (CCV) p-values proposed by Bates et al. (2023), as opposed to the conformal values proposed by Vovk et al. (1999) (which we use in our work). Indeed, these CCV p-values cannot be directly used in our setting to obtain false alarm guarantees in Theorem 2, without a similar adjustment to the thresholds as in (19), as the p-values would be dependent through both the calibration dataset and the input.

In our proposed OOD detection test we use conformal p-values in place of the actual p-values. In order to compute the conformal p-values, we maintain a calibration set $\mathcal{T}_{\text{cal}}$.

In this work, we aim to provide conditional false alarm guarantees, i.e., if X_{test} is an in-distribution sample (all $\mathrm{H}_{0,i}$ are true in (11)), then

$$\mathrm{P_F}(\mathcal{T}_{\text{cal}}) = \mathrm{P}_{\mathrm{H}_0}(\text{declare OOD } |\mathcal{T}_{\text{cal}}) = \mathrm{P}_{\mathrm{H}_0}(\text{reject at least one } \mathrm{H}_{0,i}|\mathcal{T}_{\text{cal}}) \tag{17}$$

is controlled with high probability. As discussed earlier in this section, such conditional guarantees are essential for the safe deployment of OOD detection algorithms. Note that in the literature on multiple testing, the marginal false alarm probability $\mathrm{P}_{\mathrm{H}_0}(\text{declare OOD })$ is equivalent to the Family Wise Error Rate (FWER) or False Discovery Rate (FDR) when all the null hypotheses are true in (11) (detailed discussion provided in the Appendix).

We compute the scores of these K statistics for the samples in the calibration set $\mathcal{T}_{\text{cal}}$. Using these, we calculate the conformal p-values $\hat{Q}^1, \hat{Q}^2, \ldots, \hat{Q}^K$ for the new sample as in (74), and order the conformal p-values in increasing order as $\hat{Q}^{(1)}, \hat{Q}^{(2)}, \ldots, \hat{Q}^{(K)}$. Let $\epsilon > 0$ be a parameter of the OOD detection algorithm, and let $\alpha > 0$, and let

$$m = \max \left\{ i : \hat{Q}^{(i)} \leq \frac{\alpha i}{C(K)K} \right\}, \tag{18}$$

where

$$C(K) = (1+\epsilon) \sum_{j=1}^{K} \frac{1}{j}. \tag{19}$$

The factor of $\sum_{j=1}^{K} \frac{1}{j}$ is included in order to obtain false alarm guarantees for any arbitrary dependence between the test statistics. The factor of $(1+\epsilon)$ is a constant related to the size of the calibration dataset, and is introduced to provide strong conditional false alarm guarantees, conditioned on the calibration set (discussed further in the proof of the results below). While choosing a smaller value of ϵ improves the power of the proposed OOD detection test, it increases the size of the calibration set needed to provide the conditional false alarm guarantees. The OOD detection test declares the instance X_{test} as OOD if $m \geq 1$, i.e., if any of the $\mathrm{H}_{0,i}$ are rejected. The pseudo-code is described in Algorithm 1.

For instance, consider a deep neural network (DNN) with L layers. Let $T^1, \ldots, T^L$ denote the Mahalanobis scores (Lee et al., 2018). Let $T^{L+1}, \ldots, T^{2L}$ denote the Gram deviation scores (Sastry and Oore, 2020). Lee et al. (2018) use outlier exposure to combine $T^1, \ldots, T^L$ into a single score for a threshold-based test, and Sastry and Oore (2020) use the sum of $T^{L+1}, \ldots, T^{2L}$ for a similar test. However, it is not straightforward to determine how to combine the L Mahalanobis scores and the L Gram deviation scores for OOD detection without outlier exposure. In Algorithm 1, we provide a systematic way to construct a test that uses all these $2L$ contrasting scores. In addition, we provide a systematic way to design the test thresholds to meet a given false alarm constraint as presented below in Theorem 2.

3.3 Theoretical Guarantees

On running Algorithm 1, we can guarantee that the conditional probability of the false alarm is bounded by α with high probability. In order to provide this guarantee, we need to enforce certain sample complexity conditions on the size of the calibration set n_{cal}, as detailed in the Lemma below.

Algorithm 1 BH based OOD detection test with conformal p-values

Inputs:
New input X_{test};
Scores over $\mathcal{T}_{cal}$ as $\left\{\{T_j^1 = s^1(X_j) : j \in \mathcal{T}_{\text{cal}}\}, \ldots, \{T_j^K = s^K(X_j) : j \in \mathcal{T}_{\text{cal}}\}\right\}$;
ML model $f(\mathbf{W}, .)$;
Desired conditional probability of false alarm $\alpha \in (0, 1)$.
Algorithm:
For X_{test}, compute scores T_{test}^i.
Calculate conformal p-values as:

$$\hat{Q}^i = \frac{1 + |\{j \in \mathcal{T}_{\text{cal}} : T_j^i \geq T_{\text{test}}^i\}|}{1 + |\mathcal{T}_{\text{cal}}|}. \tag{20}$$

Order them as $\hat{Q}^{(1)} \leq \hat{Q}^{(2)} \leq \ldots \leq \hat{Q}^{(K)}$.
Calculate $m = \max\left\{i : \hat{Q}^{(i)} \leq \frac{\alpha i}{C(K)K}\right\}$.
Output:
Declare OOD if $m \geq 1$.

Lemma 1 *Let $\epsilon > 0$, K and α be as in Algorithm 1. Let $a_j = \lfloor (n_{\text{cal}} + 1)\frac{\alpha j}{C(K)K} \rfloor$, $b_j = (n_{\text{cal}} + 1) - a_j$, and $\mu_j = \frac{a_j}{a_j + b_j}$. For a given $\delta > 0$, let n_{cal} be such that*

$$\min_{j=1,2,\ldots,K} I_{(1+\epsilon)\mu_j}(a_j, b_j) \geq 1 - \frac{\delta}{K^2}, \tag{21}$$

where $I_x(a, b)$ is the regularized incomplete beta function (the CDF of a Beta distribution with parameters a, b). Then for random variables $r_j^i \sim \text{Beta}(a_j, b_j)$ for $j = 1, \ldots, K$,

$$P\left\{\bigcap_{i=1}^{K}\bigcap_{j=1}^{K}\left\{r_j^i \leq (1+\epsilon)\frac{\alpha j}{C(K)K}\right\}\right\} \geq 1 - \delta. \tag{22}$$

Proof When the condition on n_{cal} in (21) is satisfied, we have that

$$\begin{aligned} P\left\{r_j^i \leq (1+\epsilon)\frac{\alpha j}{C(K)K}\right\} &= I_{(1+\epsilon)\frac{\alpha j}{C(K)K}}(a_j, b_j) \\ &\geq I_{(1+\epsilon)\mu_j}(a_j, b_j) \\ &\geq 1 - \frac{\delta}{K^2}, \end{aligned} \tag{23}$$

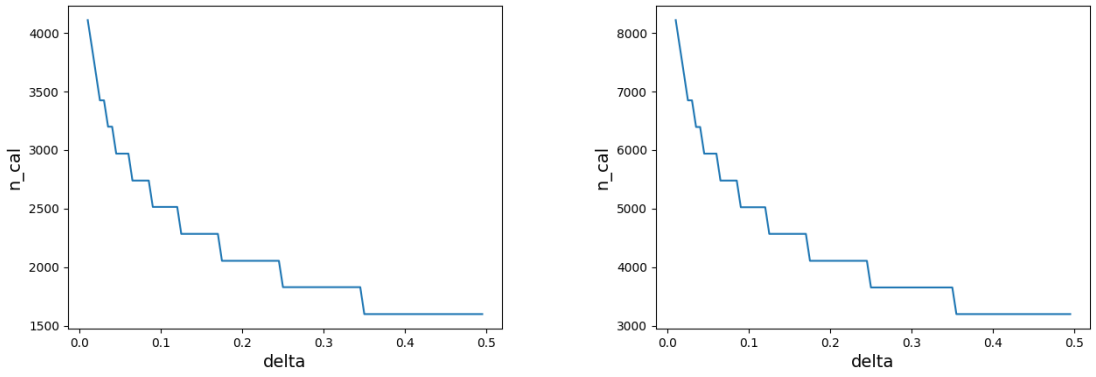
where $I_x(a, b)$ is the CDF of a Beta distribution with parameters a, b, and the second inequality follows since μ_j is upper bounded by $\frac{\alpha j}{C(K)K}$. From the Union Bound, we have

that,

$$
\begin{aligned}
1 - P\left\{\bigcap_{i=1}^{K}\bigcap_{j=1}^{K}\left\{r_j^i \leq (1+\epsilon)\frac{\alpha j}{C(K)K}\right\}\right\} &\leq \sum_{i=1}^{K}\sum_{j=1}^{K} P\left\{r_j^i \geq (1+\epsilon)\frac{\alpha j}{C(K)K}\right\} \\
&\leq \sum_{i=1}^{K}\sum_{j=1}^{K}\frac{\delta}{K^2} \\
&\leq \delta.
\end{aligned} \tag{24}
$$

Thus, we have the desired result in Lemma 1. ■

The condition on n_{cal} in Lemma 1 is due to the fact that the CDF of the conformal p-values conditioned on the calibration dataset follows a Beta distribution (see Vovk et al., 1999), and is essential to provide the guarantees in Theorem 2. Due to the form of the CDF of the Beta distribution, it is difficult to characterize the dependence of n_{cal} on α, δ, ϵ and K in closed form. We plot the calibration dataset sizes n_{cal} as given by Lemma 1 for $\epsilon = 1$ and $K = 5$ for different values of δ in Figure 1. Note that $\epsilon = 1$ is conservative.



(a) $\alpha = 0.1$ (b) $\alpha = 0.05$

Figure 1: Calibration dataset sizes that guarantees Theorem 1 with probability $1 - \delta$

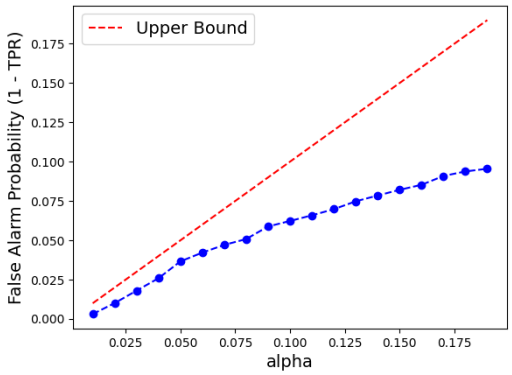
In the following result, we formally present the conditional false alarm guarantee for Algorithm 1.

Theorem 2 *Let $\alpha, \delta \in (0, 1)$. Let $\mathcal{T}_{\text{cal}}$ be a calibration set, and let n_{cal} be large enough (as defined in the Lemma 1). Then, for a new input X_{test} and an ML model $f(\mathbf{W}, .)$, the probability of incorrectly detecting X_{test} as OOD conditioned on $\mathcal{T}_{\text{cal}}$ while using Algorithm 1 is bounded by α, i.e.,*

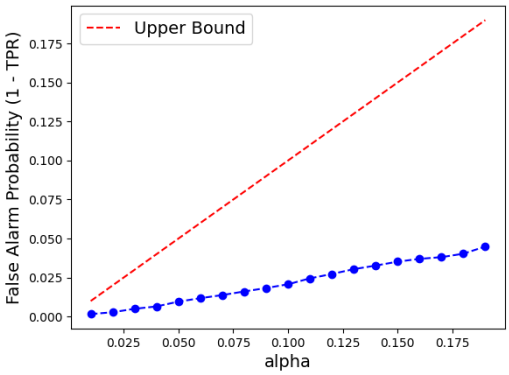
$$
\mathrm{P_F}(\mathcal{T}_{\text{cal}}) = \mathrm{P_{H_0}}\left(\textit{declare OOD} \,|\mathcal{T}_{\text{cal}}\right) \leq \alpha, \tag{25}
$$

with probability $1 - \delta$.

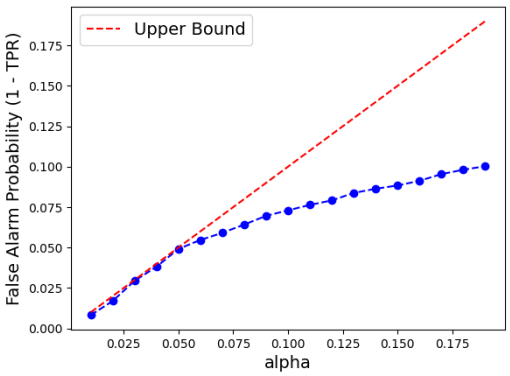
We adapt the proof of FDR control for the BH procedure provided by Benjamini and Yekutieli (2001) for our algorithm, to the use of conformal p-values estimated from the calibration set instead of the actual p-values in Algorithm 1. The details of the proof are presented in the Appendix.



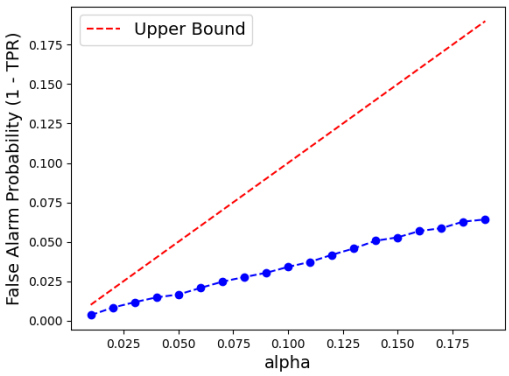
(a) ResNet with CIFAR10



(b) DenseNet with CIFAR10



(c) ResNet with SVHN



(d) DenseNet with SVHN

Figure 2: False Alarm probabilities with CIFAR10 and SVHN as in-distribution datasets for ResNet and DenseNet

3.4 Empirical Validation

We verify the results in Theorem 2 through experiments with CIFAR10 and SVHN as in-distribution datasets, and ResNet and DenseNet architectures (more details on the experimental setup are given in Section 4). In Figure 2, we plot the false alarm probabilities when the thresholds for comparing the conformal p-values are set according to Algorithm 1. The dashed line represents the theoretical upper bound on the false alarm probability. As seen in Figure 2, the false alarm probability is bounded by the theoretical upper bound as stated in Theorem 2 for all settings considered. Note that the results in this paper hold for

any given ML model, and while the bound may be conservative for certain settings (e.g., DenseNet with CIFAR10), it is tight in other cases (e.g., ResNet with SVHN).

Such strong theoretical guarantees are absent in most prior work on OOD detection. A few works that have suggested the use of conformal p-values for OOD detection, such as the work by Kaur et al. (2022), provide marginal false alarm guarantees of the form:

$$\mathrm{P}_{\mathrm{H}_0}\{\text{declare OOD }\} = \mathrm{E}\left[\mathrm{P}_{\mathrm{H}_0}\{\text{declare OOD } |\mathcal{T}_{\mathrm{cal}}\}\right] \leq \alpha \tag{26}$$

where the expectation is over all possible calibration sets. (See also the discussion surrounding (16).) However, this does not guarantee that the false alarm level α is maintained with high probability for the particular calibration dataset used. In addition, it does not provide any information on the size of the calibration dataset to be used.

4. Experimental Evaluation

In the previous section, we have provided guarantees on the strong probability of false alarm for Algorithm 1. However, since it is not possible to theoretically analyze the power of such a test (due to the structure of the alternate hypothesis), we evaluate the power of our proposed approach through experiments. In addition, since we do not know beforehand what kind of OOD samples might arise at inference time, an effective OOD detection test must perform uniformly well across different OOD datasets, for a given deep neural network (DNN) architecture. In our experiments, we evaluate both of these metrics to demonstrate the effectiveness of our approach.

Following the standard protocol for OOD detection (Lee et al., 2018; Sastry and Oore, 2020; Liu et al., 2020), we consider settings with CIFAR10 and SVHN as the in-distribution datasets.

- For CIFAR10 as the in-distribution dataset, we study SVHN, LSUN, ImageNet, and iSUN as OOD datasets.
- For SVHN as the in-distribution dataset, we study LSUN, ImageNet, CIFAR10 and iSUN as OOD datasets.

We evaluate the detection performance on two pre-trained architectures: ResNet34 (He et al., 2016) and DenseNet (Huang et al., 2017). The calibration dataset in each case is a subset of 5000 samples of the in-distribution training dataset.

We evaluate the proposed approach, and compare it with baseline methods based on the standard metric of probability of detection $\mathrm{P_D}$ or power (i.e., probability of correctly detecting an OOD sample) at probability of false alarm $\mathrm{P_F}$ at 0.1. Note that in some prior works on OOD detection, the probability of detection is referred to as the True Negative Rate (TNR) and $1 - \mathrm{P_F}$ as the True Positive Rate (TPR), where the in-distribution samples are considered positives, and OOD samples are considered negatives.

Recall that we focus exclusively on methods that do **not** have any outlier exposure to OOD samples (see Hendrycks et al., 2019; Liang et al., 2018; Lee et al., 2018), and can be applied to *any* pre-trained ML model. We compare our approach against baselines: Mahalanobis (Lee et al., 2018), Gram matrix (Sastry and Oore, 2020), and Energy (Liu et al., 2020). For the Mahalanobis baseline, we use the scores from the penultimate layer of the network to maintain uniformity.

To evaluate our proposed method, we systematically combine the following test statistics using our multiple testing approach as detailed in Algorithm 1:

1. Mahalanobis distances from individual DNN layers (Lee et al., 2018): Let $g_i(X), i = 1, \ldots, L$ denote the outputs of the intermediate layers of the neural network for an input X. We estimate μ_i^c, the class-wise mean of $g_i(.)$, as the empirical class-wise mean from the training dataset:

$$\mu_i^c = \frac{1}{n_c} \sum_{j:Y_j=c} g_i(X_j), \tag{27}$$

where n_c is the number of points with label c. We estimate the common covariance Σ for all classes as

$$\Sigma = \frac{1}{n_c} \sum_{c} \sum_{j:Y_j=c} \left(g_i(X_j) - \mu_c\right) \left(g_i(X_j) - \mu_c\right)^T. \tag{28}$$

The Mahalanobis score for layer i is calculated as:

$$\max_c - \left(g_i(X_j) - \mu_c\right) \Sigma^{-1} \left(g_i(X_j) - \mu_c\right)^T. \tag{29}$$

We calculate 5 Mahalanobis scores from the intermediate layers for the ResNet34 architecture, and 4 scores for the DenseNet architecture.

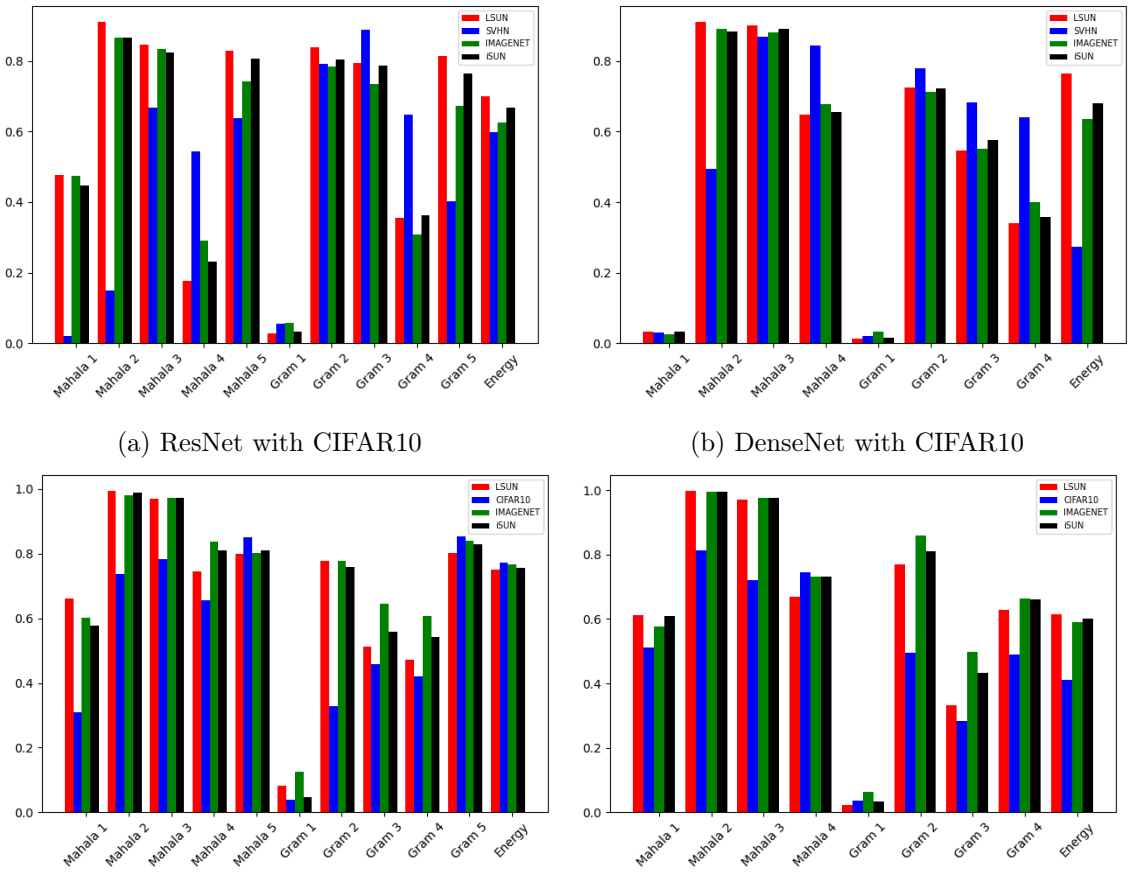
2. Gram matrix deviations from the individual DNN layers (Sastry and Oore, 2020): For each intermediate layer i, the Gram matrix of order p is calculated as:

$$M_i^p(x) = \left(g_i^p g_i^{pT}\right)^{\frac{1}{p}}, \tag{30}$$

where the power is calculated element-wise. For each flattened upper triangular Gram matrix $\overline{M_i^p}$, there are n_i correlations. The class-specific minimum and maximum values for the correlation j (i.e., j-th element of $\overline{M_i^p}$), class c, layer i and power p are estimated from the training dataset as $\min[c][i][p][j]$ and $\max[c][i][p][j]$, respectively. For a new input X, the deviation for correlation j, layer i, power p is calculated with respect to the predicted class c_X as

$$\delta_X(i,p,j) = \begin{cases} \frac{\overline{M_i^p(X)}[j] - \min[c_X][i][p][j]}{|\min[c_X][i][p][j]|} & \text{if } \overline{M_i^p(X)}[j] > \min[c_X][i][p][j] \\ \frac{\max[c_X][i][p][j] - \overline{M_i^p(X)}[j]}{|\max[c_X][i][p][j]|} & \text{if } \overline{M_i^p(X)}[j] < \max[c_X][i][p][j] \\ 0 & \text{otherwise.} \end{cases} \tag{31}$$

As proposed by Sastry and Oore (2020), the Gram matrix score for layer i is then calculated as the sum of $\delta_X(i,p,j)$ over values of p from 1 to 10, for all values of j, and normalized by the empirical mean of $\delta_X(i,p,j)$. We calculate 5 Gram scores from the intermediate layers for the ResNet34 architecture, and 4 scores for the DenseNet architecture.



(a) ResNet with CIFAR10

(b) DenseNet with CIFAR10

(c) ResNet with SVHN

(d) DenseNet with SVHN

Figure 3: Probabilities of scores rejected with CIFAR10 and SVHN as in-distribution datasets for ResNet and DenseNet

3. Energy statistic (Liu et al., 2020): The energy score is a temperature scaled log-sum-exponent of the softmax scores

$$-T \log \sum_{i=1}^{C} e^{\sigma_i(X)/T}, \tag{32}$$

where C is the number of classes, $\sigma_i(.)$ are the softmax scores, and T is the temperature parameter. In our experiments, we set the temperature parameter T to 100 for all in-distribution datasets, DNN architectures and OOD datasets (as stated by Liu et al. (2020), the energy score is not sensitive to the temperature parameter).

We use a subset of 45000 points from the training dataset (with no overlap with the calibration dataset) to calculate the class-wise empirical means and shared covariance for the Mahalanobis scores, and the minimum and maximum correlations for the Gram scores.

For CIFAR10 and SVHN as the in-distribution datasets, we use our proposed method in Algorithm 1 to combine the Mahalanobis scores and Gram scores across layers, and the energy score, to detect OOD samples. There are 11 scores (i.e., $K = 11$) in total for the ResNet34 architecture, and 9 scores (i.e., $K = 9$) for the DenseNet architecture. Recall that Algorithm 1 declares an input to be an OOD sample if any of the K null hypotheses corresponding to the K scores are rejected. For different OOD datasets, we empirically study the probability of each null hypothesis being rejected by Algorithm 1. In Figure 3, we plot the empirical probability of each score i being rejected, i.e., the proportion of data points in each OOD dataset for which the corresponding null hypothesis $H_{0,i}$ was rejected. The Mahalanobis score and Gram score of layer i are denoted by 'Mahala i' and 'Gram i' respectively, and the energy score is denoted by 'Energy'. We observe that while the probability of a hypothesis corresponding to a particular score being rejected is high for certain OOD datsets, there exist instances from other OOD datasets for which it is quite low. For example, in the Resnet34 architecture with CIFAR10 as the in-distribution dataset, while the Mahalanobis scores of layers 2, 3 and 5, and the Gram scores of layer 5 are useful to detect OOD instances from the LSUN, ImageNet and iSUN datasets, they are not likely to be useful in detecting OOD instances from the SVHN dataset. On the other hand, the Mahalanobis and Gram scores from layer 4 of the network are more useful in detecting OOD instances from the SVHN dataset than from the LSUN, ImageNet and iSUN datasets. This study provides evidence that a single score may not be useful to detect all kinds of OOD instances that an ML model might encounter at inference time, and combining different scores systematically, as proposed in Algorithm 1, might lead to a more robust OOD detection method. We demonstrate an improvement in detection performance and the robustness of our proposed OOD detection method through extensive experiments presented further in this section.

In addition to baseline methods from previous works on OOD detection, we further compare our proposed method of combining multiple scores in Algorithm 1 with a baseline that combines scores naively through an averaging rule. This naive OOD detection test maintains thresholds $\tau_1, \ldots, \tau_K$ for the K scores. Let γ_i be the weight for the i-th score where

$$\gamma_i = \mathbf{1}_{\{T_i \geq \tau_i\}}, \tag{33}$$

and let γ be defined as

$$\gamma = \frac{1}{K} \sum_{i=1}^{K} \gamma_i. \tag{34}$$

The naive averaging OOD detection rule declares an input to be an OOD sample if $\gamma \geq \frac{1}{2}$. The thresholds $\tau_1, \ldots, \tau_K$ are set to ensure a false alarm probability of 0.1.

It is also possible to construct an OOD detection test adapted from the Bonferroni procedure similar to Algorithm 1, by replacing m with:

$$m = \left| \left\{ i : \hat{Q}^i \leq \frac{\alpha}{(1+\epsilon)K} \right\} \right|, \tag{35}$$

i.e., calculating m as the number of hypotheses i for which the corresponding conformal p-value is smaller than the constant $\frac{\alpha}{(1+\epsilon)K}$. A sample is declared as OOD if $m \geq 1$. This procedure is detailed in Algorithm 2 for completeness. We can provide guarantees on the conditional false alarm probability for Algorithm 2 as well (see Appendix D).

The detection power performances for CIFAR10 and SVHN as in-distribution datasets are presented in Tables 1 and 2, for the Mahalanobis, Gram and Energy baselines, naive averaging method, Bonferroni-inspired procedure (Algorithm 2) and our BH-inspired method (Algorithm 1) of combining different statistics. Both the naive averaging rule and the Bonferroni inspired method use the Mahalanobis and Gram scores from all the layers, and the energy score. We annotate our method with the number of statistics used, e.g., Mahalanobis, Gram and Energy (5/4+5/4+1) uses 5,4 layers in ResNet34, DenseNet architectures respectively, for both Mahalanobis and Gram, and the energy score. For each in-distribution dataset, we consider 8 cases, comprising of 4 OOD Datasets and 2 different DNN architectures.

1. **Improvement in probability of detection across OOD datasets and DNN architectures:** The best probability of detection in all 8 cases with CIFAR10 as in-distribution correspond to our method of combining statistics. Similarly, with SVHN as in-distribution, our method of combining statistics gives the best probability of detection in all 8 cases. Thus, our approach leads to an improvement across OOD datasets and DNN architectures.

2. **Lower variation in detection probability across OOD datasets and DNN architectures:** Detection probabilities of baselines Mahalanobis, Gram and Energy exhibit a much higher variation across different kinds of OOD samples as compared to the combination of all statistics.

   With CIFAR10 as the in-distribution dataset, for the ResNet34 architecture: the variation in $\mathrm{P_D}$ is $82.77 - 90.97$ for the Mahalanobis baseline, $92.34 - 96.04$ for the Gram baseline, and $73.21 - 81.16$ for the energy baseline. In contrast, our method of combining all statistics has a variation of $97.03 - 98.00$. For DenseNet, the variation in $\mathrm{P_D}$ is $82.81 - 92.98$ for the Mahalanobis baseline, $80.04 - 89.97$ for the Gram baseline, and $42.40 - 96.89$ for the energy baseline. Our method of combining all statistics has a variation of $94.57 - 97.78$.

   A similar trend is seen with SVHN as the in-distribution dataset. Thus, we see that while the baseline methods have a high variation in the detection performance across different OOD datasets, our method of combining all statistics performs uniformly well across OOD datasets. This is a key improvement, as the kind of OOD samples encountered at inference time is unknown, and our proposed method shows very little variation across different OOD datasets.

3. **Comparison with naive averaging:** We observe that the naive averaging method does not perform as well as our proposed method of combining statistics, and indeed has a high variation in its detection performance across different OOD datasets. Thus, we see that while it is imperative to combine multiple scores for effective and robust OOD detection, combining them in an ad hoc manner such as uniform averaging does not yield good results.

4. **Comparison with Bonferroni-inspired procedure:** In general, the Bonferroni procedure has been observed to have a smaller detection power as compared to the BH procedure (see Sec 4 Benjamini and Hochberg, 1995). Indeed, we observe that the

Table 1: Comparison with baseline methods for CIFAR10 as in-distribution. Each entry is $P_D(\%)$ **at** $P_F = 10\%$.

OOD Dataset	**Method**	**ResNet34**	**DenseNet**
SVHN	Mahala (penultimate layer)	82.77	92.98
	Gram (sum across layers)	96.04	89.97
	Energy	73.21	42.40
	Naive Averaging (5/4 + 5/4 + 1)	81.13	83.28
	Bonferroni - Mahala, Gram and Energy (5/4+5/4+1)	96.41	91.13
	Ours - Mahala (5/4)	87.92	93.16
	Ours - Gram (5/4)	95.61	89.90
	Ours - Mahala, Energy (5/4 + 1)	91.88	94.03
	Ours - Gram, Energy (5/4 + 1)	96.78	90.77
	Ours - Mahala, Gram (5/4 + 5)	96.23	94.21
	Ours - Mahala, Gram and Energy (5/4+5/4+1)	97.13	94.57
ImageNet	Mahala (penultimate layer)	85.45	82.81
	Gram (sum across layers)	92.34	80.04
	Energy	76.76	94.93
	Naive Averaging (5/4 + 5/4 + 1)	86.45	80.96
	Bonferroni - Mahala, Gram and Energy (5/4+5/4+1)	95.92	95.89
	Ours - Mahala (5/4)	96.90	95.19
	Ours - Gram (5/4)	92.60	80.12
	Ours - Mahala, Energy (5/4 + 1)	97.28	98.09
	Ours - Gram, Energy (5/4 + 1)	94.53	95.19
	Ours - Mahala, Gram (5/4 + 5)	96.38	92.81
	Ours - Mahala, Gram and Energy (5/4+5/4+1)	97.03	97.20
LSUN	Mahala (penultimate layer)	90.97	84.11
	Gram (sum across layers)	95.94	81.83
	Energy	81.16	96.89
	Naive Averaging (5/4 + 5/4 + 1)	91.31	83.79
	Bonferroni - Mahala, Gram and Energy (5/4+5/4+1)	96.99	96.53
	Ours - Mahala (5/4)	98.11	96.38
	Ours - Gram (5/4)	96.16	81.67
	Ours - Mahala, Energy (5/4 + 1)	97.87	98.20
	Ours - Gram, Energy (5/4 + 1)	96.61	96.43
	Ours - Mahala, Gram (5/4 + 5/4)	98.02	94.40
	Ours - Mahala, Gram and Energy (5/4+5/4+1)	98.00	97.78
iSUN	Mahala (penultimate layer)	89.99	83.19
	Gram (sum across layers)	95.10	81.47
	Energy	80.11	95.10
	Naive Averaging (5/4 + 5/4 + 1)	89.22	81.70
	Bonferroni - Mahala, Gram and Energy (5/4+5/4+1)	96.76	94.79
	Ours - Mahala (5/4)	97.24	95.26
	Ours - Gram (5/4)	95.11	81.09
	Ours - Mahala, Energy (5/4 + 1)	97.17	97.12
	Ours - Gram, Energy (5/4 + 1)	96.19	94.73
	Ours - Mahala, Gram (5/4 + 5/4)	97.36	92.93
	Ours - Mahala, Gram and Energy (5/4+5/4+1)	97.67	96.34

Table 2: Comparison with baseline methods for SVHN as in-distribution. Each entry is $P_D(\%)$ **at** $P_F = 10\%$.

OOD Dataset	**Method**	**ResNet34**	**DenseNet**
ImageNet	Mahala (penultimate layer)	96.12	96.34
	Gram (sum across layers)	97.52	93.57
	Energy	85.14	70.53
	Naive Averaging (5/4 + 5/4 + 1)	97.08	95.67
	Bonferroni - Mahala, Gram and Energy (5/4+5/4+1)	99.72	99.79
	Ours - Mahala (5/4)	99.91	99.95
	Ours - Gram (5/4)	97.68	94.38
	Ours - Mahala, Energy (5/4 + 1)	99.89	99.93
	Ours - Gram, Energy (5/4 + 1)	97.85	95.01
	Ours - Mahala, Gram (5/4 + 5/4)	99.83	99.91
	Ours - Mahala, Gram and Energy (5/4+5/4+1)	99.84	99.89
LSUN	Mahala (penultimate layer)	93.74	94.17
	Gram (sum across layers)	96.20	88.25
	Energy	81.30	71.36
	Naive Averaging (5/4 + 5/4 + 1)	95.00	92.81
	Bonferroni - Mahala, Gram and Energy (5/4+5/4+1)	99.89	99.97
	Ours - Mahala (5/4)	99.98	100.0
	Ours - Gram (5/4)	96.54	89.02
	Ours - Mahala, Energy (5/4 + 1)	99.96	99.99
	Ours - Gram, Energy (5/4 + 1)	96.82	90.56
	Ours - Mahala, Gram (5/4 + 5/4)	99.96	99.98
	Ours - Mahala, Gram and Energy (5/4+5/4+1)	99.95	100.0
iSUN	Mahala (penultimate layer)	95.23	96.01
	Gram (sum across layers)	96.50	91.46
	Energy	82.79	71.20
	Naive Averaging (5/4 + 5/4 + 1)	96.00	94.53
	Bonferroni - Mahala, Gram and Energy (5/4+5/4+1)	99.88	99.98
	Ours - Mahala (5/4)	99.98	100.0
	Ours - Gram (5/4)	96.80	91.89
	Ours - Mahala, Energy (5/4 + 1)	99.93	100.0
	Ours - Gram, Energy (5/4 + 1)	97.21	92.69
	Ours - Mahala, Gram (5/4 + 5/4)	99.88	99.98
	Ours - Mahala, Gram and Energy (5/4+5/4+1)	99.88	99.98
CIFAR10	Mahala (penultimate layer)	96.09	94.25
	Gram (sum across layers)	91.58	69.77
	Energy	83.31	54.07
	Naive Averaging (5/4 + 5/4 + 1)	86.10	77.22
	Bonferroni - Mahala, Gram and Energy (5/4+5/4+1)	95.84	91.77
	Ours - Mahala (5/4)	98.31	97.64
	Ours - Gram (5/4)	92.39	72.84
	Ours - Mahala, Energy (5/4 + 1)	98.13	97.16
	Ours - Gram, Energy (5/4 + 1)	92.91	78.03
	Ours - Mahala, Gram (5/4 + 5)	97.15	94.83
	Ours - Mahala, Gram and Energy (5/4+5/4+1)	97.35	95.23

Bonferroni inspired procedure in Algorithm 2 does not perform as well as our proposed BH-inspired method of combining statistics from Algorithm 1.

5. **Impact of combining all the scores using Algorithm 1:** For CIFAR10 as the in-distribution dataset, in all 8 cases, combining all the scores - Mahalanobis and Gram from individual layers, and the energy score, is either the best method, or within 1% of the best performance.

   Similarly, with SVHN as the in-distribution dataset, in 7 out of 8 cases, combining all the scores is either the best method, or within 1% of the best performance (the gap is 2.41% in the remaining case).

   Thus, in contrast to existing methods, combining all the statistics using Algorithm 1 is robust to different kinds of OOD samples across DNN architectures.

In some of the prior work on OOD detection, the Area Under the Receiver Operating Characteristic (AUROC) metric has been used to compare different tests (Liang et al., 2018; Liu et al., 2020; Sastry and Oore, 2020; Lee et al., 2018). However, it is not clear that this measure is useful in such a comparison, especially when the ROC is being estimated through simulations. It is possible for a test (say, Test 1) to have a larger AUROC than another test (say, Test 2), with Test 2 having a larger detection power than Test 1 for all values of false alarm less than some threshold (equivalently, all values of TPR greater than some threshold). Nevertheless, we provide the AUROC numbers for our experimental setups below for completeness.

Table 3 contains the AUROC numbers for CIFAR10 as the in-distribution dataset, and Table 4 contains the AUROC numbers for SVHN as the in-distribution dataset. We observe similar patterns in the AUROC numbers as the above observations on the detection power at a fixed false alarm probability. The Mahalanobis, Gram and Energy baselines have a high variability across different kinds of OOD samples and DNN architectures, whereas our proposed method of combining all statistics has a low variability. Our proposed method of combining all statistics either has the best AUROC performance or within 1% of the best performance in all 8 cases for CIFAR10 and SVHN as the in-distribution datasets.

5. Conclusion

While empirical methods for OOD detection have been studied extensively in recent literature, a formal characterization of OOD is lacking. We proposed a characterization for the notion of OOD that includes both the input distribution and the ML model. This provided insights for the construction of effective OOD detection tests. Our approach, inspired by *multiple hypothesis testing*, allows us to systematically combine any number of different statistics derived from the ML model with an arbitrary dependence structure.

Furthermore, our analysis allows us to set the test thresholds to meet given constraints on the probability of incorrectly classifying an in-distribution sample as OOD (false alarm probability). We provide strong theoretical guarantees on the probability of false alarm in OOD detection, conditioned on the dataset used for computing the conformal p-values.

In our experiments, we observe that no single score is useful for detecting different kinds of OOD instances. We demonstrated that our proposed method outperforms threshold-based

Table 3: Comparison with baseline OOD detection techniques for CIFAR10 as in-distribution. Each entry is **AUROC**.

OOD Dataset	**Method**	**ResNet34**	**DenseNet**
SVHN	Mahala (penultimate layer)	93.86	96.72
	Gram (sum across layers)	97.28	94.31
	Energy	90.24	77.92
	Naive Averaging (5/4 + 5/4 + 1)	88.81	88.03
	Bonferroni - Mahala, Gram and Energy (5/4 + 5/4 + 1)	97.83	96.72
	Ours - Mahala (5/4)	95.34	96.70
	Ours - Gram (5/4)	97.47	94.28
	Ours - Mahala, Energy (5/4 + 1)	95.84	96.99
	Ours - Gram, Energy (5/4 + 1)	97.90	96.20
	Ours - Mahala, Gram (5/4 + 5/4)	97.56	96.98
	Ours - Mahala, Gram and Energy (5/4+5/4+1)	97.76	97.24
ImageNet	Mahala (penultimate layer)	94.84	93.12
	Gram (sum across layers)	95.90	89.83
	Energy	91.40	96.03
	Naive Averaging (5/4 + 5/4 + 1)	91.26	84.65
	Bonferroni - Mahala, Gram and Energy (5/4 + 5/4 + 1)	97.47	97.54
	Ours - Mahala (5/4)	97.89	97.32
	Ours - Gram (5/4)	96.09	89.75
	Ours - Mahala, Energy (5/4 + 1)	97.97	98.13
	Ours - Gram, Energy (5/4 + 1)	97.07	96.79
	Ours - Mahala, Gram (5/4 + 5/4)	97.55	96.67
	Ours - Mahala, Gram and Energy (5/4+5/4+1)	97.64	97.70
LSUN	Mahala (penultimate layer)	96.28	90.00
	Gram (sum across layers)	97.31	87.97
	Energy	92.35	96.83
	Naive Averaging (5/4 + 5/4 + 1)	94.30	87.64
	Bonferroni - Mahala, Gram and Energy (5/4 + 5/4 + 1)	97.81	97.69
	Ours - Mahala (5/4)	98.20	97.54
	Ours - Gram (5/4)	97.46	87.76
	Ours - Mahala, Energy (5/4 + 1)	98.07	98.16
	Ours - Gram, Energy (5/4 + 1)	97.76	97.14
	Ours - Mahala, Gram (5/4 + 5/4)	97.99	96.82
	Ours - Mahala, Gram and Energy (5/4+5/4+1)	97.96	97.74
iSUN	Mahala (penultimate layer)	96.07	93.71
	Gram (sum across layers)	97.01	90.48
	Energy	92.05	96.25
	Naive Averaging (5/4 + 5/4 + 1)	92.98	85.85
	Bonferroni - Mahala, Gram and Energy (5/4 + 5/4 + 1)	97.71	97.18
	Ours - Mahala (5/4)	97.95	97.39
	Ours - Gram (5/4)	97.15	90.36
	Ours - Mahala, Energy (5/4 + 1)	97.93	97.89
	Ours - Gram, Energy (5/4 + 1)	97.66	96.71
	Ours - Mahala, Gram (5/4 + 5/4)	97.79	96.76
	Ours - Mahala, Gram and Energy (5/4+5/4+1)	97.83	97.47

Table 4: Comparison with baseline OOD detection techniques for SVHN as in-distribution. Each entry is AUROC.

OOD Dataset	**Method**	**ResNet34**	**DenseNet**
LSUN	Mahala (penultimate layer)	96.06	96.22
	Gram	97.23	94.17
	Energy	87.58	86.01
	Naive Averaging (5/4 + 5/4 + 1)	96.85	95.30
	Bonferroni - Mahala, Gram and Energy (5/4 + 5/4 + 1)	98.17	99.07
	Ours - Mahala (5/4)	99.00	98.92
	Ours - Gram (5/4)	97.19	94.11
	Ours - Mahala, Energy (5/4 + 1)	98.76	98.94
	Ours - Gram, Energy (5/4 + 1)	97.47	95.69
	Ours - Mahala, Gram (5/4 + 5/4)	98.82	99.08
	Ours - Mahala, Gram and Energy (5/4+5/4+1)	98.21	99.06
ImageNet	Mahala (penultimate layer)	96.81	97.01
	Gram	97.75	96.34
	Energy	90.33	85.76
	Naive Averaging (5/4 + 5/4 + 1)	97.91	96.90
	Bonferroni - Mahala, Gram and Energy (5/4 + 5/4 + 1)	98.28	99.00
	Ours - Mahala (5/4)	98.99	98.89
	Ours - Gram (5/4)	97.73	96.32
	Ours - Mahala, Energy (5/4 + 1)	98.79	98.91
	Ours - Gram, Energy (5/4 + 1)	98.01	97.11
	Ours - Mahala, Gram (5/4 + 5/4)	98.87	99.04
	Ours - Mahala, Gram and Energy (5/4+5/4+1)	98.25	99.02
iSUN	Mahala (penultimate layer)	96.49	96.85
	Gram	97.40	95.47
	Energy	88.75	85.69
	Naive Averaging (5/4 + 5/4 + 1)	97.15	96.41
	Bonferroni - Mahala, Gram and Energy (5/4 + 5/4 + 1)	98.19	99.05
	Ours - Mahala (5/4)	98.99	98.91
	Ours - Gram (5/4)	97.37	95.42
	Ours - Mahala, Energy (5/4 + 1)	98.76	98.94
	Ours - Gram, Energy (5/4 + 1)	97.63	96.40
	Ours - Mahala, Gram (5/4 + 5/4)	98.82	99.07
	Ours - Mahala, Gram and Energy (5/4+5/4+1)	98.21	99.05
CIFAR10	Mahala (penultimate layer)	96.90	96.59
	Gram	95.35	87.06
	Energy	89.09	77.72
	Naive Averaging (5/4 + 5/4 + 1)	92.42	85.96
	Bonferroni - Mahala, Gram and Energy (5/4 + 5/4 + 1)	96.83	96.23
	Ours - Mahala (5/4)	97.63	97.50
	Ours - Gram (5/4)	95.35	87.21
	Ours - Mahala, Energy (5/4 + 1)	97.68	97.43
	Ours - Gram, Energy (5/4 + 1)	95.94	91.15
	Ours - Mahala, Gram (5/4 + 5/4)	97.32	96.91
	Ours - Mahala, Gram and Energy (5/4+5/4+1)	97.10	96.98

tests for OOD detection proposed in prior work. Across different kinds of OOD examples, we observed that the state-of-the-art methods from prior work exhibit high variability across OOD instances and neural network architectures in their probability of detection of OOD samples. In contrast, our proposed method is robust and provides uniformly good performance (with respect to both detection power and AUROC) across different kinds of OOD samples and neural network architectures. This robustness is important, since a useful OOD detection algorithm should perform well regardless of the type of OOD instance encountered at inference time.

Acknowledgments

This work was supported by the U.S. Army Research Laboratory Cooperative Research Agreement W911NF-17-2-0196, and the U.S. National Science Foundation(NSF) grant #2106727. Any opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the Department of Defense or the United States Government. The authors would like to thank Aditya Deshmukh for useful discussions.

Appendix A. Preliminaries on Multiple Testing

Multiple hypothesis testing (a.k.a. multiple testing) refers to the inference problem testing between multiple binary hypotheses, e.g., $\mathrm{H}_{0,i}$ versus $\mathrm{H}_{1,i}$, $i = 1, 2, \ldots, K$. For a given multiple testing procedure, let R be the number of null hypotheses rejected (i.e., number of tests declared as the alternative $H_{1,i}$), out of which V is the number of true null hypotheses. Some measures of performance for multiple testing procedures are as follows:

1. Family Wise Error Rate (FWER): The probability of rejecting at least one null hypothesis when all of them are true.

2. False Discovery Rate (FDR): Expected ratio of number of true null hypotheses rejected (V) and the total number of hypotheses rejected (R), i.e.,

$$\mathrm{FDR} = \mathrm{E}\left[\frac{V}{R}\mathbf{1}_{\{R>0\}}\right]. \tag{36}$$

   where the expectation is taken over the joint distribution of the statistics involved in the multiple testing problem.

When all the null hypotheses are true, $V = R$ with probability 1, and:

$$\mathrm{FDR} = \mathrm{E}\left[\mathbf{1}_{\{R>0\}}\right] = \mathrm{P}\{R > 0\} = \mathrm{FWER}.$$

Various multiple testing procedures have been proposed in literature depending on the quantity of interest to be controlled. Widely used multiple testing procedures involve calculating the p-values for each test i as Q^i, and combining these p-values to give decisions for each hypothesis. Let $\alpha > 0$. One of the earliest tests proposed to control the FWER

is the Bonferroni test. In this test, each Q^i is computed, and for each $i = 1, \ldots, K$, the corresponding hypothesis $\mathrm{H}_{0,i}$ is rejected if

$$Q^i \leq \frac{\alpha}{K}. \tag{37}$$

This test controls the FWER at α for any joint distribution of the test statistics of the K hypotheses. However, the power of this test has been observed to be low, and hence the test is considered to be conservative. The FDR measure was proposed by Benjamini and Hochberg (1995), who also proposed a procedure to control the FDR. Let the p-values for each test i be Q^i, and let the ordered p-values be denoted by $Q^{(1)}, Q^{(2)}, \ldots, Q^{(K)}$. Let

$$m = \max \left\{ i : Q^{(i)} \leq \frac{\alpha i}{K} \right\}. \tag{38}$$

The Benjamini-Hochberg (BH) procedure rejects hypotheses $\mathrm{H}_{0,1}, \ldots, \mathrm{H}_{0,m}$, and controls the FDR at level α when the test statistics are independent. Benjamini and Yekutieli (2001) showed that the constants in the BH procedure can be modified to $\frac{\alpha i}{K \sum_{j=1}^{K} \frac{1}{j}}$ instead of $\frac{\alpha i}{K}$ to control the FDR at level α for arbitrarily dependent test statistics. Note that the Bonferroni procedure and the BH procedure can be used to test against the global null H_0 (all $\mathrm{H}_{0,\mathrm{i}}$ are true), where the probability of false alarm $\mathrm{P}_{\mathrm{H}_0}(\text{reject } \mathrm{H}_0)$ is equal to the FWER and FDR. Our proposed algorithm for the OOD detection problem builds on the BH procedure with the modified constants, and conformal p-values calculated using a calibration dataset $\mathcal{T}_{\mathrm{cal}}$, where we aim to control the conditional probability of false alarm $\mathrm{P}_{\mathrm{F}}(\mathcal{T}_{\mathrm{cal}}) = \mathrm{P}_{\mathrm{H}_0}(\text{reject } \mathrm{H}_0 | \mathcal{T}_{\mathrm{cal}})$ with high probability.

Appendix B. Proposed OOD Modeling

In Section 2, we conclude that functions of the input from the ML model apart from the final output are required for the OOD formulation presented above. Note that this does not violate the data-processing inequality, as the out-distribution $\mathrm{P}_{X,\hat{Y}}$ characterizes the input and the model, and these functions of the input give us additional information regarding the ML model. In addition, these functions give us information to differentiate between the null and the alternate hypothesis.

Appendix C. Proof of Theorem 2

For $\ell = 1, 2, \ldots, K$, let

$$\alpha_\ell = \frac{\alpha \ell}{C(K) K}, \tag{39}$$

where

$$C(K) = (1 + \epsilon) \sum_{j=1}^{K} \frac{1}{j} \tag{40}$$

and let $\alpha_0 = 0$. As in (17), the probability of false alarm conditioned on the calibration set $\mathcal{T}_{\mathrm{cal}}$ is given by

$$\mathrm{P}_{\mathrm{F}}(\mathcal{T}_{\mathrm{cal}}) = \mathrm{P}_{\mathrm{H}_0}(\text{reject } \mathrm{H}_0 | \mathcal{T}_{\mathrm{cal}}) = \mathrm{P}_{\mathrm{H}_0}(m \geq 1 | \mathcal{T}_{\mathrm{cal}}), \tag{41}$$

where m is as defined in Algorithm 1. Here H_0 denotes the global null hypothesis, which corresponds to all the $\mathrm{H}_{0,i}$ being true. Note that $m \geq 1$ signifies that $\mathrm{H}_{0,(1)}, \ldots, \mathrm{H}_{0,(m)}$ are being rejected. Let

$$A_\ell = \{\text{exactly } \ell \text{ of the } \mathrm{H}_{0,i}\text{'s are rejected}\}.$$

Then,

$$\mathrm{P_F}(\mathcal{T}_{\text{cal}}) = \sum_{\ell=1}^{K} \mathrm{P}_{\mathrm{H}_0}(A_\ell | \mathcal{T}_{\text{cal}}). \tag{42}$$

The following lemma is useful in deriving an upper bound for $\mathrm{P_F}$.

Lemma 3 *For $\ell = 1, \ldots, K$,*

$$\mathrm{P}_{\mathrm{H}_0}(A_\ell) = \frac{1}{\ell} \sum_{i=1}^{K} \mathrm{P}_{\mathrm{H}_0}(\{\hat{Q}^i \leq \alpha_\ell\} \cap A_\ell), \tag{43}$$

where $\hat{Q}^i$ is as defined in Section 3.

Proof

Let

$$\mathcal{W}_\ell = \{\text{all subsets of } \{1, 2, \ldots, K\} \text{ with } \ell \text{ elements}\}.$$

Let $A_\ell^{\mathcal{V}}$ be the subset of A_ℓ where the ℓ null hypotheses rejected correspond to the indices in $\mathcal{V} \in \mathcal{W}_\ell$. Then

$$A_\ell = \bigcup_{\mathcal{V} \in \mathcal{W}_\ell} A_\ell^{\mathcal{V}}. \tag{44}$$

Note that if ℓ null hypotheses corresponding to the indices in $\mathcal{V} \in \mathcal{W}_\ell$ are rejected, then the conformal p-values corresponding to these ℓ tests are less than or equal to α_ℓ (since the maximum among them is less than or equal to α_ℓ), and the conformal p-values corresponding to the remaining $K - \ell$ tests are greater than α_ℓ, i.e.,

$$\hat{Q}^i \leq \max_{j \in \mathcal{V}} \hat{Q}^j \leq \alpha_\ell \quad \text{for } i \in \mathcal{V}, \tag{45}$$

and

$$\hat{Q}^i > \alpha_\ell \qquad \text{for } i \notin \mathcal{V}. \tag{46}$$

Thus,

$$\mathrm{P}_{\mathrm{H}_0}(\{\hat{Q}^i \leq \alpha_\ell\} \cap A_\ell^{\mathcal{V}}) = \begin{cases} \mathrm{P}_{\mathrm{H}_0}(A_\ell^{\mathcal{V}}) & \text{if } i \in \mathcal{V} \\ 0 & \text{else.} \end{cases} \tag{47}$$

Then,

$$\sum_{i=1}^{K} \mathrm{P}_{\mathrm{H}_0}(\{\hat{Q}^i \leq \alpha_\ell\} \cap A_\ell) = \sum_{i=1}^{K} \sum_{\mathcal{V} \in \mathcal{W}_\ell} \mathrm{P}_{\mathrm{H}_0}(\{\hat{Q}^i \leq \alpha_\ell\} \cap A_\ell^{\mathcal{V}}) \tag{48}$$

$$= \sum_{\mathcal{V} \in \mathcal{W}_\ell} \sum_{i=1}^{K} \mathrm{P}_{\mathrm{H}_0}(\{\hat{Q}^i \leq \alpha_v\} \cap A_\ell^{\mathcal{V}}) \tag{49}$$

$$= \sum_{\mathcal{V} \in \mathcal{W}_\ell} \sum_{i=1}^{K} \mathbf{1}_{\{i \in \mathcal{V}\}} \mathrm{P}_{\mathrm{H}_0}(A_\ell^{\mathcal{V}}) \tag{50}$$

$$= \sum_{\mathcal{V} \in \mathcal{W}_\ell} \mathrm{P}_{\mathrm{H}_0}(A_\ell^{\mathcal{V}}) \sum_{i=1}^{K} \mathbf{1}_{\{i \in \mathcal{V}\}} \tag{51}$$

$$= \sum_{\mathcal{V} \in \mathcal{W}_\ell} \mathrm{P}_{\mathrm{H}_0}(A_\ell^{\mathcal{V}})\, \ell \tag{52}$$

$$= \ell\, \mathrm{P}_{\mathrm{H}_0}(A_\ell), \tag{53}$$

where the first equality arises from the fact that A_ℓ is the union of disjoint sets $A_\ell^{\mathcal{V}}$ for $\mathcal{V} \in \mathcal{W}_\ell$, and the third equality follows from (47). ■

Using the result from Lemma 3 in the expression for $\mathrm{P}_{\mathrm{F}}(\mathcal{T}_{\mathrm{cal}})$ in (42), we obtain that

$$\mathrm{P}_{\mathrm{F}}(\mathcal{T}_{\mathrm{cal}}) = \sum_{\ell=1}^{K} \sum_{i=1}^{K} \frac{1}{\ell} \mathrm{P}_{\mathrm{H}_0}(\{\hat{Q}^i \leq \alpha_\ell\} \cap A_\ell | \mathcal{T}_{\mathrm{cal}}). \tag{54}$$

Note that by definition, $\alpha_0 < \alpha_1 < \ldots, \alpha_K$. Thus,

$$\{\hat{Q}^i \leq \alpha_\ell\} \cap A_\ell = \cup_{j=1}^{\ell} \{\hat{Q}^i \in (\alpha_{j-1}, \alpha_j]\} \cap A_\ell, \tag{55}$$

and

$$\mathrm{P}_{\mathrm{F}}(\mathcal{T}_{\mathrm{cal}}) = \sum_{i=1}^{K} \sum_{\ell=1}^{K} \frac{1}{\ell} \sum_{j=1}^{\ell} \mathrm{P}_{\mathrm{H}_0}(\{\hat{Q}^i \in (\alpha_{j-1}, \alpha_j]\} \cap A_\ell | \mathcal{T}_{\mathrm{cal}}) \tag{56}$$

$$= \sum_{i=1}^{K} \sum_{j=1}^{K} \sum_{\ell=j}^{K} \frac{1}{\ell} \mathrm{P}_{\mathrm{H}_0}(\{\hat{Q}^i \in (\alpha_{j-1}, \alpha_j]\} \cap A_\ell | \mathcal{T}_{\mathrm{cal}}) \tag{57}$$

$$\leq \sum_{i=1}^{K} \sum_{j=1}^{K} \sum_{\ell=j}^{K} \frac{1}{j} \mathrm{P}_{\mathrm{H}_0}(\{\hat{Q}^i \in (\alpha_{j-1}, \alpha_j]\} \cap A_\ell | \mathcal{T}_{\mathrm{cal}}) \tag{58}$$

$$\leq \sum_{i=1}^{K} \sum_{j=1}^{K} \frac{1}{j} \sum_{\ell=1}^{K} \mathrm{P}_{\mathrm{H}_0}(\{\hat{Q}^i \in (\alpha_{j-1}, \alpha_j]\} \cap A_\ell | \mathcal{T}_{\mathrm{cal}}). \tag{59}$$

Note that the events A_ℓ are disjoint for $\ell = 1, \ldots, K$. Thus,

$$\sum_{\ell=1}^{K} \mathrm{P}_{\mathrm{H}_0}(\{\hat{Q}^i \in (\alpha_{j-1}, \alpha_j]\} \cap A_\ell | \mathcal{T}_{\mathrm{cal}}) = \mathrm{P}_{\mathrm{H}_0}(\{\hat{Q}^i \in (\alpha_{j-1}, \alpha_j]\} \cap (\cup_{\ell=1}^{K} A_\ell) | \mathcal{T}_{\mathrm{cal}}) \tag{60}$$

$$\leq \mathrm{P}_{\mathrm{H}_0}(\{\hat{Q}^i \in (\alpha_{j-1}, \alpha_j]\} | \mathcal{T}_{\mathrm{cal}}) \tag{61}$$

$$= \mathrm{P}_{\mathrm{H}_0}(\{\hat{Q}^i \leq \alpha_j\} | \mathcal{T}_{\mathrm{cal}}) - \mathrm{P}(\{\hat{Q}^i \leq \alpha_{j-1}\} | \mathcal{T}_{\mathrm{cal}}). \tag{62}$$

Using this in (59),we get that

$$\mathrm{P}_{\mathrm{F}}(\mathcal{T}_{\mathrm{cal}}) \leq \sum_{i=1}^{K} \sum_{j=1}^{K} \frac{1}{j} (\mathrm{P}_{\mathrm{H}_0}(\{\hat{Q}^i \leq \alpha_j\} | \mathcal{T}_{\mathrm{cal}}) - \mathrm{P}_{\mathrm{H}_0}(\{\hat{Q}^i \leq \alpha_{j-1}\} | \mathcal{T}_{\mathrm{cal}})). \tag{63}$$

Let $r_j^i = \mathrm{P}_{\mathrm{H}_0}(\{\hat{Q}^i \leq \alpha_j\} | \mathcal{T}_{\mathrm{cal}})$. Then, rearranging the terms from above, we get

$$\mathrm{P}_{\mathrm{F}}(\mathcal{T}_{\mathrm{cal}}) = \sum_{i=1}^{K} \left[\sum_{j=1}^{K-1} \frac{r_j^i}{j(j+1)} + \frac{r_K^i}{K} \right]. \tag{64}$$

Note that P_{F} is a function of $\mathcal{T}_{\mathrm{cal}}$ only through random variables r_j^i. We have from Vovk (2012); Bates et al. (2023) that r_j^i follows a Beta distribution, i.e., $r_j^i \sim \mathrm{Beta}(a_j, b_j)$, where

$$a_j = \left\lfloor (n_{\mathrm{cal}} + 1) \frac{\alpha j}{C(K)K} \right\rfloor \tag{65}$$

$$b_j = (n_{\mathrm{cal}} + 1) - a_j. \tag{66}$$

The mean of this distribution is $\mu_j = \frac{a_j}{a_j + b_j}$. Let E denote the event

$$\bigcap_{i=1}^{K} \bigcap_{j=1}^{K} \left\{ r_j^i \leq (1+\epsilon) \frac{\alpha j}{C(K)K} \right\}. \tag{67}$$

When the condition on n_{cal} in Lemma 1 is satisfied, we have that

$$\mathrm{P}(E) \geq 1 - \delta. \tag{68}$$

Under the event E, we have that

$$\mathrm{P}_{\mathrm{F}}(\mathcal{T}_{\mathrm{cal}}) = \sum_{i=1}^{K} \left[\sum_{j=1}^{K-1} \frac{r_j^i}{j(j+1)} + \frac{r_K^i}{K} \right] \tag{69}$$

$$\leq \sum_{i=1}^{K} \left[\sum_{j=1}^{K-1} \frac{(1+\epsilon)\alpha}{(j+1)C(K)K} + \frac{(1+\epsilon)\alpha}{C(K)K} \right] \tag{70}$$

$$= \frac{(1+\epsilon)\alpha}{C(K)} \left[\sum_{j=1}^{K-1} \frac{1}{j+1} + 1 \right] \tag{71}$$

$$= \frac{(1+\epsilon)\alpha}{C(K)} \left(\sum_{j=1}^{K} \frac{1}{j} \right) = \alpha. \tag{72}$$

Thus, with probability greater than $1-\delta$, we have that

$$\mathrm{P_F}(\mathcal{T}_{\mathrm{cal}}) \leq \alpha. \tag{73}$$

Appendix D. Comparison with Bonferroni inspired test

The Bonferroni inspired procedure is detailed in Algorithm 2. We can provide guarantees on the conditional false alarm probability similar to Theorem 2 for Algorithm 2 as well.

Algorithm 2 Bonferroni based OOD detection test with conformal p-values

Inputs:
New input X_{test};
Scores over $\mathcal{T}_{cal}$ as $\left\{\{T_j^1 = s^1(X_j) : j \in \mathcal{T}_{\mathrm{cal}}\}, \ldots, \{T_j^K = s^K(X_j) : j \in \mathcal{T}_{\mathrm{cal}}\}\right\}$;
ML model $f(\mathbf{W}, .)$;
Desired conditional probability of false alarm $\alpha \in (0,1)$.
Algorithm:
For X_{test}, compute scores T_{test}^i.
Calculate conformal p-values as:

$$\hat{Q}^i = \frac{1 + |\{j \in \mathcal{T}_{\mathrm{cal}} : T_j^i \geq T_{\mathrm{test}}^i\}|}{1 + |\mathcal{T}_{\mathrm{cal}}|}. \tag{74}$$

Calculate $m = \left|\left\{i : \hat{Q}^i \leq \frac{\alpha}{(1+\epsilon)K}\right\}\right|$.
Output:
Declare OOD if $m \geq 1$.

Theorem 4 *Let $\alpha, \delta \in (0,1)$. Let $\mathcal{T}_{\mathrm{cal}}$ be a calibration set, and let n_{cal} be such that for a given $\delta > 0$,*

$$I_{(1+\epsilon)\mu}(a,b) \geq 1 - \frac{\delta}{K}, \tag{75}$$

where $a = \left\lfloor (n_{\mathrm{cal}}+1)\frac{\alpha}{(1+\epsilon)K} \right\rfloor$, $b = (n_{\mathrm{cal}}+1) - a$, $\mu = \frac{a}{a+b}$, and $I_x(a,b)$ is the CDF of a Beta distribution with parameters a, b. Then, for a new input X_{test} and a ML model $f(\mathbf{W}, .)$, the probability of incorrectly detecting X_{test} as OOD conditioned on $\mathcal{T}_{\mathrm{cal}}$ while using Algorithm 2 is bounded by α, i.e.,

$$\mathrm{P_{H_0}}\left(\textit{declare OOD} \,|\mathcal{T}_{\mathrm{cal}}\right) \leq \alpha, \tag{76}$$

with probability $1-\delta$.

Proof We have that

$$\begin{aligned}
\mathrm{P_F}(\mathcal{T}_{\mathrm{cal}}) &= \mathrm{P}_{\mathrm{H}_0}(\text{reject } \mathrm{H}_0|\mathcal{T}_{\mathrm{cal}}) && (77)\\
&= \mathrm{P}_{\mathrm{H}_0}(m \geq 1|\mathcal{T}_{\mathrm{cal}}) && (78)\\
&= \mathrm{P}_{\mathrm{H}_0}\left(\bigcup_{i=1}^{K}\left\{\hat{Q}^i \leq \frac{\alpha}{(1+\epsilon)K}\right\}\middle|\mathcal{T}_{\mathrm{cal}}\right) && (79)\\
&\leq \sum_{i=1}^{K}\mathrm{P}_{\mathrm{H}_0}\left(\left\{\hat{Q}^i \leq \frac{\alpha}{(1+\epsilon)K}\right\}\middle|\mathcal{T}_{\mathrm{cal}}\right) && (80)
\end{aligned}$$

Let $r^i = \mathrm{P}_{\mathrm{H}_0}\left(\left\{\hat{Q}^i \leq \frac{\alpha}{(1+\epsilon)K}\right\}\middle|\mathcal{T}_{\mathrm{cal}}\right)$. Thus,

$$\mathrm{P_F}(\mathcal{T}_{\mathrm{cal}}) = \sum_{i=1}^{K} r^i. \tag{81}$$

We have from Vovk (2012); Bates et al. (2023) that r^i follows a Beta distribution, i.e., $r^i \sim \mathrm{Beta}(a, b)$, where

$$\begin{aligned}
a &= \left\lfloor (n_{\mathrm{cal}}+1)\frac{\alpha}{(1+\epsilon)K}\right\rfloor && (82)\\
b &= (n_{\mathrm{cal}}+1) - a. && (83)
\end{aligned}$$

The mean of this distribution is $\mu = \frac{a}{a+b}$. Let E denote the event

$$E = \bigcap_{i=1}^{K}\left\{r^i \leq (1+\epsilon)\frac{\alpha}{(1+\epsilon)K}\right\}. \tag{84}$$

When n_{cal} satisfies the condition in (75), we have that

$$\begin{aligned}
1 - \mathrm{P}_{\mathrm{H}_0}(E) &= 1 - \mathrm{P}_{\mathrm{H}_0}\left(\bigcap_{i=1}^{K}\left\{r^i \leq (1+\epsilon)\frac{\alpha}{(1+\epsilon)K}\right\}\right) && (85)\\
&\leq \sum_{i=1}^{K}\mathrm{P}_{\mathrm{H}_0}\left\{r^i \geq (1+\epsilon)\frac{\alpha}{(1+\epsilon)K}\right\} && (86)\\
&= \sum_{i=1}^{K} 1 - I_{\frac{\alpha}{K}}(a, b) && (87)\\
&\leq \sum_{i=1}^{K} 1 - I_{(1+\epsilon)\mu}(a, b) \leq \sum_{i=1}^{K}\frac{\delta}{K} \leq \delta. && (88)
\end{aligned}$$

Thus, under event E, i.e., with probability greater than $1-\delta$, we have that

$$\begin{aligned} \mathrm{P_F}(\mathcal{T}_{\mathrm{cal}}) &= \sum_{i=1}^{K} r^i \\ &\leq \sum_{i=1}^{K} (1+\epsilon)\frac{\alpha}{(1+\epsilon)K} \\ &= \alpha. \end{aligned} \tag{89}$$

■

Appendix E. Additional Experimental Results

All experiments presented in this paper were run on a single NVIDIA GTX-1080Ti GPU with PyTorch.

In addition, we provide the detection probabilities for CIFAR100 as the in-distribution dataset in Table 5. We consider the Mahalanobis scores and Gram scores from the individual layers for the same. Recall that the energy score is a temperature scaled log-sum-exponent of the softmax scores, i.e., $-T\log\sum_{i=1}^{c} e^{\sigma_i(x)/T}$ where c is the number of classes, σ_i are the softmax scores, and T is the temperature parameter. We do not consider the energy score as one of the statistics for CIFAR100 as in-distribution, as we do not expect it to give a good representation of the in-distribution data. As the number of classes in CIFAR100 is quite large (100), we expect the softmax scores to not provide a reliable confidence score for distinguishing in-distribution points from OOD samples. Table 6 contains the AUROC numbers for CIFAR100 as the in-distribution dataset.

Table 5: Comparison with baseline OOD detection techniques for CIFAR100 as in-distribution dataset. Each entry is $P_D(\%)$ **at** $P_F = 10\%$ for the corresponding detection method, OOD dataset and DNN architecture.

OOD Dataset	**Method**	**ResNet34**	**DenseNet**
SVHN	Mahala (penultimate layer)	61.75	62.21
	Gram	71.60	77.87
	Ours - Mahala (5/4)	64.55	62.81
	Ours - Gram (5/4)	58.54	78.15
	Ours - Mahala, Gram (all) (5/4 + 1)	72.81	70.80
ImageNet	Mahala (penultimate layer)	35.03	89.05
	Gram	82.42	86.42
	Ours - Mahala (5/4)	86.04	90.72
	Ours - Gram (5/4)	74.43	86.85
	Ours - Mahala, Gram (all) (5/4 + 1)	85.64	90.15
LSUN	Mahala (penultimate layer)	34.00	92.17
	Gram	78.36	88.93
	Ours - Mahala (5/4)	86.19	92.86
	Ours - Gram (5/4)	66.62	89.20
	Ours - Mahala, Gram (all) (5/4 + 1)	84.81	92.66
iSUN	Mahala (penultimate layer)	36.01	88.89
	Gram	83.15	84.82
	Ours - Mahala (5/4)	99.35	99.82
	Ours - Gram (5/4)	53.71	83.01
	Ours - Mahala, Gram (all) (5/4 + 1)	99.42	99.85

Table 6: Comparison with baseline OOD detection techniques for CIFAR100 as in-distribution dataset. Each entry is AUROC for the corresponding detection method, OOD dataset and DNN architecture.

OOD Dataset	**Method**	**ResNet34**	**DenseNet**
SVHN	Mahala (penultimate layer)	89.35	85.81
	Gram	91.85	91.33
	Ours - Mahala (5/4)	89.42	86.59
	Ours - Gram (5/4)	88.86	91.23
	Ours - Mahala, Gram (all) (5/4 + 1)	91.53	89.98
ImageNet	Mahala (penultimate layer)	78.81	95.38
	Gram	94.10	94.13
	Ours - Mahala (5/4)	94.96	95.65
	Ours - Gram (5/4)	92.00	94.04
	Ours - Mahala, Gram (all) (5/4 + 1)	94.96	95.66
LSUN	Mahala (penultimate layer)	78.90	96.39
	Gram	93.06	95.33
	Ours - Mahala (5/4)	94.91	96.13
	Ours - Gram (5/4)	90.00	95.19
	Ours - Mahala, Gram (all) (5/4 + 1)	94.73	96.18
iSUN	Mahala (penultimate layer)	81.38	95.43
	Gram	94.71	94.36
	Ours - Mahala (5/4)	98.12	98.04
	Ours - Gram (5/4)	89.77	93.85
	Ours - Mahala, Gram (all) (5/4 + 1)	98.04	97.90

References

Anastasios N Angelopoulos, Stephen Bates, Emmanuel J Candès, Michael I Jordan, and Lihua Lei. Learn then test: Calibrating predictive algorithms to achieve risk control. *arXiv preprint arXiv:2110.01052*, 2021.

Vineeth Balasubramanian, Shen-Shyang Ho, and Vladimir Vovk. *Conformal Prediction for Reliable Machine Learning: Theory, Adaptations and Applications*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1st edition, 2014. ISBN 0123985374.

Stephen Bates, Emmanuel Candès, Lihua Lei, Yaniv Romano, and Matteo Sesia. Testing for outliers with conformal p-values. *The Annals of Statistics*, 51(1):149 – 178, 2023.

Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, 57(1):289–300, 1995. ISSN 00359246.

Yoav Benjamini and Daniel Yekutieli. The control of the false discovery rate in multiple testing under dependency. *The Annals of Statistics*, 29(4):1165–1188, 2001. ISSN 00905364.

Federico Bergamin, Pierre-Alexandre Mattei, Jakob Drachmann Havtorn, Hugo Senetaire, Hugo Schmutz, Lars Maaløe, Soren Hauberg, and Jes Frellsen. Model-agnostic out-of-distribution detection using combined statistical tests. In *International Conference on Artificial Intelligence and Statistics*, pages 10753–10776. PMLR, 2022.

Ronald A. Fisher. *Statistical Methods for Research Workers*, pages 66–70. Springer New York, New York, NY, 1992. ISBN 978-1-4612-4380-9.

Matan Haroush, Tzviel Frostig, Ruth Heller, and Daniel Soudry. A statistical framework for efficient out of distribution detection in deep neural networks. *arXiv preprint arXiv:2102.12967*, 2021.

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.

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, Mantas Mazeika, and Thomas Dietterich. Deep anomaly detection with outlier exposure. In *International Conference on Learning Representations*, 2019.

Sture Holm. A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 6(2):65–70, 1979. ISSN 03036898, 14679469.

Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4700–4708, 2017.

Rui Huang, Andrew Geng, and Yixuan Li. On the importance of gradients for detecting distributional shifts in the wild. In *Advances in Neural Information Processing Systems*, volume 34, pages 677–689. Curran Associates, Inc., 2021.

Ramneet Kaur, Susmit Jha, Anirban Roy, Sangdon Park, Edgar Dobriban, Oleg Sokolsky, and Insup Lee. iDECODe: In-distribution equivariance for conformal out-of-distribution detection. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(7):7104–7114, Jun. 2022.

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*, volume 31. Curran Associates, Inc., 2018.

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

Ziyi Liang, Matteo Sesia, and Wenguang Sun. Integrative conformal p-values for powerful out-of-distribution testing with labeled outliers. *arXiv preprint arXiv:2208.11111*, 2022.

Weitang Liu, Xiaoyun Wang, John Owens, and Yixuan Li. Energy-based out-of-distribution detection. In *Advances in Neural Information Processing Systems*, volume 33, pages 21464–21475. Curran Associates, Inc., 2020.

Anh Nguyen, Jason Yosinski, and Jeff Clune. Deep neural networks are easily fooled: High confidence predictions for unrecognizable images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 427–436, 2015.

Aleksandr Podkopaev and Aaditya Ramdas. Tracking the risk of a deployed model and detecting harmful distribution shifts. *arXiv preprint arXiv:2110.06177*, 2021.

Chandramouli S. Sastry and Sageev Oore. Detecting out-of-distribution examples with Gram matrices. In *Proceedings of the 37th International Conference on Machine Learning*, ICML'20. JMLR.org, 2020.

R John Simes. An improved Bonferroni procedure for multiple tests of significance. *Biometrika*, 73(3):751–754, 1986.

Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*, 2013.

Vladimir Vovk. Conditional validity of inductive conformal predictors. In *Proceedings of the Asian Conference on Machine Learning*, volume 25 of *Proceedings of Machine Learning Research*, pages 475–490. PMLR, 2012.

Volodya Vovk, Alexander Gammerman, and Craig Saunders. Machine-learning applications of algorithmic randomness. In *Proceedings of the Sixteenth International Conference on Machine Learning*, ICML '99, page 444–453, San Francisco, CA, USA, 1999. Morgan Kaufmann Publishers Inc. ISBN 1558606122.

Lily Zhang, Mark Goldstein, and Rajesh Ranganath. Understanding failures in out-of-distribution detection with deep generative models. In *International Conference on Machine Learning*, pages 12427–12436. PMLR, 2021.