# UNDERSTANDING BIAS IN ANOMALY DETECTION: A SEMI-SUPERVISED VIEW WITH PAC GUARANTEES

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

Anomaly detection presents a unique challenge in machine learning, due to the scarcity of labeled anomaly data. Existing work attempts to mitigate such problems via semi-supervised learning, i.e., augmenting unsupervised anomaly detection models with additional labeled anomaly samples. However, the labeled data often does not align with the target distribution and therefore introduces harmful bias to the trained model. In this paper, we aim to understand the effect of a biased anomaly set for anomaly detection. In particular, we formally state the anomaly detection problem as a semi-supervised learning task. We focus on the anomaly detector's recall at a given false positive rate as the main performance metric. Given two different anomaly score functions, we formally define their difference in performance as the relative *scoring bias* of the anomaly detectors. We establish the first finite sample rates for estimating the relative scoring bias for semi-supervised anomaly detection. We then empirically validate our theoretical results on both synthetic and real-world datasets. Furthermore, we provide extensive empirical study on how a biased training anomaly set affects the anomaly score function and therefore the resulting detection performance. Our case study demonstrates scenarios in which the biased anomaly set can be useful, and provides a solid benchmark for future research.

## 1 INTRODUCTION

Anomaly detection (Pimentel et al. (2014)) trains a formal model to identify unexpected or anomalous instances in incoming data, whose behaviors differ from normal instances. It is particularly useful for detecting problematic events such as digital fraud, structural defects, and system malfunctions. Building accurate anomaly detection models is a well-known challenge in machine learning, due to the scarcity of labeled anomaly data. The classical and most common approach is to train anomaly detection models using unsupervised methods, i.e., first train a model using a corpus of normal data to capture *normal* behaviors, then configure the model to flag instances with large deviations as anomalies. Researchers have also developed deep learning methods to better capture the complex structure in the data (Ruff et al. (2018); Wang et al. (2019a); Zhou & Paffenroth (2017)).

Recently, a new line of anomaly detection models attempts to mitigate the shortage of labeled anomalies via semi-supervised learning, which augments unsupervised anomaly detection models with additional labeled anomaly samples as they become available (Ruff et al. (2020b); Yamanaka et al. (2019); Ruff et al. (2020a); Hendrycks et al. (2019a)). Existing works show that these new models achieve considerable performance improvements beyond their unsupervised versions.

When exploring semi-supervised models, we found that when the labeled anomalies (used to train the model) do not align with the target distribution, they could introduce harmful bias to the trained model. Specifically, when comparing the performance of a semi-supervised anomaly detector to its unsupervised baseline, the performance difference varies significantly across test anomaly data, some better and some worse. That is, semi-supervised learning can introduce large variance (or bias) in anomaly detection outcomes.

In this paper, we aim to understand the effect of a biased training set on semi-supervised anomaly detection models. We formally state the anomaly detection problem as a semi-supervised binary classification task, and focus on the anomaly detector's recall at a given false positive rate as the main performance metric. We factor the contribution of the labeled anomalies by the detector's anomaly

scoring function, and show that different types of labeled anomalies produce different anomaly scoring functions. Next, given two different anomaly scoring functions, we formally define their difference in performance as the relative *scoring bias* of the anomaly detectors. Our novel notion of scoring bias for anomaly detection aligns with the notion of bias in the classical supervised learning setting, with the key difference being the different performance metric (i.e., recall at a given false positive rate). Along this line, we establish the first finite sample rates for estimating the relative scoring bias for semi-supervised anomaly detection. We empirically validate our assumptions and theoretical results on both synthetic and real-world datasets.

Furthermore, we provide an empirical study on how a biased training anomaly set affects the anomaly score function and therefore the resulting detection performance. We consider three real-world datasets (Fashion-MNIST, Statlog (Landsat Satellite), and ImageNet) and six deep-learning based anomaly detection models. Our case study demonstrates scenarios in which the biased anomaly set can be useful, and provides a solid benchmark for future research.

In this paper, we introduce a formal analysis on the effect of a biased training set on semi-supervised anomaly detection. Our main contributions are the following:

- We discover the issue of large performance variance in semi-supervised anomaly detectors, caused by the use of the biased anomaly set as training data.
- We model the effect of biased training as relative scoring bias, and establish the first finite sample rates for estimating the relative scoring bias of the trained models.
- We conduct empirical experiments to verify and characterize the impact of the relative scoring bias on six popular anomaly detection models, and three real-world datasets.

## 2 Related Work

**Anomaly Detection Models.** While the literature on anomaly detection models is extensive, the most relevant to our work are deep learning based unsupervised and semi-supervised anomaly detection models. Here we define unsupervised models as those trained only on normal data (Ruff et al. (2018); Sakurada & Yairi (2014); Zhou & Paffenroth (2017)), and semi-supervised models as those trained on both normal data and a small set of labeled anomalies (Pang et al. (2019); Daniel et al. (2019); Yamanaka et al. (2019); Ruff et al. (2020b;a)). Under this context, we categorize existing deep anomaly detection models by two major types: hypersphere based (Ruff et al. (2018; 2020a;b)) and autoencoder (or reconstruction) based models (Zhou & Paffenroth (2017); Yamanaka et al. (2019)). Another line of recent work proposes to use synthetic or auxiliary anomalies to train anomaly detection models (Golan & El-Yaniv (2018); Hendrycks et al. (2019c); Lee et al. (2018); Hendrycks et al. (2019b)), "forcing" the model to learn a more compact representation of the normal data. While the existing work has shown empirically that the choice of abnormal data in training can help detect some unseen abnormal distributions, it does not offer any theoretical explanation for the phenomenon, nor does it consider the countercases when additional abnormal data in training hurt the detection performance.

**Bias in Anomaly Detection.** To the best of our knowledge, we are the first to identify the presence of bias caused by additional labeled anomaly set in semi-supervised anomaly detection models, especially when there exist a mismatch between the anomalies present in training and the anomalies encountered in testing (as shown in Section 5). Existing works have explored the presence of bias in unsupervised anomaly detection models when there exists defective normal data in training, like outliers and simple-to-reconstruct examples (Tong et al., 2019), or examples with background noise (Liu & Ma, 2019). There is also literature on the bias-variance tradeoff for ensembles of unsupervised anomaly detection models (Aggarwal & Sathe, 2015; Rayana et al., 2016). But little work has been done on the bias of anomaly detection in the semi-supervised setting. Another line of work in transfer learning has identified the value of additional labeled data in training (Kpotufe & Martinet, 2018; Hanneke & Kpotufe, 2019) and the performance bias on target data by transferring knowledge from a less related source (Wang et al., 2019b; Wu et al., 2020). Yet most work only considered the cases of classifier models.

**PAC guarantees for Anomaly Detection.** Despite significant progress on developing theoretical guarantees for classification tasks (Valiant (1984); Kearns et al. (1994)), little has been done for anomaly detection tasks. Siddiqui et al. (2016) first establishes a PAC framework for anomaly detection models using the notion of pattern space; however, it is hard to apply such pattern spaces to deep learning models with complex latent spaces. Liu et al. (2018) proposes a model-agnostic

approach to provide the PAC guarantee for anomaly detection performance, by analyzing the convergence for the cumulative distribution of anomaly scores. We follow the basic setting from this line of work to address the convergence of the relative scoring bias. In contrast to prior work, our proof relies on a novel adaption of the key theoretical tool from Massart (1990), which allows us to extend our theory to characterize the notion of scoring bias as defined in Section 3.2.

# 3 Problem Formulation

We formulate the anomaly detection problem in a (semi-) supervised binary classification setting. Consider a model class $\Theta$ for anomaly detection, and a (labeled) training set $D$ sampled from a mixture distribution $\mathcal{D}$ over the normal and anomalous instances. In the context of anomaly detection, a model $\theta$ maps each input instance $x$ to a continuous output, which corresponds to anomaly score $s_\theta(x)$. The model further uses a threshold $\tau_\theta$ on the score function to produce a binary label for input $x$.

For a given threshold value $\tau_\theta$, we can define the False Positive Rate (FPR) of the model $\theta$ on the input data distribution as $\text{FPR}(s_\theta, \tau_\theta) = \mathbb{P}\left[s_\theta(x) > \tau_\theta \mid y = 0\right]$, and the True Positive Rate (TPR, a.k.a. Recall) as $\text{TPR}(s_\theta, \tau_\theta) = \mathbb{P}\left[s_\theta(x) > \tau_\theta \mid y = 1\right]$. The FPR and TPR are competing objectives—therefore, a key challenge for anomaly detection algorithms is to identify a configuration of the score, threshold pair $(s_\theta, \tau_\theta)$ that strikes a balance between the two performance metrics. W.l.o.g.[1], in this paper we focus on the following scenario, where the objective is to maximize TPR subject to achieving a target FPR. Formally, let $q$ be the target FPR; we define the optimal anomaly detector as

$$(s_\theta^*, \tau_\theta^*) \in \underset{(s_\theta, \tau_\theta):\theta\in\Theta}{\arg\max}\ \text{TPR}(s_\theta, \tau_\theta) \quad \text{s.t. } \text{FPR}(s_\theta, \tau_\theta) \leq q \tag{3.1}$$

## 3.1 A General Anomaly Detection Framework

Note that the performance metric (namely TPR) in Problem 3.1 is statistics that depends on the entire predictive distribution, and can not be easily evaluated on any single data point. Therefore, rather than directly solving Problem 3.1, practical anomaly detection algorithms (such as OCSVM (Schölkopf et al., 1999), Deep SAD (Ruff et al., 2020b), etc) often rely on a two-stage process: (1) learning the score function $s_\theta$ from training data via a surrogate loss, and (2) given $s_\theta$ from the previous step, computing the threshold function $\tau_\theta$ on the training data. Formally, given a model class $\Theta$, a training set $D$, a loss function $\ell$, and a target FPR $q$, a two-staged anomaly detection algorithm outputs

$$\begin{cases} \hat{s}_\theta \in \arg\min_{s_\theta:\theta\in\Theta} \ell(s_\theta, D) \\ \hat{\tau}_\theta \in \arg\max_{\tau_\theta:\theta\in\Theta} \text{TPR}(\hat{s}_\theta, \tau_\theta) \quad \text{s.t. } \text{FPR}(\hat{s}_\theta, \tau_\theta) \leq q \end{cases} \tag{3.2}$$

Here, the loss function $\ell$ could be instantiated into latent-space-based losses (e.g., Deep SAD), margin-based losses (e.g., OCSVM), or reconstruction-based losses (e.g., ABC (Yamanaka et al., 2019)); therefore, many contemporary anomaly detection models fall into this framework. To set the threshold $\hat{\tau}_\theta$, we consider using the distribution of the anomaly scores $\hat{s}_\theta(\cdot)$ from a labeled validation set $D^{\text{val}} \sim \mathcal{D}$. Let $D^{\text{val}} := D_0^{\text{val}} \cup D_a^{\text{val}}$ where $D_0^{\text{val}}$ and $D_a^{\text{val}}$ denote the subset of normal data and the subset of abnormal data of $D^{\text{val}}$. Denote the empirical CDFs for anomaly scores assigned to $x$ in $D_0^{\text{val}}$ and $D_a^{\text{val}}$ as $\hat{F}_0$ and $\hat{F}_a$, respectively. Then, given a target FPR value $q$, following a similar argument as Liu et al. (2018), one can compute the threshold as $\hat{\tau}_\theta = \max\{u \in \mathbb{R} : \hat{F}_0(u) \leq q\}$. The steps for solving the second part of Problem 3.2 is summarized in Algorithm 1.

## 3.2 Scoring Bias

Given a model class $\Theta$ and a training set $D$, we define the *scoring bias* of a detector $(\hat{s}_\theta, \hat{\tau}_\theta)$ to be the difference in TPR between $(\hat{s}_\theta, \hat{\tau}_\theta)$ and $(s_\theta^*, \tau_\theta^*)$:

$$\text{bias}(\hat{s}_\theta, \hat{\tau}_\theta) := \underset{(s_\theta, \tau_\theta):\theta\in\Theta}{\arg\max}\ \text{TPR}(s_\theta, \tau_\theta) - \text{TPR}(\hat{s}_\theta, \hat{\tau}_\theta) \tag{3.3}$$

[1] Our results can be easily extended to the setting where the goal is to minimize FPR subject to a given TPR.

**Algorithm 1:** Computing the anomaly detection threshold for Problem 3.2

**Data:** A validation dataset $D^{\text{val}}$ and a scoring function $s(\cdot)$.
**Result:** A score threshold achieving a target FPR and the corresponding recall on $D^{\text{val}}$.

1 Get anomaly score $s(x)$ for each x in $D^{\text{val}}$.
2 Compute empirical CDF $\hat{F}_0(x)$ and $\hat{F}_a(x)$ for anomaly scores of $x$ in $D_0^{\text{val}}$ and $D_a^{\text{val}}$.
3 Output detection threshold $\hat{\tau} = \max\{u \in \mathbb{R} : \hat{F}_0(u) \leq q\}$.
4 Output FPR (recall) on $D_a^{\text{val}}$ as $\hat{r} = 1 - \hat{F}_a(\hat{\tau})$.

We call $(\hat{s}_\theta, \hat{\tau}_\theta)$ a *biased* detector if $\text{bias}(\hat{s}_\theta, \hat{\tau}_\theta) > 0$. In practice, due to biased training distribution, and the fact that the two-stage process in Problem 3.2 is not directly optimizing TPR, the resulting anomaly detectors are often biased by construction. Therefore, one practically significant performance measure is the *relative bias*, defined as the difference in TPR between two anomaly detectors, subject to the constraints in Problem 3.2. It captures the relative strength of two algorithms in detecting anomalies, and therefore is an important indicator for model evaluation and model selection. Formally, given two anomaly score functions $s, s'$ and the corresponding threshold function $\tau, \tau'$ obtained from Algorithm 1, we define the *relative scoring bias* between $s$ and $s'$ as:

$$\xi(s, s') := \text{bias}(s, \tau) - \text{bias}(s', \tau') = \text{TPR}(s', \tau') - \text{TPR}(s, \tau) \tag{3.4}$$

Note that when $s' = s_\theta^*$, the relative scoring bias (equation 3.4) reduces to the scoring bias (equation 3.3). We further define the *empirical relative scoring bias* between $s$ and $s'$ as

$$\hat{\xi}(s, s') := \widehat{\text{TPR}}(s', \tau') - \widehat{\text{TPR}}(s, \tau) \tag{3.5}$$

where $\widehat{\text{TPR}}(s, \tau) = \frac{1}{n}\sum_{j=1}^{n} \mathbf{1}_{s(x_j) \leq \tau; y_j = 1}$ denotes the TPR/recall estimated on a finite validation set of size $n$. In the following sections, we will investigate both the theoretical properties and the empirical behavior of the empirical relative scoring bias for contemporary anomaly detectors.

## 4 Finite Sample Analysis for Empirical Relative Scoring Bias

In this section, we show how one can estimate the relative scoring bias (equation 3.4) given two scoring functions $s, s'$ learned in Section 3.1. As an example, $s$ could be a scoring function learned via an unsupervised anomaly detector, and $s'$ could be a scoring function learned via a semi-supervised anomaly detector with a biased anomaly set. In the following, we provide a finite sample analysis of the convergence rate of the empirical relative scoring bias, and validate our theoretical analysis via a simulated case study.

### 4.1 Finite Sample Guarantee

**Notations.** Concretely, we assume that when determining the threshold in Line 3 of Algorithm 1, both scoring functions $s, s'$ are evaluated on the unbiased (marginal) empirical distribution of the normal data. Furthermore, the empirical TPR in Line 4 are estimated on the unbiased empirical distribution of the abnormal data. Let $\{s_i := s(x_i) \mid x_i, y_i = 0\}_{i=1}^{n_0}$ denote a set of anomaly scores evaluated by $s(\cdot)$ on $n_0$ *i.i.d.*random normal data points. Following the notation in Section 3.1, we use $F_0(t) := \mathbb{P}[s(x) < t \mid y = 0]$ to denote the CDF of $s(x)$, and use $\hat{F}_0(t) := \frac{1}{n_0}\sum_{i=1}^{n_0} \mathbf{1}_{s_i \leq t; y_i = 0}$ to denote the corresponding empirical CDF. For $n_1$ *i.i.d.*samples $\{s_j := s(x_j) \mid x_j, y_j = 1\}_{j=0}^{m}$ with CDF $F_a(t) := \mathbb{P}[s(x) < t \mid y = 1]$, the corresponding emprical CDF is $\hat{F}_a(t) := \frac{1}{n_1}\sum_{j=1}^{n_1} \mathbf{1}_{s_j \leq t; y_j = 1}$. Similarly, we denote the CDF and emiprical CDF for $\{s'_i \mid y_i = 0\}_{i=0}^{n_0}$ as $F'_0(t)$ and $\hat{F}'_0(t)$, and for $\{s'_j \mid y_j = 1\}_{j=0}^{n_1}$ as as $F'_a(t)$ and $\hat{F}'_a(t)$, respectively.

**Infinite sample case.** In the limit of infinite data (both normal and abnormal), $\hat{F}_0, \hat{F}_a, \hat{F}'_0, \hat{F}'_a$ will converge to the true CDFs (Parzen, 1980), and hence the empirical relative scoring bias will also converge. The following Proposition establishes a connection between the CDFs and the scoring bias.

**Proposition 1.** *Given two scoring functions $s, s'$ and a target FPR $q$, the relative scoring bias* $\xi(s, s') = F_a(F_0^{-1}(q)) - F'_a(F_0'^{-1}(q))$.

Here, $F^{-1}(\cdot)$ is the quantile function. The proof of Proposition 1 follows from the fact that for corresponding choice of $\tau, \tau'$ in Algorithm 1, TPR$(s, \tau) = 1 - F_a(F_0^{-1}(q))$, and TPR$s', \tau' = 1 - F'_a(F_0'^{-1}(q))$.

Next, a direct corollary of the above result shows that, for the special cases where both the scores for normal and abnormal data are Gaussian distributed, one can directly compute the relative scoring bias. The proof is listed in Appendix A.

**Corollary 2.** *Fixed a target FPR $q$. Given two scoring functions $s, s'$, assume that $s(x) \mid y = 0 \sim \mathcal{N}(\mu_0, \sigma_0)$, $s(x) \mid y = 1 \sim \mathcal{N}(\mu_a, \sigma_a)$, $s'(x) \mid y = 0 \sim \mathcal{N}(\mu'_0, \sigma'_0)$, $s'(x) \mid y = 1 \sim \mathcal{N}(\mu'_a, \sigma'_a)$. The relative scoring bias*

$$\xi(s, s') = \Phi\left(\frac{\sigma_0 \Phi^{-1}(q)}{\sigma_a} + \frac{\mu_0 - \mu_a}{\sigma_a}\right) - \Phi\left(\frac{\sigma'_0 \Phi^{-1}(q)}{\sigma'_a} + \frac{\mu'_0 - \mu'_a}{\sigma'_a}\right)$$

**Finite sample case.** In practical scenarios, when comparing the performance of two scoring functions $s$ and $s'$, we would only have access to finite samples from the validation set, and hence it is crucial to bound the estimation error due to insufficient samples. We now establish a finite sample guarantee for the estimating the relative scoring bias. Our result extends the analysis of Liu et al. (2018), where we follow the convention to assume that there is the anomaly data amounts to an $\alpha$ fraction of the mixture distribution. The validation set contains a mixture of $n = n_0 + n_1$ i.i.d. samples, with $n_0$ normal samples and $n_1$ abnormal samples where $\frac{n_1}{n} = \alpha$.

The following result shows that under mild assumptions of the continuity of the CDFs and quantile functions $F_a, F'_a, F_0^{-1}, F_0'^{-1}$, the sample complexity for achieving $|\hat{\xi} - \xi| \leq \epsilon$:

**Theorem 3.** *Assume that $F_a, F'_a, F_0^-, F_0'^-$ are Lipschitz continuous with Lipschitz constant $\ell_a, \ell'_a, \ell_0^-, \ell_0'^-$, respectively. Let $\alpha$ be the fraction of abnormal data among $n$ i.i.d.samples from the mixture distribution. Then, w.p. at least $1 - \delta$, with*

$$n \geq \frac{8}{\epsilon^2} \cdot \left(\log \frac{2}{1 - \sqrt{1 - \delta}} \cdot \left(\frac{2 - \alpha}{\alpha}\right)^2 + \log \frac{2}{\delta} \cdot \frac{1}{1 - \alpha}\left(\left(\frac{\ell_a}{\ell_0^-}\right)^2 + \left(\frac{\ell'_a}{\ell_0'^-}\right)^2\right)\right)$$

*the empirical relative scoring bias satisfies $|\hat{\xi} - \xi| \leq \epsilon$.*

We defer the proof of Theorem 3 to Appendix B. Similarly with the open category alien detection setting as discussed in Liu et al. (2018), the sample complexity for estimating the relative scoring bias $n$ grows as $\mathcal{O}\left(\frac{1}{\alpha^2 \epsilon^2} \log \frac{1}{\delta}\right)$. Note the analysis of our bound involves a novel two-step process which first bounds the estimation of the threshold for the given FPR, and then leverages the Lipschitz continuity condition to derive the final bound.

### 4.2 Simulated Case Study

We conduct a simulated case study to validate our main results in this section, by training anomaly detection models using a synthetic dataset (Liu et al., 2018). Our case study considers six anomaly detection models listed in Table 1, and they lead to similar results. For brevity, we show results when using Deep SVDD (Ruff et al., 2018) as the unsupervised model and Deep SAD (Ruff et al., 2020b) as the semi-supervised model. Later in Appendix D, we include results of other model pairs, including Deep SVDD vs. Hypersphere Classifier (HSC) (Ruff et al., 2020a), Autoencoder (AE) vs. Semi-supervised Autoencoder (SAE)[2], and AE vs. ABC (Yamanaka et al., 2019).

Similar to Liu et al. (2018), we generate our synthetic dataset by sampling data from a mixture data distribution $S$, w.p. $1 - \alpha$ generating the normal data distribution $S_0$ and w.p. $\alpha$ generating the abnormal data distribution $S_a$. Data points in $S_0$ are sampled randomly from a 9-dimensional Gaussian distribution, where each dimension is independently distributed as $\mathcal{N}(0, 1)$. Data points

[2]We design SAE by forcing the reconstruction errors to be maximized for additional labeled anomalies encountered in training the autoencoder.

| Type | Unsupervised) | Semi-supervised) |
|---|---|---|
| Hypersphere-based | Deep SVDD | Deep SAD, Hypersphere Classifier (HSC) |
| Reconstruction-based | Autoencoder (AE) | Semi-supervised AE (SAE), Autoencoding Binary Classifier (ABC) |

Table 1: The anomaly detection models considered in our case study. Deep SAD and HSC are the semi-supervised version of Deep SVDD; SAE and ABC are the semi-supervised version of AE.

in $S_a$ are sampled from another 9-dimensional distribution, which w.p. 0.4 have 3 dimensions (uniformly chosen at random) distributed as $\mathcal{N}(1.6, 0.8)$, w.p. 0.6 have 4 dimensions (uniformly chosen at random) distributed as $\mathcal{N}(1.6, 0.8)$, and have the remaining dimensions distributed as $\mathcal{N}(0, 1)$. This ensures meaningful feature relevance, point difficulty and variation for the abnormal data distribution as discussed in Emmott et al. (2015).

We obtain two score functions $s$ and $s'$ by training Deep SVDD and Deep SAD respectively on samples from the synthetic dataset (10K data from $S_0$, 1K data from $S_a$). We configure the training, validation and test set so there is no data overlap in them. Thus the training procedure will not affect the sample complexity for estimating the relative scoring bias. To set the anomaly threshold, we fix the target FPR to be 0.05, and vary the number of normal data in the validation set $n$ from $\{100, 1\text{K}, 10\text{K}\}$. We then test the score function and threshold on a fixed test dataset with a large number (20K) of normal data and $\alpha \times 20\text{K}$ of abnormal data. We vary $\alpha$ from $\{0.01, 0.05, 0.1, 0.2\}$.

**Distribution of anomaly scores.** We first study the distribution of anomaly scores. Figure 1 is a sample plot of score distributions on the test set with $\alpha = 0.1$. We plot the scores for normal and abnormal test data separately, for both scoring functions (derived from Deep SVDD and Deep SAD models respectively). We make two key observations. First, all the distribution curves follow a rough bell shape. Second and more importantly, while the abnormal score distribution closely mimics the normal score distribution under the unsupervised model, it deviates largely from the normal score distribution after semi-supervised learning (i.e., similar mean but much higher variance). This confirms that semi-supervised learning does introduce additional bias in anomaly scores.

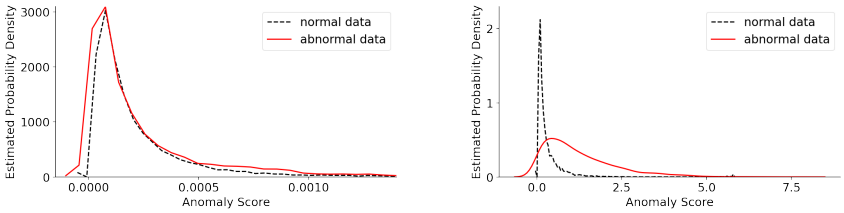


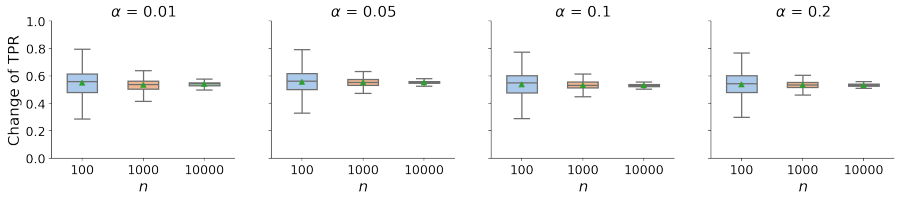

Figure 1: Anomaly score distributions for Deep SVDD (left) and Deep SAD (right) trained and tested using the synthetic dataset, estimated by Kernel Density Estimation.

We also examine the anomaly score distributions for models trained using a real-world dataset (Fashion-MNIST). While the score distributions are less close to Gaussian, we do observe the same trend where normal and abnormal score distributions become significantly different after applying semi-supervised learning. The result is shown in Figure 7 in Appendix D.

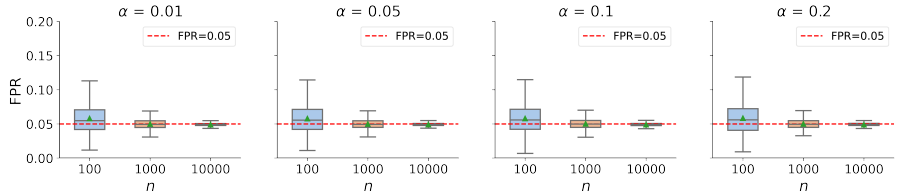
**Convergence of relative scoring bias ($\xi$) and FPR.** Next we examine the convergence of FPR obtained from $s'$ (semi-supervised model) and the relative scoring bias $\xi$ (computed as the difference of TPR) obtained from $s$ (unsupervised model) and $s'$ (semi-supervised model).

Similar to our theory results, we observe a consistent trend of convergence in FPR and $\xi$ as the sample complexity goes up. More specifically, for FPR, as $n$ goes up, it converges to the prefixed value of $0.05$; for $\xi$, it also converges to a certain level. These can also be observed from the two examples in Figure 2, in terms of the quantile distribution of $\xi$ between Deep SVDD (unsupervised) and Deep SAD (semi-supervised), and the quantile distribution of Deep SAD's FPR. Results for other models are in Appendix D, and show consistent trends.

We also examine the rate of convergence w.r.t to $n$. Earlier in Section 4.1 we show that $n$ required for estimating $\xi$ grows in the same order as $\frac{1}{\alpha^2\epsilon^2}\log\frac{1}{\delta}$. That is, the estimation error $\epsilon$ decreases at the rate of $\frac{1}{\sqrt{n}}$; furthermore, as $\alpha$ increases, the value of $n$ required for estimating $\xi$ decreases. This can be seen from Figure 2(a) where at $n = 10000$, the variation of $\xi$ at $\alpha = 0.2$ is 50% less than that at $\alpha = 0.01$.




(a) Relative Scoring Bias ($\xi$) between Deep SVDD and Deep SAD




(b) FPR of Deep SAD

Figure 2: The quantile distribution of (a) relative scoring bias $\xi$ and (b) FPR, computed on the test set over 1500 runs. We vary $n$ =100, 1000 or 10000, and $\alpha$ =0.01, 0.05, 0.1, or 0.5. The triangle in each boxplot marks the mean value. For (b) FPR, the red dotted line marks the target FPR of 0.05.

## 5 Impact of Scoring Bias on Anomaly Detection Performance

We perform empirical experiments to study the end-to-end impact of relative scoring bias on semi-supervised anomaly detection models. Our goal is to understand the type and severity of performance variations introduced by different anomaly training sets.

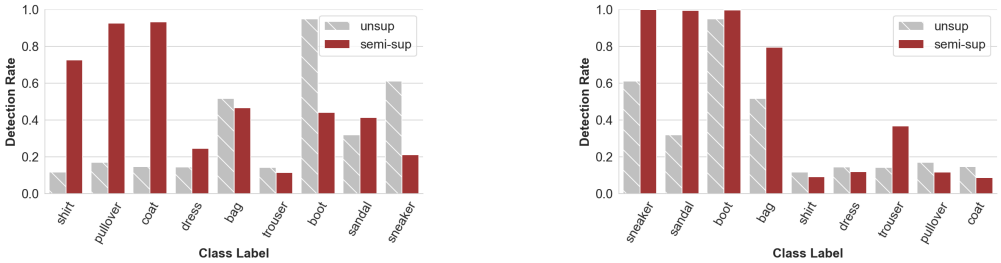
**Experiment setup.** We consider six anomaly detection models previously listed in Table 1, and three real-world datasets (Fashion-MNIST, Statlog (Landsat Satellite) and ImageNet). For each dataset, we build normal data by choosing a single class (e.g, class `top` in Fashion-MNIST), and treat the other classes as the abnormal classes. From those abnormal class, we pick a single class as the abnormal training data, and the rest as the abnormal test data on which we test separately. Given the data, we train $\theta_0 := (s_{\theta_0}, \tau_{\theta_0})$, a unsupervised anomaly detector using normal training data, and $\theta_s := (s_{\theta_s}, \tau_{\theta_s})$ a semi-supervised anomaly detector using both normal training data and abnormal training data (with a 10:1 normal vs. anomaly ratio). We follow the original paper of each model to implement the model and its training. For each trained model, we configure the anomaly score threshold to reach a target false positive rate (FPR) of 0.05. We then test these trained models against various abnormal test classes, and record the recall (TPR) value per class. We repeat the above by selecting different abnormal training data. Detailed descriptions of these datasets and training configurations are listed in Appendix C.

We evaluate the potential bias by comparing the model recall (TPR) value of both $\theta_0$ and $\theta_s$ against different abnormal test data. We define the bias to be upward ($\uparrow$) if TPR$(\theta_s)$ $>$ TPR$(\theta_0)$, and downward ($\downarrow$) if TPR$(\theta_s)$ $<$ TPR$(\theta_0)$.

We group our experiments into two scenarios: (1) when abnormal training data is visually similar to normal training data, and (2) when abnormal training data is visually dissimilar to normal training data. Here we compute visual similarity as the $L^2$ distance. The similarity result between any two classes is listed in Appendix E.

We observe similar trends across all three datasets and all six anomaly detection models. For brevity, we summarize our observations below, and further illustrate them using examples that consider two models (Deep SVDD as $\theta_0$ and Deep SAD as $\theta_s$), and one dataset (Fashion-MNIST). We list additional results on other models and datasets in Appendix E.

**Scenario 1: Abnormal training data visually similar to normal training data.** In this scenario, the use of semi-supervised training does improve detection of abnormal data in the training class, but also creates considerable performance changes, both upward and downward, for other classes of abnormal test data. The direction of change depends heavily on the similarity of the abnormal




(a) Scenario 1 (b) Scenario 2

Figure 3: Model TPR under Scenario 1 and 2, for Fashion-MNIST. We mark $\theta_0$ = Deep SVDD as "unsup" and $\theta_s$ = Deep SAD as "semi-sup". We arrange abnormal test data (identified by their class label) in decreasing similarity with training abnormal data. The leftmost entry in each figure is the class used for abnormal training.

test data to the training abnormal data. The model performance on test data similar to the training abnormal data moves *upward* significantly while that on test data dissimilar to the training abnormal moves *downward* significantly.

We illustrate this observation using an example of Fashion-MNISST. Here the normal training class is `top` and the abnormal training class is `shirt`, which are similar to each other. Figure 3 plots the recalls of model $\theta_0$ and $\theta_s$ for all abnormal test classes, arranged by a decreasing similarity to the training abnormal class (`shirt`). We see that TPR$(\theta_s)$ on classes similar to `shirt` (including itself) is significantly higher than TPR$(\theta_0)$ (e.g. increased from <0.2 to 0.9 for `pullover`). But for classes dissimilar from the training abnormal class, TPR$(\theta_s)$ is either similar or significantly lower (e.g., reduced from 0.9 to 0.4 for `boot`).

**Scenario 2: Abnormal training data visually dissimilar to normal training data.** Like Scenario 1, the use of semi-supervised training does improve the detection of abnormal data belonging to the training class and those similar to the training class. But different from Scenario 1, we observe very little downward changes at abnormal classes dissimilar to the training abnormal.

This is illustrated using another Fashion-MNIST example in Figure 3(b). While the normal training class is still `top`, we use a new abnormal training class of `sneaker` that is quite dissimilar from `top`. We see that TPR$(\theta_s)$ on `sneaker`, `sandal`, `boot` and `bag` are largely elevated to 0.8 and higher, while TPR$(\theta_s)$ on other classes are relatively stable (except for `trouser` which more than doubles).

**Summary of observations.** Our empirical study shows that a biased (anomaly) training set can introduce significant impact on semi-supervised anomaly detection models. When the labeled anomalies are similar to the normal instances, the trained model is likely to face large performance degradation on unseen anomalies that are different from the labeled anomalies. But when the labeled anomalies are dissimilar to the normal instances, the semi-supervised model becomes more useful and produce a consistent improvement in unseen anomalies.

## 6 Conclusions and Future Work

To the best of our knowledge, our work provides the first formal analysis on how a biased anomaly training set affects deep learning based semi-supervised anomaly detection. We define and formulate its impact on anomaly detector's recall (or TPR) as the relative *scoring bias* of the detector when comparing to its unsupervised baseline. We then establish the first finite sample rates for estimating the relative scoring bias for semi-supervised anomaly detection, and empirically validate our theoretical results on both synthetic and real-world datasets. Finally, we also empirically study how such relative scoring bias translates into variance in detector performance against different unseen anomalies, and demonstrate scenarios in which the biased anomaly set can be useful or harmful.

Our work exposes a new challenge in building semi-supervised models. An open question is how to construct an unbiased anomaly detector, even when having access to the true anomaly distribution. In addition, as future work, we plan to develop new training procedures that can leverage labeled anomalies to exploit upward scoring bias while avoiding downward scoring bias.

## REFERENCES


Charu C Aggarwal and Saket Sathe. Theoretical foundations and algorithms for outlier ensembles. *Acm Sigkdd Explorations Newsletter*, 17(1):24–47, 2015.

Tal Daniel, Thanard Kurutach, and Aviv Tamar. Deep variational semi-supervised novelty detection. *arXiv preprint arXiv:1911.04971*, 2019.

Taylor Denouden, Rick Salay, Krzysztof Czarnecki, Vahdat Abdelzad, Buu Phan, and Sachin Vernekar. Improving reconstruction autoencoder out-of-distribution detection with mahalanobis distance. *arXiv preprint arXiv:1812.02765*, 2018.

Andrew Emmott, Shubhomoy Das, Thomas Dietterich, Alan Fern, and Weng-Keen Wong. A meta-analysis of the anomaly detection problem. *arXiv preprint arXiv:1503.01158*, 2015.

Izhak Golan and Ran El-Yaniv. Deep anomaly detection using geometric transformations. In *Advances in Neural Information Processing Systems*, pp. 9758–9769, 2018.

Steve Hanneke and Samory Kpotufe. On the value of target data in transfer learning. In *Advances in Neural Information Processing Systems*, pp. 9871–9881, 2019.

Dan Hendrycks, Kimin Lee, and Mantas Mazeika. Using pre-training can improve model robustness and uncertainty. *Proceedings of the International Conference on Machine Learning*, 2019a.

Dan Hendrycks, Mantas Mazeika, and Thomas Dietterich. Deep anomaly detection with outlier exposure. *Proceedings of the International Conference on Learning Representations*, 2019b.

Dan Hendrycks, Mantas Mazeika, Saurav Kadavath, and Dawn Song. Using self-supervised learning can improve model robustness and uncertainty. In *Advances in Neural Information Processing Systems*, pp. 15663–15674, 2019c.

Michael J Kearns, Umesh Virkumar Vazirani, and Umesh Vazirani. *An introduction to computational learning theory*. MIT press, 1994.

Samory Kpotufe and Guillaume Martinet. Marginal singularity, and the benefits of labels in covariate-shift. *arXiv preprint arXiv:1803.01833*, 2018.

Kimin Lee, Honglak Lee, Kibok Lee, and Jinwoo Shin. Training confidence-calibrated classifiers for detecting out-of-distribution samples. *Proceedings of the International Conference on Learning Representations*, 2018.

Kun Liu and Huadong Ma. Exploring background-bias for anomaly detection in surveillance videos. In *Proceedings of the 27th ACM International Conference on Multimedia*, pp. 1490–1499, 2019.

Si Liu, Risheek Garrepalli, Thomas G Dietterich, Alan Fern, and Dan Hendrycks. Open category detection with pac guarantees. *arXiv preprint arXiv:1808.00529*, 2018.

P. Massart. The tight constant in the dvoretzky-kiefer-wolfowitz inequality. *Ann. Probab.*, 18(3): 1269–1283, 07 1990. doi: 10.1214/aop/1176990746. URL `https://doi.org/10.1214/aop/1176990746`.

Guansong Pang, Chunhua Shen, and Anton van den Hengel. Deep anomaly detection with deviation networks. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pp. 353–362, 2019.

Emanuel Parzen. Quantile functions, convergence in quantile, and extreme value distribution theory. Technical report, TEXAS A AND M UNIV COLLEGE STATION INST OF STATISTICS, 1980.

Marco AF Pimentel, David A Clifton, Lei Clifton, and Lionel Tarassenko. A review of novelty detection. *Signal Processing*, 99:215–249, 2014.

Shebuti Rayana, Wen Zhong, and Leman Akoglu. Sequential ensemble learning for outlier detection: A bias-variance perspective. In *2016 IEEE 16th International Conference on Data Mining (ICDM)*, pp. 1167–1172. IEEE, 2016.

Lukas Ruff, Robert A. Vandermeulen, Nico Görnitz, Lucas Deecke, Shoaib A. Siddiqui, Alexander Binder, Emmanuel Müller, and Marius Kloft. Deep one-class classification. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80, pp. 4393–4402, 2018.

Lukas Ruff, Robert A. Vandermeulen, Billy Joe Franks, Klaus-Robert Müller, and Marius Kloft. Rethinking assumptions in deep anomaly detection, 2020a.

Lukas Ruff, Robert A. Vandermeulen, Nico Görnitz, Alexander Binder, Emmanuel Müller, Klaus-Robert Müller, and Marius Kloft. Deep semi-supervised anomaly detection. In *Proc. of ICLR*, 2020b.

Mayu Sakurada and Takehisa Yairi. Anomaly detection using autoencoders with nonlinear dimensionality reduction. In *Proceedings of the MLSDA 2014 2nd Workshop on Machine Learning for Sensory Data Analysis*, pp. 4–11, 2014.

Bernhard Schölkopf, Robert Williamson, Alex Smola, John Shawe-Taylor, and John Platt. Support vector method for novelty detection. In *Proceedings of the 12th International Conference on Neural Information Processing Systems*, pp. 582–588, 1999.

Md Amran Siddiqui, Alan Fern, Thomas G Dietterich, and Shubhomoy Das. Finite sample complexity of rare pattern anomaly detection. In *UAI*, 2016.

Ashwin Srinivasan. Uci machine learning repository: statlog (landsat satellite) data set, 2010.

Alexander Tong, Roozbah Yousefzadeh, Guy Wolf, and Smita Krishnaswamy. Fixing bias in reconstruction-based anomaly detection with lipschitz discriminators. *arXiv preprint arXiv:1905.10710*, 2019.

Leslie G Valiant. A theory of the learnable. *Communications of the ACM*, 27(11):1134–1142, 1984.

Siqi Wang, Yijie Zeng, Xinwang Liu, En Zhu, Jianping Yin, Chuanfu Xu, and Marius Kloft. Effective end-to-end unsupervised outlier detection via inlier priority of discriminative network. In *Advances in Neural Information Processing Systems 32*, pp. 5960–5973. Curran Associates, Inc., 2019a.

Zirui Wang, Zihang Dai, Barnabás Póczos, and Jaime Carbonell. Characterizing and avoiding negative transfer. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 11293–11302, 2019b.

Sen Wu, Hongyang R Zhang, and Christopher Ré. Understanding and improving information transfer in multi-task learning. *arXiv preprint arXiv:2005.00944*, 2020.

Han Xiao, Kashif Rasul, and Roland Vollgraf. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms, 2017.

Yuki Yamanaka, Tomoharu Iwata, Hiroshi Takahashi, Masanori Yamada, and Sekitoshi Kanai. Autoencoding binary classifiers for supervised anomaly detection, 2019.

Chong Zhou and Randy C. Paffenroth. Anomaly detection with robust deep autoencoders. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2017.

# A PROOF OF COROLLARY 2

*Proof of Corollary 2.* Assuming the score functions are Gaussian distributed, we can denoted $F_0(s)$ as $\Phi(\frac{s-\mu_0}{\sigma_0})$, $\tilde{F}_0(s)$ as $\Phi(\frac{s-\tilde{\mu}_0}{\tilde{\sigma}_0})$, $F_a(s)$ as $\Phi(\frac{s-\mu_a}{\sigma_a})$, and $\tilde{F}_a(s)$ as $\Phi(\frac{s-\tilde{\mu}_a}{\tilde{\sigma}_a})$.

Therefore, we have $\Delta_0 = |(\sigma_0\Phi^{-1}(q) + \mu_0) - (\tilde{\sigma}_0\Phi^{-1}(q) + \tilde{\mu}_0)|$.

Thus,

$$\begin{aligned}\xi_N &:= \tilde{r} - r \\ &= F_a(F_0^{-1}(q)) - \tilde{F}_a(\tilde{F_0^{-1}}(q)) \\ &= \Phi(\frac{\sigma_0\Phi^{-1}(q)}{\sigma_a} + \frac{\mu_0 - \mu_a}{\sigma_a}) - \Phi(\frac{\tilde{\sigma}_0\Phi^{-1}(q)}{\tilde{\sigma}_a} + \frac{\tilde{\mu}_0 - \tilde{\mu}_a}{\tilde{\sigma}_a})\end{aligned}$$

□

# B PROOF OF THEOREM 3

*Proof of Theorem 3.* Our proof builds upon and extends the analysis framework of Liu et al. (2018), which relies on one key result from Massart (1990),

$$\mathbb{P}\left[\sqrt{n}\sup_x |\hat{F}(x) - F(x)| > \lambda\right] \leq 2\exp(-2\lambda^2). \tag{B.1}$$

Here, $\hat{F}(x)$ is the empirical CDF calculated from $n$ samples. Given a fixed threshold function, Liu et al. (2018) showed that it required

$$n > \frac{1}{2\epsilon_1^2}\log\frac{2}{1-\sqrt{1-\delta}} \cdot \left(\frac{2-\alpha}{\alpha}\right)^2 \tag{B.2}$$

examples, in order to guarantee $|\hat{F}_a(x) - F_a(x)| \leq \epsilon_1$ with probability at least $1-\delta$ (recall that $\alpha$ denotes the fraction of abnormal data among the $n$ samples).

We further assume the Lipschitz continuity for the CDFs/quantile functions:

$$|F_a(x) - F_a(x')| \leq \ell_a|x - x'| \tag{B.3}$$

$$|F_a'(x) - F_a'(x')| \leq \ell_a'|x - x'| \tag{B.4}$$

$$|F_0{}^-(x) - F_0{}^-(x')| \leq \ell_0^-|x - x'| \tag{B.5}$$

$$|F_0'^-(x) - F_0'^-(x')| \leq \ell_0'^-|x - x'| \tag{B.6}$$

Combining the above inequalities equation B.5 with equation B.1, we obtain

$$\mathbb{P}\left[\sup_q \left|\hat{F}_0^{-1}(q) - F_0^{-1}(q)\right| \geq \frac{\lambda}{\sqrt{n_0}}\right] \leq \mathbb{P}\left[\sup_q \left|F_0\left(\hat{F}_0^{-1}(q)\right) - F_0\left(F_0^{-1}(q)\right)\right| \geq \frac{\lambda}{\sqrt{n_0}\ell_0^-}\right] \tag{B.7}$$

Let $q = \hat{F}_0(x)$, then equation B.7 becomes

$$\begin{aligned}&\mathbb{P}\left[\left|F_0\left(\hat{F}_0^{-1}(\hat{F}_0(x))\right) - F_0\left(F_0^{-1}(\hat{F}_0(x))\right)\right| \geq \frac{\lambda}{\sqrt{n_0}\ell_0^-}\right] \\ =&\mathbb{P}\left[\left|F_0\left(x\right) - \hat{F}_0(x))\right| \geq \frac{\lambda}{\sqrt{n_0}\ell_0^-}\right] \\ \leq& 2e^{-\frac{2}{n_0}\left(\frac{\lambda}{\ell_0^-}\right)^2}\end{aligned} \tag{B.8}$$

Therefore, in order for $\mathbb{P}\left[\sup_q \left|\hat{F}_0^{-1}(q) - F_0^{-1}(q)\right| \geq \epsilon_2\right] \leq \delta$ to hold, it suffices to set

$$n_0 = n(1-\alpha) > \frac{1}{2}\frac{1}{\epsilon_2^2}\frac{1}{\ell_0^{-2}}\log\frac{2}{\delta} \tag{B.9}$$

Furthermore, combining equation B.1, equation B.2, and equation B.3, we get

$$\hat{F}_a(\tau_n) \leq F_a(\tau) + (\tau - \tau_n)\ell_a + \epsilon_1 \tag{B.10}$$

Subtitute $\tau_n = \hat{F}_0^{-1}(q)$, and $\tau = F_0^{-1}(q)$ in the above inequality, and set $\epsilon_1 = \frac{\epsilon}{4}$, $\epsilon_2 = \frac{\epsilon}{4\ell_a}$, we get

$$\left|\hat{F}_a\left(\hat{F}_0^{-1}(q)\right) - F_a\left(F_0^{-1}(q)\right)\right| \leq \epsilon/2. \tag{B.11}$$

Similary, we repeat the same procedure for $s'$, and can get

$$\left|\hat{F'}_a\left(\hat{F}_0^{'-1}(q)\right) - F'_a\left(F_0^{'-1}(q)\right)\right| \leq \epsilon/2. \tag{B.12}$$

with probability at least $1 - \delta$.

Therefore, with $n \geq \frac{8}{\epsilon^2} \cdot \left(\log \frac{2}{1-\sqrt{1-\delta}} \cdot \left(\frac{2-\alpha}{\alpha}\right)^2 + \log \frac{2}{\delta} \cdot \frac{1}{1-\alpha}\left(\left(\frac{\ell_a}{\ell_0^-}\right)^2 + \left(\frac{\ell'_a}{\ell_0'^-}\right)^2\right)\right)$ examples we can get $|\hat{\xi} - \xi| \leq \epsilon$ with probability at least $1 - \delta$.

□

## C Datasets and Training Configurations

**Fashion-MNIST.** This dataset (Xiao et al. (2017)) is a collection of 70,000 grayscale images on fashion objects (a training set of 60,000 examples and a test set of 10,000 examples), evenly divided into 10 classes (7000 images per class). Each image is of 28 pixels in height and 28 pixels in width, and each pixel-value is an integer between 0 and 255. The 10 classes are denoted as `top`, `trouser`, `pullover`, `dress`, `coat`, `sandal`, `shirt`, `sneaker`, `bag`, `boot`.

To train our anomaly detection models, we pick one class as the normal training class, and another class as the abnormal training class, the rest as the abnormal testing class. We use the full training set of the normal class (6000), and a random 10% of the training set of the abnormal training class (600) to train the unsupervised and semi-supervised anomaly detection models. We use the test data of the normal class (1000) to configure the anomaly scoring thresholds to meet a 5% false positive rate (FPR). We then test the models on the full data of each abnormal testing class, as well as on the untrained fraction of the data of the abnormal training class.

**StatLog.** This dataset (Srinivasan (2010)) is a collection of 6435 NASA satellite images, each of $82 \times 100$ pixels, valued between 0 and 255. The six labeled classes are denoted as `red soil`, `cotton crop`, `grey soil`, `damp grey soil`, `soil with vegetation stubble`, and `very damp grey soil`.

If not otherwise specified, we follow the same procedure (as the above) to train the unsupervised and semi-supervised anomaly detection models. The normal training data includes 80% data of the designated class, and the abnormal training data is 10% of the normal training data in size. Due to the insufficient amount of the data, we use the full data of the normal data to configure the anomaly scoring thresholds to meet a 5% false positive rate (FPR). We then test the models on the full data of each abnormal testing class, as well as the untrained fraction of the data of the abnormal training class.

**ImageNet.** This is a very large collection of colored images on various objects. We select 10 classes, each with 1300 images in total. The 10 classes are denoted as: `thunder snake`, `ringneck snake`, `basketball`, `baseball`, `African crocodile`, `American alligator`, `laptop`, `desktop`, `hammerhead shark`, `tiger shark`.

We follow the same procedure (as the above) to train the unsupervised and semi-supervised anomaly detection models, set the thresholds and test on the data. The normal training data includes 85% of 1300 images, and the abnormal training data is 10% of the normal training data in size (111).

## D ADDITIONAL EXPERIMENT RESULTS OF SECTION 4.2

**Anomaly score distributions for models trained on the synthetic dataset.** We plot in Figure 4–6 the anomaly score distributions for unsupervised and semi-supervised models, all trained on the synthetic dataset.

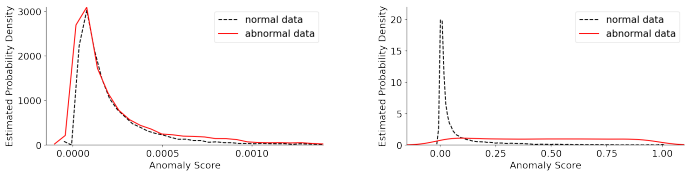



Figure 4: Anomaly score distributions for Deep SVDD (left) and HSC (right), trained on the synthetic dataset ($\alpha = 0.1$, $n = 10000$), estimated by Kernel Density Estimation.

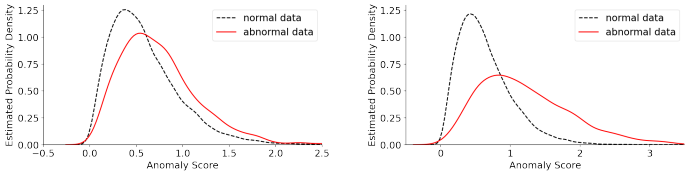



Figure 5: Anomaly score distributions for AE (left) and SAE (right), trained on the synthetic dataset ($\alpha = 0.1$, $n = 10000$), estimated by Kernel Density Estimation.

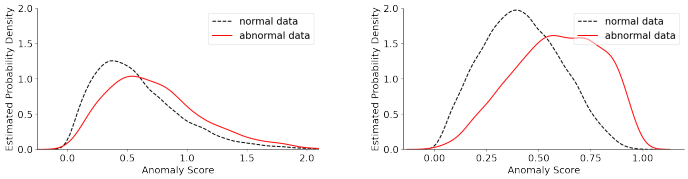



Figure 6: Anomaly score distributions for AE (left) and ABC (right), trained on the synthetic dataset ($\alpha = 0.1$, $n = 10000$), estimated by Kernel Density Estimation.

**Anomaly score distributions for models trained on a real-world dataset.** We plot in Figure 7 the anomaly score distributions for unsupervised and semi-supervised models, when trained on the Fashion-MNIST dataset.

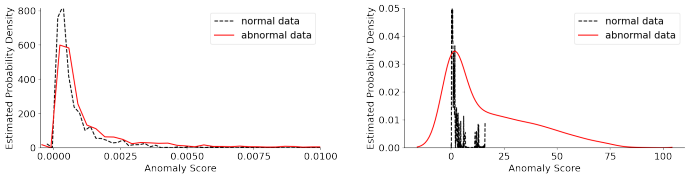
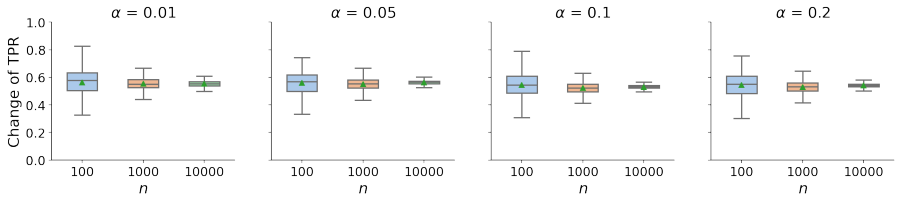
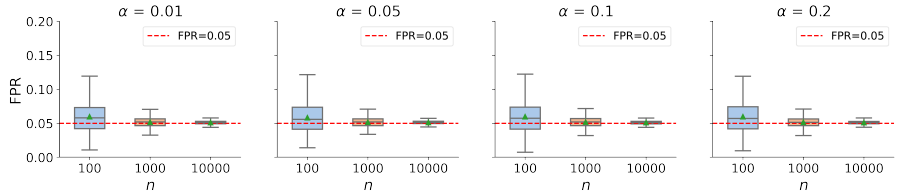



Figure 7: Anomaly score distributions for Deep SVDD (left) and Deep SAD (right), trained on the Fashion-MNIST dataset (using the class `top` as the normal class and `shirt` as the abnormal class), estimated by Kernel Density Estimation.

**Convergence of relative scoring bias $\xi$ and FRP.** We plot in Figure 8–10 the additional results on Deep SVDD vs. HSC, AE vs. SAE, and AE vs. ABC. Overall, they show a consistent trend on convergence.

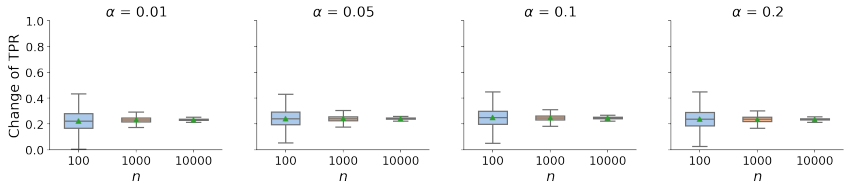



(a) Relative Scoring Bias ($\xi$) between Deep SVDD and HSC

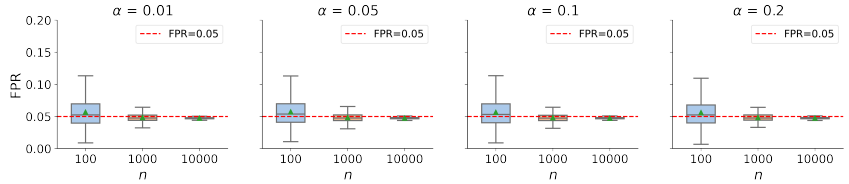



(b) FPR of HSC

Figure 8: The quantile distribution of (a) relative scoring bias $\xi$ and (b) FPR, computed on the test set over 1500 runs, for Deep SVDD and HSC.

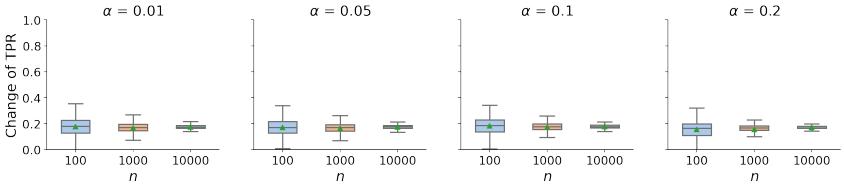



(a) Relative Scoring Bias ($\xi$) between AE and SAE

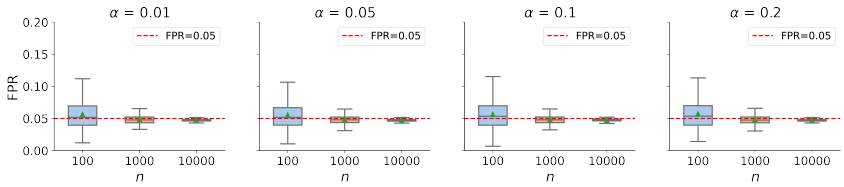



(b) FPR of SAE

Figure 9: The quantile distribution of (a) relative scoring bias $\xi$ and (b) FPR, computed on the test set over 1500 runs, for AE and SAE.




(a) Relative Scoring Bias ($\xi$) between AE and ABC




(b) FPR of ABC

Figure 10: The quantile distribution of (a) relative scoring bias $\xi$ and (b) FPR, computed on the test set over 1500 runs, for AE and ABC.

# E ADDITIONAL RESULTS OF SECTION 5

**Scenario 1.** We consider scenario 1 where the training normal set is visually similar to the training abnormal set. We include the detailed recall result of all six models, and three real-life datasets in Table 2 – 4. In each table, the abnormal test classes are sorted by decreasing similarity to the abnormal training class (measured by $L^2$, small value = visually similar). Also, ↑ indicates that the semi-supervised model has a higher recall than the unsupervised model. ↓ indicates the other direction. Overall we observe both upward and downward bias across the test abnormal classes, and the direction depends on the similarity to the train abnormal class.

We also observe that when using the reconstruction based models (AE, SAE, ABC), the performance for Statlog and ImageNet is much worse than the hypersphere based models. We consider this as a result potentially arising from specific latent spaces of data on the reconstruction based models. As discussed in (Denouden et al., 2018), reconstruction based approaches can sometimes fail to capture the anomalies which are away from the normal class in the latent space yet those near the normal class in the manifold defined by AE models. In addition, for ImageNet, the input data is much more complex than those in other datasets, which adds more difficulty for AE models to reconstruct. We leave improvement of these reconstruction models to future work.

training normal = `top`, training abnormal = `shirt`

| Test data | **Deep SVDD** | **Deep SAD** | **HSC** | **AE** | **SAE** | **ABC** | $L^2$ |
|---|---|---|---|---|---|---|---|
| `shirt` | 0.12 | 0.73 ↑ | 0.73↑ | 0.12 | 0.72↑ | 0.72↑ | 0 |
| `pullover` | 0.17 | 0.93↑ | 0.89↑ | 0.21 | 0.83↑ | 0.83↑ | 0.009 |
| `coat` | 0.15 | 0.93↑ | 0.91↑ | 0.16 | 0.90↑ | 0.89↑ | 0.012 |
| `dress` | 0.14 | 0.24↑ | 0.27↑ | 0.13 | 0.23↑ | 0.19↑ | 0.035 |
| `bag` | 0.52 | 0.47↓ | 0.46↓ | 0.71 | 0.55↓ | 0.51↓ | 0.036 |
| `trouser` | 0.14 | 0.11↓ | 0.06↓ | 0.64 | 0.10↓ | 0.10↓ | 0.062 |
| `boot` | 0.95 | 0.44↓ | 0.24↓ | 0.98 | 0.93↓ | 0.89↓ | 0.085 |
| `sandal` | 0.32 | 0.41↑ | 0.23↓ | 0.83 | 0.53↓ | 0.50↓ | 0.088 |
| `sneaker` | 0.61 | 0.21↓ | 0.18↓ | 0.72 | 0.30↓ | 0.51↓ | 0.103 |

Table 2: The model TPR under scenario 1, Fashion-MNIST. The normal class `top` is similar to the abnormal training class `shirt`.

training normal = `very damp grey soil`, training abnormal = `damp gray soil`

| Test data | **Deep SVDD** | **Deep SAD** | **HSC** | **AE** | **SAE** | **ABC** | $L^2$ |
|---|---|---|---|---|---|---|---|
| `damp grey soil` | 0.09 | 0.80 ↑ | 0.82↑ | 0.00 | 0.11↑ | 0.10↑ | 0 |
| `red soil` | 0.45 | 0.94↑ | 0.92↑ | 0.00 | 0.00 | 0.00 | 4.39 |
| `grey soil` | 0.32 | 0.93↑ | 0.89↑ | 0.01 | 0.03↑ | 0.01 | 4.42 |
| `vegetable soil` | 0.27 | 0.20↓ | 0.27 | 0.35 | 0.35 | 0.34 | 5.44 |
| `cotton crop` | 0.99 | 0.71↓ | 0.71↓ | 0.89 | 0.89 | 0.90 | 11.46 |

Table 3: The model TPR under scenario 1, StatLog. The normal class `very damp grey soil` is similar to the abnormal training class `damp gray soil`.

training normal = `tiger shark`, training abnormal = `hammerhead shark`

| Test data | **Deep SVDD** | **Deep SAD** | **HSC** | **AE** | **SAE** | **ABC** | $L^2$ |
|---|---|---|---|---|---|---|---|
| `hammerhead shark` | 0.17 | 0.42 ↑ | 0.51↑ | 0.10 | 0.10 | 0.12 | 0 |
| `desktop computer` | 0.87 | 0.81↓ | 0.86↓ | 0.14 | 0.25↑ | 0.30↑ | 0.396 |
| `laptop` | 0.75 | 0.72↓ | 0.77↑ | 0.10 | 0.15↑ | 0.18↑ | 0.409 |
| `alligator` | 0.68 | 0.70↑ | 0.72↑ | 0.02 | 0.03 | 0.05 | 0.451 |
| `crocodile` | 0.71 | 0.72↑ | 0.76↑ | 0.01 | 0.02 | 0.02 | 0.561 |
| `baseball` | 0.79 | 0.72↓ | 0.77↓ | 0.06 | 0.10↑ | 0.15↑ | 0.612 |
| `basketball` | 0.87 | 0.74↓ | 0.77↓ | 0.05 | 0.11↑ | 0.22↑ | 0.634 |
| `ring snake` | 0.76 | 0.71↓ | 0.72↓ | 0.02 | 0.03 | 0.06 | 0.644 |
| `thunder snake` | 0.77 | 0.72↓ | 0.75↓ | 0.01 | 0.02 | 0.04 | 0.699 |

Table 4: The model TPR under scenario 1, ImageNet. The normal class `tiger shark` is similar to the abnormal training class `hammerhead shark`.

**Scenario 2.** We consider scenario 2 where the training normal set is visually dissimilar to the training abnormal set. We include the detailed TPR result of all six models, and three real-life datasets in Table 5 – 7. Like the above, in each table, the abnormal test classes are sorted by decreasing similarity to the abnormal training class. ↑ indicates that the semi-supervised model has a higher TPR than the unsupervised model. ↓ indicates the other direction.

Different from Scenario 1, here we observe mostly upward changes, i.e, the semi-supervised learning gets higher TPR than its unsupervised version. Again we observe poorer performance of AE, SAE, ABC on Statlog and ImageNet compared to the hypersphere-based models.

training normal = `top`, training abnormal = `sneaker`

| Test data | **Deep SVDD** | **Deep SAD** | **HSC** | **AE** | **SAE** | **ABC** | $L^2$ |
|---|---|---|---|---|---|---|---|
| `sneaker` | 0.61 | 1.00 ↑ | 1.00↑ | 0.72 | 1.00↑ | 1.00↑ | 0 |
| `sandal` | 0.32 | 0.99↑ | 0.96↑ | 0.83 | 1.00↑ | 0.99↑ | 0.019 |
| `boot` | 0.95 | 1.00↑ | 1.00↑ | 0.98 | 1.00↑ | 1.00↑ | 0.066 |
| `bag` | 0.51 | 0.80↑ | 0.66↑ | 0.71 | 0.84↑ | 0.80↑ | 0.070 |
| `shirt` | 0.12 | 0.09↓ | 0.11↓ | 0.12 | 0.13↑ | 0.19↑ | 0.103 |
| `dress` | 0.15 | 0.12↓ | 0.05↓ | 0.13 | 0.12↓ | 0.11↓ | 0.116 |
| `trouser` | 0.14 | 0.37↑ | 0.14 | 0.64 | 0.60↓ | 0.55↓ | 0.125 |
| `pullover` | 0.17 | 0.12↓ | 0.15↓ | 0.21 | 0.19↓ | 0.16↓ | 0.127 |
| `coat` | 0.15 | 0.09↓ | 0.13↓ | 0.16 | 0.15↓ | 0.13↓ | 0.142 |

Table 5: The model TPR under scenario 2, Fashion-MNIST. The normal class `top` is dissimilar to the abnormal training class `sneaker`.

training normal = `very damp grey soil`, training abnormal = `red soil`

| Test data | **Deep SVDD** | **Deep SAD** | **HSC** | **AE** | **SAE** | **ABC** | $L^2$ |
|---|---|---|---|---|---|---|---|
| `red soil` | 0.45 | 1.00↑ | 1.00↑ | 0.00 | 0.25↑ | 0.21↑ | 0 |
| `damp grey soil` | 0.09 | 0.32↑ | 0.21↑ | 0.00 | 0.00 | 0.00 | 4.39 |
| `grey soil` | 0.32 | 0.85↑ | 0.41↑ | 0.01 | 0.01 | 0.02 | 5.48 |
| `vegetable soil` | 0.27 | 0.61↑ | 0.82↑ | 0.35 | 0.42↑ | 0.42↑ | 6.63 |
| `cotton crop` | 0.99 | 0.99 | 1.00 | 0.89 | 0.94↑ | 0.93↑ | 8.97 |

Table 6: The model TPR under scenario 2, Statlog. The normal class `very damp grey soil` is dissimilar to the abnormal training class `red soil`.

training normal `tiger shark`, training abnormal = `basketball`

| Test data | **Deep SVDD** | **Deep SAD** | **HSC** | **AE** | **SAE** | **ABC** | $L^2$ |
|---|---|---|---|---|---|---|---|
| `basketball` | 0.87 | 1.00 ↑ | 0.99↑ | 0.05 | 0.67 ↑ | 0.62 ↑ | 0 |
| `crocodile` | 0.71 | 0.88↑ | 0.80↑ | 0.01 | 0.06↑ | 0.07↑ | 0.020 |
| `alligator` | 0.68 | 0.81↑ | 0.70↑ | 0.02 | 0.08↑ | 0.11 ↑ | 0.026 |
| `ring snake` | 0.76 | 0.95↑ | 0.90↑ | 0.02 | 0.21↑ | 0.25↑ | 0.035 |
| `laptop` | 0.75 | 0.90↑ | 0.78↑ | 0.10 | 0.35 ↑ | 0.26 ↑ | 0.039 |
| `thunder snake` | 0.77 | 0.94↑ | 0.88↑ | 0.01 | 0.15↑ | 0.18↑ | 0.058 |
| `baseball` | 0.79 | 0.96↑ | 0.90↑ | 0.06 | 0.30 ↑ | 0.24 ↑ | 0.072 |
| `desktop` | 0.87 | 0.95↑ | 0.80↓ | 0.14 | 0.43↑ | 0.34↑ | 0.129 |
| `hammerhead shark` | 0.17 | 0.17↑ | 0.11↑ | 0.10 | 0.11 | 0.11 | 0.634 |

Table 7: The model TPR under scenario 2, ImageNet. The normal class is `tiger shark` is dissimilar to the abnormal training class `basketball`.