Fairness-aware Anomaly Detection via Fair Projection

Feng Xiao[1] **Xiaoying Tang**[1] **Jicong Fan**[2*]
[1]School of Science and Engineering, The Chinese University of Hong Kong (Shenzhen),
Longgang, Shenzhen, Guangdong, 518172, P.R. China
[2]School of Data Science, The Chinese University of Hong Kong (Shenzhen),
Longgang, Shenzhen, Guangdong, 518172, P.R. China

Abstract

Unsupervised anomaly detection is a critical task in many high-social-impact applications such as finance, healthcare, social media, and cybersecurity, where demographics involving age, gender, race, disease, etc. are used frequently. In these scenarios, possible bias from anomaly detection systems can lead to unfair treatment for different groups and even exacerbate social bias. In this work, first, we thoroughly analyze the feasibility and necessary assumptions for ensuring group fairness in unsupervised anomaly detection. Second, we propose a novel fairness-aware anomaly detection method FairAD. From the normal training data, FairAD learns a projection to map data of different demographic groups to a common target distribution that is simple and compact, and hence provides a reliable base to estimate the density of the data. The density can be directly used to identify anomalies while the common target distribution ensures fairness between different groups. Furthermore, we propose a threshold-free fairness metric that provides a global view for model's fairness, eliminating dependence on manual threshold selection. Experiments on real-world benchmarks demonstrate that our method achieves an improved trade-off between detection accuracy and fairness under both balanced and skewed data across different groups.

1 Introduction

As machine learning techniques are increasingly being applied in high-social-impact fields such as finance and justice, the fairness of machine learning systems receives a surge of attention. There are growing studies [Larson et al., 2016, Hendricks et al., 2018, Dastin, 2022] that exhibit discrimination in real-world machine learning systems. For instance, analysis [Larson et al., 2016] on the COMPAS (Correctional Offender Management Profiling for Alternative Sanctions), a recidivism risk prediction system, shows a strong correlation between recidivism prediction and race, where African-American individuals have a much higher risk of recidivism than Caucasian. It is vital to eliminate or mitigate the possible disparity of machine learning algorithms between different demographic groups to ensure social fairness.

In response, many researchers [Dwork et al., 2012a, Zemel et al., 2013, Tzeng et al., 2014, Hardt et al., 2016, Zafar et al., 2017a, Agarwal et al., 2019, Oh et al., 2022, Jovanović et al., 2023] have begun to propose fairness-aware machine learning algorithms, where one of the most common paradigms is to learn a fair representation depending on a formal fairness principle [Hajian et al., 2016, Hardt et al., 2016, Zafar et al., 2017b, Gajane, 2017]. However, previous works on algorithm fairness often focused on supervised machine learning tasks [Hardt et al., 2016, Agarwal et al., 2018, 2019] and the study on the fairness issue in unsupervised anomaly detection [Pang et al., 2021] is scarce. Anomaly detection (AD), aiming at identifying anomalous samples in data, plays a crucial

*Corresponding author

39th Conference on Neural Information Processing Systems (NeurIPS 2025).

role in many important fields such as finance, healthcare, social media, and cybersecurity, where demographics, like age, gender, race, ethnicity, and disease, are used frequently. Given multiple groups (≥ 2) partitioned by a protected social variable with multiple attribute values, a fair AD method is supposed to ensure equal probabilities of samples being detected as anomalous (or normal) across different demographic groups[2]. However, as shown by [Zhang and Davidson, 2021], existing AD methods suffer from unfairness to some extent. In addition, Meissen et al. [2023] and Wu et al. [2024] found that the fairness of unsupervised anomaly detection models is easily affected by sample proportion across different demographic groups.

Surprisingly, studies in the literature focusing on fairness in unsupervised AD are quite limited and incomplete. Although there have been several attempts such as [Deepak and Abraham, 2020, Zhang and Davidson, 2021, Shekhar et al., 2021, Han et al., 2023], we still encounter, at least, the following four problems or difficulties:

I The feasibility and necessary assumptions of ensuring fairness in unsupervised anomaly detection haven't been clearly discussed in the literature.
II The trade-off between fairness and utility criteria (such as accuracy) does not meet pressing practical needs.
III There is still a lack of global and convenient (threshold-free) evaluation metrics of fairness for unsupervised anomaly detection methods.
IV There is a lack of research on the fairness issues under different data splitting strategies including balanced and skewed data splitting across demographic groups.

In this work, we attempt to address the four difficulties. We present a novel fairness-aware unsupervised AD method called FairAD. FairAD learns to map data from different demographic groups to a common target distribution, which ensures statistical parity for different groups in the target distribution space. The chosen target distribution is supposed to be simple and compact, where simplicity ensures that sampling from the target distribution is easy, and compactness aims to obtain a reliable decision boundary to distinguish between normal and abnormal samples. Furthermore, in order to effectively evaluate fairness in anomaly detection, we propose a novel threshold-free evaluation metric. Our main contributions are as follows.

- We discuss the group fairness issue and introduce two fundamental assumptions for any fairness-aware (group fairness) unsupervised AD methods. Furthermore, we empirically demonstrate that the assumptions are reasonable in real-world scenarios. (for Difficulty I)
- We propose FairAD without introducing additional fairness regularization, which achieves a coordinated optimization process for detection task and group fairness and improves the fairness-utility trade-off, in comparison to existing methods. (for Difficulty II)
- We introduce a threshold-free fairness evaluation metric that holistically quantifies model fairness across the entire decision spectrum, eliminating dependence on manual threshold selection. (for Difficulty III)
- We consider both balanced and skewed data-splitting strategies across different demographic groups and evaluate all baselines in the two settings. (for Difficulty IV)

The experiments on real-world datasets show that our method achieves an improved trade-off between detection accuracy and fairness and the results also verify the effectiveness of the proposed evaluation metrics of fairness. The source code is provided in supplementary materials.

2 Related Work

2.1 Fair Representation Learning

Fair representation learning (FRL) [Cerrato et al., 2024] focuses on mitigating biases and ensuring fairness in machine learning systems by transforming data into latent space where sensitive attributes (e.g., race, gender) have minimal or no influence on outcomes. It aims to achieve equity in predictions across different demographic groups while maintaining task accuracy. The main technical routes include: (1) adversarial learning to promote independence between latent features and sensitive attributes [Xie et al., 2017, Madras et al., 2018, Zhang et al., 2018]; and (2) mutual information and variational inference [Louizos et al., 2015, Moyer et al., 2018, Creager et al., 2019, Oh et al., 2022];

[2]In this work, we focus on group fairness rather than individual fairness.

and (3) introducing fairness constraints Agarwal et al. [2018]. To some extent, these FRL techniques constitute a key approach to building fair machine learning systems.

Building on the FRL techniques, a straightforward strategy for achieving fair anomaly detection is to construct two-stage pipelines: (1) generating fair embeddings using existing FRL methods, followed by (2) training an anomaly detector on these embeddings. However, this paradigm exhibits two critical limitations in unsupervised anomaly detection scenario: (1) (**Task Compatibility**) most FRL methods are not disentangled with downstream tasks (typically classification), making them incompatible with unsupervised anomaly detection where all training samples belong to a single (normal) class and lack auxiliary label information, and (2) (**Detection Efficacy**) on the other hand, the task-agnostic representation is not guaranteed to be useful in distinguishing between normal samples and anomalies, leading to low detection accuracy. In contrast, our proposed method introduces an end-to-end framework with fairness mechanisms specifically designed for unsupervised anomaly detection. Empirical results (see Section 4) also demonstrated that our proposed method achieves significantly better detection accuracy than the two-stage pipelines while maintaining comparable fairness.

2.2 Fairness in Anomaly Detection

Despite so many works on anomaly detection problems [Ruff et al., 2018, Cai and Fan, 2022, Han et al., 2022, Bouman et al., 2024], the studies on the fairness of anomaly detection are limited. To the best of our knowledge, Davidson and Ravi [2020] first studied the fairness issue of outlier detection and proposed a framework to determine whether the output of an outlier detection algorithm is fair. Subsequently, Deepak and Abraham [2020] studied the fairness problem of LOF (Local Outlier Factor) [Breunig et al., 2000] and proposed a strategy to mitigate the unfairness of LOF on tabular datasets. Zhang and Davidson [2021] studied the fairness problem of Deep SVDD [Ruff et al., 2018] and proposed Deep Fair SVDD, which used an adversarial network to de-correlate the relationships between the sensitive attributes and the learned representations. Shekhar et al. [2021] added statistical parity regularization and group fidelity regularization on AutoEncoder (AE) [Hinton and Salakhutdinov, 2006] to mitigate the unfairness of AE-based anomaly detection methods. More recently, Han et al. [2023] studied counterfactual fairness, which is to ensure the consistency of the detection outcome in the factual and counterfactual world to different demographic groups.

Adversarial training [Zhang and Davidson, 2021] is known to be unstable, and fairness regularization terms [Shekhar et al., 2021] often compromise detection performance, and counterfactual-based methods [Han et al., 2023] introduce extra training complexity. In contrast to these methods, our proposed method achieves a simple and coordinated optimization process for the detection task and group fairness. Empirical results (See Section 4) also demonstrated that our method achieves superior detection accuracy compared to these methods, while maintaining comparable or even better fairness.

3 Fairness-aware Anomaly Detection

3.1 Preliminary Knowledge

Unsupervised AD Let $\mathcal{X} = \{\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_n\} \subseteq \mathbb{R}^d$ be a set of n samples drawn from an unknown distribution $\mathcal{D}_\mathbf{x}$ and the samples from $\mathcal{D}_\mathbf{x}$ are deemed as normal data. A point $\mathbf{x} \in \mathbb{R}^d$ is deemed to be anomalous if $\mathbf{x}$ is not drawn from $\mathcal{D}_\mathbf{x}$. Then, the goal of the unsupervised AD is to obtain a decision function $f : \mathbb{R}^d \to \{0, 1\}$ by utilizing only $\mathcal{X}$, such that $f(\mathbf{x}) = 0$ if $\mathbf{x}$ is drawn from $\mathcal{D}_\mathbf{x}$ and $f(\mathbf{x}) = 1$ if $\mathbf{x}$ is not drawn from $\mathcal{D}_\mathbf{x}$. Note that this is a standard setting of anomaly detection, followed by most unsupervised AD methods [Ruff et al., 2018, Goyal et al., 2020, Han et al., 2022, Fu et al., 2024, Xiao et al., 2025], where models are trained exclusively on normal data. The main difference among unsupervised AD methods is the design of the decision function f.

Group Fairness Demographic parity, a.k.a. statistical parity [Dwork et al., 2012b], demands the existence of parity between different demographic groups, such as those defined by gender or race. We use $S \in \mathcal{S} := \{s_1, s_2, \ldots, s_K\}$ to denote the sensitive or protected attribute and $|\mathcal{S}| = K \geq 2$. There are the following definitions of group fairness.

Definition 1 (Demographic parity [Agarwal et al., 2018])**.** *A predictor $f : \mathcal{X} \to \mathcal{Y}$ achieves demographic parity under a distribution over $(\mathbf{x}, S, y)$ where $y \in \{0, 1\}$ be the data label if its*

prediction $\hat{y} := f(\mathbf{x})$ is statistically independent of the protected attribute S— that is, if $\mathbb{P}[f(\mathbf{x}) = \hat{y} \mid S = s] = \mathbb{P}[f(\mathbf{x}) = \hat{y}]$ for all s and $\hat{y}$.

Definition 2 (Equal opportunity [Gajane, 2017])**.** *Use the same notations of Definition 1. A predictor f achieves equal opportunity under a distribution over $(\mathbf{x}, S, y)$ if its prediction $\hat{y} := f(\mathbf{x})$ is conditionally independent of the protected attribute S given the label $y = 1$— that is, if $\mathbb{P}[\hat{y} = 1 \mid S = s, y = 1] = \mathbb{P}[\hat{y} = 1 \mid y = 1]$ for all s.*

In unsupervised AD, $y = 0$ and $y = 1$ represent normality and anomaly respectively. However, the training data do not include any labeled anomalous samples, which means the predictor f will never be guaranteed to learn sufficient information about the anomaly pattern. In Definition 2 (Equal opportunity), $y = 1$ is presented explicitly and in Definition 1 (Demographic parity), $y = 1$ is presented implicitly because $\mathbb{P}[f(\mathbf{x}) = \hat{y} \mid S = s]$ and $\mathbb{P}[f(\mathbf{x}) = \hat{y}]$ depend on both normal samples ($y = 0$) and anomalous samples ($y = 1$), that is $\mathbb{P}[f(\mathbf{x}) = \hat{y}] = \mathbb{P}[f(\mathbf{x}) = \hat{y} \mid y = 0] \times \mathbb{P}[y = 0] + \mathbb{P}[f(\mathbf{x}) = \hat{y} \mid y = 1] \times \mathbb{P}[y = 1]$. Therefore, we have the following claim (proved in Appendix A.1).

Claim 1. *In unsupervised AD, neither demographic parity nor equal opportunity can be meaningfully guaranteed if without any additional assumptions.*

Note that "meaningfully" emphasizes the trivial solutions $f(\mathbf{x}) \equiv 0$ or 1 for any $\mathbf{x}$ are excluded. In fact, without additional assumptions, any fairness involving anomalous data cannot be guaranteed in unsupervised settings. It is worth noting that Shekhar et al. [2021] considered the equal opportunity in unsupervised AD, where a fairness-unaware AD model (base model) is utilized first to predict pseudo-label $\hat{y}$ and then they used $\hat{y}$ to ensure equal opportunity. Obviously, such a strategy has a significant limitation: the result of equal opportunity depends on the detection performance of the base model, of which fairness cannot be guaranteed. Therefore, for unsupervised AD, without additional assumptions, it is only possible to guarantee the following fairness.

Definition 3 (Predictive equality [Chouldechova, 2017])**.** *Let $\hat{y} := f(\mathbf{x})$. The false positive error rate balance, a.k.a. predictive equality, is defined as*

$$\mathbb{P}[\hat{y} = 1 \mid S = s_i, y = 0] = \mathbb{P}[\hat{y} = 1 \mid S = s_j, y = 0] \tag{1}$$

where $s_i, s_j \in \mathcal{S},\ i \neq j$.

3.2 Fairness Discussion in Unsupervised Anomaly Detection

As evidenced by the preceding analysis, it is intractable to ensure group fairness both on normal and abnormal data in unsupervised anomaly detection (UAD). However, existing works [Zhang and Davidson, 2021, Shekhar et al., 2021, Han et al., 2023] have empirically demonstrated that group fairness (demographic parity or equal opportunity) can be achieved to some extents by introducing adversarial training, adding fairness constraints to optimization objective or generating counterfactual samples for training. This suggests that there must be some natural conditions implicitly contributing to group fairness in unsupervised anomaly detection. In real-world scenarios, anomalous instances are not completely unrelated to normal instances; otherwise, the detection task would be trivial. It is quite common that anomalous samples emerge as perturbed normal samples and the evolution from normality to anomaly is gradual. For instance, in chemical engineering, flow control valves will gradually block, leading to failure; in the mechanical field, bearings will gradually deform, resulting in abnormal vibration signals. Meanwhile, some normal samples are naturally close to anomalous samples. Therefore, it is possible to learn some patterns of anomaly from the normal training data and we make the following assumption.

Assumption 1 (Learnable abnormality)**.** *Suppose $\mathcal{X} = \{\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_n\}$ are independently drawn from $\mathcal{D}_{\mathbf{x}}$. The abnormalities of both $\mathcal{X}$ and unknown anomalous samples can be correctly quantified by a function $\mathcal{T}^* : \mathbb{R}^d \to \mathbb{R}$, and there exists a permutation π on $\mathcal{X}$ such that $0 \leq \mathcal{T}^*(\mathbf{x}_{\pi_1}) < \mathcal{T}^*(\mathbf{x}_{\pi_2}) < \cdots < \mathcal{T}^*(\mathbf{x}_{\pi_n})$, where a larger value of $\mathcal{T}^*(\mathbf{x})$ means that $\mathbf{x}$ is not drawn from the normal distribution $\mathcal{D}_{\mathbf{x}}$ with a higher probability.*

Assumption 1 reflects the fundamental expectation that normal data should be compact in the feature space. Formally, this requires the existence of a bounding function $\mathcal{T}^* : \mathbb{R}^d \to \mathbb{R}$ such that observed normal samples satisfy $\mathcal{T}^*(\mathbf{x}_{\text{normal}}) < \mathcal{T}^*(\mathbf{x}_{\text{abnormal}})$. This assumption ensures the learnability of normal patterns while providing feasibility guarantees for the anomaly detection task. Based on Assumption 1, $\mathcal{T}^*$ servers as an anomaly score function and performs extrapolation on anomalous

samples. Most existing UAD methods [Han et al., 2022, Bouman et al., 2024] have demonstrated that it is possible to learn an approximation of $\mathcal{T}^*$ from $\mathcal{X}$ that can generalize to unseen anomalous samples. However, $\mathcal{T}^*(\mathbf{x})$ may not be independent from a sensitive attribute S and hence can lead to unfairness.

To achieve fairness on $\mathcal{X}$, we split $\mathcal{X}$ into different protected groups $\mathcal{X}_{S=s_i}$ according to the values of a sensitive attribute S. Therefore, a fair AD model $\hat{y} = f(\mathbf{x})$ on $\mathcal{X}$ is to ensure

$$\begin{aligned} &\mathbb{P}[\hat{y} \mid \mathbf{x} \in \mathcal{X}_{S=s_i}, \mathcal{T}^*(\mathbf{x}_{\pi_1}) \leq \tilde{y} \leq \mathcal{T}^*(\mathbf{x}_{\pi_n})] \\ = \ &\mathbb{P}[\hat{y} \mid \mathbf{x} \in \mathcal{X}_{S=s_j}, \mathcal{T}^*(\mathbf{x}_{\pi_1}) \leq \tilde{y} \leq \mathcal{T}^*(\mathbf{x}_{\pi_n})] \end{aligned} \tag{2}$$

where $\tilde{y} := \mathcal{T}^*(\mathbf{x})$ denotes the anomaly level of $\mathbf{x}$. Different from (1) that does not involve any information about the anomaly, (2) is associated with $\mathcal{T}^*$ (established in Assumption 1), meaning that it is possible to learn some information about the anomaly from $\mathcal{X}$. Therefore, it is reasonable to make the following assumptions for fairness on abnormal samples, where $E := \big|\mathbb{P}[\hat{y} \mid \mathbf{x} \in \mathcal{X}_{S=s_i}, \mathcal{T}^*(\mathbf{x}_{\pi_1}) \leq \tilde{y} \leq \mathcal{T}^*(\mathbf{x}_{\pi_n})] - \mathbb{P}[\hat{y} \mid \mathbf{x} \in \mathcal{X}_{S=s_j}, \mathcal{T}^*(\mathbf{x}_{\pi_1}) \leq \tilde{y} \leq \mathcal{T}^*(\mathbf{x}_{\pi_n})]\big|$.

Assumption 2 (Transferable fairness)**.** *Let $\tilde{\mathcal{X}}$ be the set of all unseen anomalous samples and $\tilde{E} = \big|\mathbb{P}[\hat{y} \mid \mathbf{x} \in \tilde{\mathcal{X}}_{S=s_i},\ \tilde{y} > \mathcal{T}^*(\mathbf{x}_{\pi_n})] - \mathbb{P}[\hat{y} \mid \mathbf{x} \in \tilde{\mathcal{X}}_{S=s_j},\ \tilde{y} > \mathcal{T}^*(\mathbf{x}_{\pi_n})]\big|$. There exists a small constant $\mu \geq 1$ such that $\tilde{E} \leq \mu E$.*

Assumption 3 (Generalizable parity)**.** *Let $\hat{\mathcal{X}}$ be the union of the set of all unseen anomalous samples and the set of all unseen normal samples and $\hat{E} = \big|\mathbb{P}[\hat{y} \mid \mathbf{x} \in \hat{\mathcal{X}}_{S=s_i}, 0 \leq \tilde{y}] - \mathbb{P}[\hat{y} \mid \mathbf{x} \in \hat{\mathcal{X}}_{S=s_j}, 0 \leq \tilde{y}]\big|$. There exists a small constant $\tau \geq 1$ such that $\hat{E} \leq \tau E$.*

Assumption 2 ensures that when $\mathcal{T}^*$ is fair on the training data, it is also fair on unseen anomalous data, provided that μ is not too large, where μ depends on real data distribution and the unknown score function $\mathcal{T}^*$. If $\tilde{E} = 0$ for any i, j, the assumption implies equal opportunity. Similarly, Assumption 3 implies demographic parity when $\hat{E} = 0$. Therefore, the evaluation of a fair AD method depends on both the fairness principle (e.g., demographic parity or equal opportunity) and assumption. Indeed, Assumptions 2, 3 have already implicitly verified by existing fairness-aware AD methods [Zhang and Davidson, 2021, Shekhar et al., 2021, Han et al., 2023]. Otherwise, it is only possible to guarantee the predictive equality (1). In our experiments, the results on real-world datasets validate the reasonability of the two assumptions again.

3.3 Model Formulation

3.3.1 Anomaly Detection via Compact Distribution Transformation

For unsupervised anomaly detection, density estimation [Silverman, 2018] is an effective strategy [Zong et al., 2018, Ruff et al., 2018, Fu et al., 2024] to distinguish between normal and abnormal instances. However, the dimensionality of the data is often high and the data distribution in the original space is complex, which makes density estimation challenging.

To solve this problem, we propose to learn a projection $\mathcal{P} : \mathbb{R}^d \to \mathbb{R}^m$ that transforms data distribution $\mathcal{D}_{\mathbf{x}}$ to a known target distribution $\mathcal{D}_{\mathbf{z}}$ that is simple and compact, while there still exists a projection $\mathcal{P}' : \mathbb{R}^m \to \mathbb{R}^d$ that can recover $\mathcal{D}_{\mathbf{x}}$ from $\mathcal{D}_{\mathbf{z}}$ approximately, ensuring that the major information of $\mathbf{x}$ can be preserved by $\mathbf{z}$, where $\mathbf{z} \sim \mathcal{D}_{\mathbf{z}}$. Therefore, we aim to solve

$$\begin{aligned} &\min_{\mathcal{P},\mathcal{P}'} \mathcal{M}(\mathcal{P}(\mathcal{D}_{\mathbf{x}}), \mathcal{D}_{\mathbf{z}}) \\ &\quad \text{s.t. } \mathcal{M}(\mathcal{P}'(\mathcal{P}(\mathcal{D}_{\mathbf{x}})), \mathcal{D}_{\mathbf{x}}) \leq c \end{aligned} \tag{3}$$

where $\mathcal{M}(\cdot, \cdot)$ denotes a distance metric between distributions and c is some positive constant. Instead of the constrained optimization problem (3), we can simply solve the following problem

$$\min_{\mathcal{P},\mathcal{P}'} \mathcal{M}(\mathcal{P}(\mathcal{D}_{\mathbf{x}}), \mathcal{D}_{\mathbf{z}}) + \beta \mathcal{M}(\mathcal{P}'(\mathcal{P}(\mathcal{D}_{\mathbf{x}})), \mathcal{D}_{\mathbf{x}}), \tag{4}$$

where $\beta > 0$ is a trade-off hyperparameter for the two terms. We use two deep neural networks h_ϕ and g_ψ with parameters ϕ, ψ to model $\mathcal{P}$ and $\mathcal{P}'$ respectively. Now, the problem (4) becomes

$$\min_{\phi,\psi} \mathcal{M}(\mathcal{D}_{h_\phi(\mathbf{x})}, \mathcal{D}_{\mathbf{z}}) + \beta \mathcal{M}(\mathcal{D}_{g_\psi(h_\phi(\mathbf{x}))}, \mathcal{D}_{\mathbf{x}}) \tag{5}$$

However, problem (5) is intractable as data distribution $\mathcal{D}_{\mathbf{x}}$ is unknown and $\mathcal{D}_{h_\phi(\mathbf{x})}, \mathcal{D}_{g_\psi(h_\phi(\mathbf{x}))}$ cannot be computed analytically. Thus, we expect to measure the distance between distributions using their finite samples because we can sample from $\mathcal{D}_{\mathbf{x}}$ and $\mathcal{D}_{\mathbf{z}}$ easily. A feasible and popular choice of $\mathcal{M}(\cdot,\cdot)$ is the Sinkhorn distance [Cuturi, 2013] between two distributions supported by $\mathcal{X} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_{n_1}\}$ and $\mathcal{Y} = \{\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_{n_2}\}$:

$$\begin{aligned} \text{Sinkhorn}(\mathcal{X}, \mathcal{Y}) := \min \langle \mathbf{P}, \mathbf{C} \rangle_F + \alpha \sum_{i,j} \mathbf{P}_{ij} \log(\mathbf{P}_{ij}), \\ \text{s.t. } \mathbf{P}\mathbf{1} = \mathbf{a}, \mathbf{P}^T\mathbf{1} = \mathbf{b}, \mathbf{P} \geq 0 \end{aligned} \tag{6}$$

where $\mathbf{P} \in \mathbb{R}^{n_1 \times n_2}$ denotes the transport plan and $\mathbf{C} \in \mathbb{R}^{n_1 \times n_2}$ is a metric cost matrix between $\mathcal{X}$ and $\mathcal{Y}$. The two probability vectors $\mathbf{a}$ and $\mathbf{b}$ satisfy $\mathbf{a}^T\mathbf{1} = 1, \mathbf{b}^T\mathbf{1} = 1, \mathbf{1} = [1, 1, \cdots, 1]^T$. Other measures, such as the maximum mean discrepancy, can also be used, which is discussed in the Appendix F. Now, we use Sinkhorn distance to replace the first term in problem (5), use reconstruction error to replace the second term in problem (5), and get the following optimization objective

$$\min_{\phi,\psi} \text{Sinkhorn}(h_\phi(\mathcal{X}), \mathcal{Z}) + \frac{\beta}{n} \sum_{i=1}^{n} \|\mathbf{x}_i - g_\psi(h_\phi(\mathbf{x}_i))\|^2 \tag{7}$$

where $\mathbf{x}_i \in \mathcal{X}$, $\mathcal{Z} = \{\mathbf{z}_i : \mathbf{z}_i \sim \mathcal{D}_z, i = 1, \ldots, n\}$, and β is a trade-off hyperparameter.

We may replace both the first and second terms in problem (5) using Sinkhorn distance, however, which easily leads to a higher computational cost. Therefore, we use reconstruction error to replace the second term in problem (5). The reconstruction error term is a stronger constraint than $\mathcal{P}'$, but it is efficient and effective to preserve the core information from the original data distribution $\mathcal{D}_{\mathbf{x}}$.

Target distribution The target distribution should be simple and compact. The compactness ensures that projected normal samples in the decision space lie in high-density regions, which contributes to a reliable decision boundary. The simplicity ensures that sampling from the target distribution is easy. Therefore, for the target distribution $\mathcal{D}_{\mathbf{z}}$, we propose to use a *truncated isotropic Gaussian* based on $\mathcal{N}(\mathbf{0}, \mathbf{I}_d)$ where truncation is to ensure that the target distribution is sufficiently compact for normal data. Based on the target distribution, we can define a score function naturally depending on density estimation.

Anomaly score Let h_{ϕ^*} be the trained fairness-aware model. In inference phase, for a new sample $\mathbf{x}_{\text{new}}$, we define a soft anomaly score ζ_{soft} (or a hard score depending on a threshold)

$$\text{Score}(\mathbf{x}_{\text{new}}) = \zeta_{\text{soft}}(h_{\phi^*}(\mathbf{x}_{\text{new}})) = \|h_{\phi^*}(\mathbf{x}_{\text{new}})\|, \tag{8}$$

which can measure the anomaly degree of $\mathbf{x}_{\text{new}}$. By considering a threshold (with a certain significance level) obtained from the training data scores, we get a hard score function ζ_{hard} with a binary output in $\{0, 1\}$, indicating whether $\mathbf{x}_{\text{new}}$ is normal or not. Actually, the density of $\mathbf{x}_{\text{new}}$ can be estimated as $(2\pi)^{-d/2} \exp\left(-\frac{1}{2}(\text{Score}(\mathbf{x}_{\text{new}}))^2\right)$.

3.3.2 Fairness via Shared Target Distribution

Directly finding a fair [3] f while maintaining strong detection ability is non-trivial. A naive strategy involves combining FRL techniques with UAD methods to construct two-stage pipelines, in which the optimization objectives in the FRL stage are not tailored for the detection task. As a result, Such pipelines often yield poor detection performance. Similarly, directly incorporating fairness constraints into the optimization objectives of UAD methods also degrades detection performance. To avoid such problems, we expect to find an end-to-end and coordinated learning process for the detection task and group fairness. Based on the framework established in Section 3.3.1, we can obtain a projection h_{ϕ^*} via compact distributional transformation, and then the detection task can be conducted effectively by ζ estimating the density of data in decision space. Thus, we obtain a detector $f = \zeta \circ h$ and $\hat{y} = f(\mathbf{x}) = \zeta \circ h(\mathbf{x})$. Notably, if h is fair for a protected variable S, we have

$$\begin{aligned} &\mathbb{P}[h(\mathbf{x}) \mid \mathbf{x} \in \mathcal{X}_{S=s_i}, \mathcal{T}^*(\mathbf{x}_{\pi_1}) \leq \tilde{y} \leq \mathcal{T}^*(\mathbf{x}_{\pi_n})] \\ = \; &\mathbb{P}[h(\mathbf{x}) \mid \mathbf{x} \in \mathcal{X}_{S=s_j}, \mathcal{T}^*(\mathbf{x}_{\pi_1}) \leq \tilde{y} \leq \mathcal{T}^*(\mathbf{x}_{\pi_n})]. \end{aligned} \tag{9}$$

[3] Unless specified otherwise, all “fair" and “fairness" in this paper pertain to group fairness.

where $s_i, s_i \in S$ denote the attribute values of protected variable S.

It is easy to show that (proved in Appendix A.2)

Proposition 1. *For any ζ, if* (9) *is attained, then* (2) *holds.*

To obtain a fair h without introducing additional fairness constraints, we propose to map data across different demographic groups into a common target distribution $\mathcal{D}_{\mathbf{z}}$. Naturally, the (3) becomes

$$\begin{aligned} \min_{\mathcal{P},\mathcal{P}'} \sum_{s\in S} & \mathcal{M}(\mathcal{P}(\mathcal{D}_{\mathcal{X}_{S=s}}), \mathcal{D}_{\mathbf{z}}) \\ \text{s.t. } & \mathcal{M}(\mathcal{P}'(\mathcal{P}(\mathcal{D}_{\mathbf{x}})), \mathcal{D}_{\mathbf{x}}) \leq c \end{aligned} \tag{10}$$

It further follows that (proved in Appendix A.3)

Proposition 2. *If $\sum_{s\in S} \mathcal{M}(\mathcal{P}(\mathcal{D}_{\mathcal{X}_{S=s}}), \mathcal{D}_{\mathbf{z}}) = 0$, then* (9) *is attained.*

Therefore, combing Proposition 1 and Proposition 2, we conclude that solving problem (10) makes the decision function f as fair as possible on the training data, in terms of predictive equality. Further, using Assumption 2 or 3, we may obtain equal opportunity or demographic parity.

Based on the above analysis, we obtain a projection h via compact distribution transformation (common target distribution for different demographic groups), where h can satisfy different fairness principles building on different assumptions, and meanwhile yield a high-detection-accuracy detector $f = \zeta \circ h$ based on a compact target distribution. Building on the framework in Section 3.3.1, we finally solve

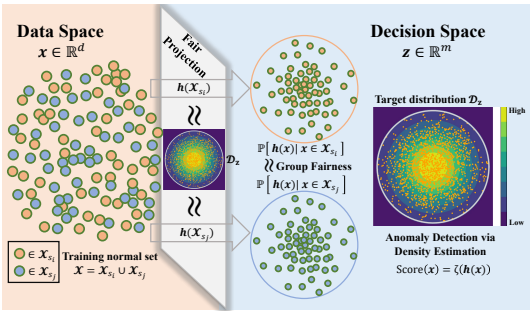

Figure 1: The illustration of Im-FairAD. For simplicity, we only visualize two different attribute values $s_i, s_j \in S$ for the protected variable S, but Im-FairAD does not impose such a restriction. The 'High' and 'Low' denote the relative density in the target distribution.

$$\min_{\phi,\psi} \sum_{s\in S} \text{Sinkhorn}(h_\phi(\mathcal{X}), \mathcal{Z}) + \frac{\beta}{n}\sum_{i=1}^{n} \|\mathbf{x}_i - g_\psi(h_\phi(\mathbf{x}_i))\|^2. \tag{11}$$

Notably, our approach achieves fairness across demographic groups of protected variables without introducing additional fairness constraints, thereby fulfilling the dual objectives of detection efficacy and group fairness. We call the method **Im-FairAD**. Figure 1 provides an illustration.

Variant of FairAD Besides the method presented by (11), we also explore another feasible optimization objective by directly minimizing the distribution distance of anomaly scores across different groups, i.e.,

$$\begin{aligned} \min_{\phi,\psi} \ & \text{Sinkhorn}(h_\phi(\mathcal{X}), \mathcal{Z}) + \frac{\beta}{n}\sum_{i=1}^{n} \|\mathbf{x}_i - g_\psi(h_\phi(\mathbf{x}_i))\|^2 \\ & + \lambda \sum_{i\neq j} \text{Sinkhorn}(\zeta(h_\phi(\mathcal{X}_{S=s_i})), \zeta(h_\phi(\mathcal{X}_{S=s_j}))). \end{aligned} \tag{12}$$

The first and second terms focus on detection accuracy and the third term ensures fairness between protected groups. We call the method **Ex-FairAD**. It is worth noting that the first term in (12), mapping data distribution into target distribution, is necessary for Ex-FairAD because it determines whether the score function (8) is feasible for Ex-FairAD. More detailed discussion and an ablation study are provided in Appendix E.

3.4 Threshold-Free Fairness Metrics

To overcome the limitations (See Appendix B.3) of *fairness ratio*, in this paper, we propose a new fairness metric called *Average Demographic Parity Difference* (ADPD):

$$\text{ADPD} := \frac{1}{n}\sum_{k=1}^{n} \Big| \mathbb{P}(\text{Score}(\mathcal{X}) > t_k | S = s_i) - \mathbb{P}(\text{Score}(\mathcal{X}) > t_k | S = s_j) \Big| \tag{13}$$

where $t_k \in \text{Score}(\mathcal{X})$ denotes the anomaly score of single sample. In our proposed methods, $t_k = \|h_{\phi^*}(\mathbf{x}_k)\|$. ADPD is a threshold-free metric (such as AUC) for measuring demographic parity. The range of ADPD is $[0, 1)$ and a smaller ADPD means a higher fairness. Although we introduce a novel threshold-free metric for fairness measure, this is not to imply that the threshold-dependent metrics are useless. More discussion on threshold-free and threshold-dependent metrics is provided in Appendix B.3. We both use threshold-free and threshold-dependent metrics to evaluate all methods in our experiments.

4 Experiments

Our experiments are conducted on six publicly available datasets from the literature on fairness in machine learning [Zhang and Davidson, 2021, Chai and Wang, 2022, Han et al., 2023, Chen et al., 2024], where there are different kinds of sensitive information. The more detailed statistics of datasets and data splitting are provided in Appendix B.1. We both use standard (fairness-unaware) UAD methods, two-stage pipelines (FRL technique + UAD methods), and end-to-end fairness-aware UAD methods as baselines. The detailed experimental settings are provided in Appendix B.2. We design experiments to answer the following questions.

- **[Q1]** Can our proposed methods achieve better detection accuracy when maintaining comparable or even better fairness?
- **[Q2]** How are the performances (including detection accuracy and fairness) of all compared methods on balanced and skewed splitting?
- **[Q3]** Can fairness-aware unsupervised AD methods achieve fairness on anomalous data in real-world scenarios, although anomalous samples are not used during the training?

4.1 Experimental Results

To answer **[Q1]**, we visualize the trade-off between AUC and ADPD under balanced splitting (same sample size across different demographic groups) in Figure 2, where the red star in the upper left corner denotes the ideal fairness-aware anomaly detection. Compared with all the baselines, our method achieves comparable or even better trade-off between detection accuracy and fairness in almost all cases. The ADPD in Figure 2 is computed on all test data, including normal and abnormal samples. The detailed numerical results and trade-off visualizations under skewed splitting (varying sample size across different demographic groups) are provided in Appendix D.1.

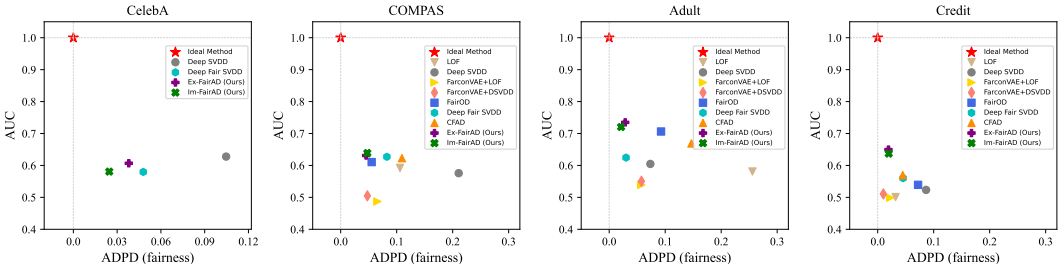

Figure 2: Accuracy-fairness trade-off on balanced splitting. Note that the baselines, FairOD and CFAD are tailored to tabular data.

Further, the empirical results in Table 5 reveal a critical trend: when transitioning from balanced to skewed splitting, fairness-aware AD methods exhibit degradation in group fairness, as evidenced by significantly larger ADPD values in most cases. This suggests that fairness of existing fairness-aware AD methods is easily affected by demographic imbalances and skewed sample proportions across groups amplify fairness violations. By contrast, our proposed methods exhibit a robust fairness preservation (slight fluctuation on ADPD) on balanced and skewed data splitting **[Q2]**.

Fairness of Im-FairAD under balanced and skewed splitting When problem (6) is solved well that is $\sum_{s\in S} \mathcal{M}(h(\mathcal{D}_{\mathcal{X}_{S=s}}), \mathcal{D}_{\mathbf{z}})$ near to zero, whatever under balanced or skewed data splitting, Im-FairAD can obtain fair $h(\mathcal{X}_{S=s})$ for different groups s. Therefore, Im-FairAD can guarantee (9) according to Proposition 2 for both balanced and skewed data splitting **[Q2]**.

Fairness of Im-FairAD on abnormal data

As there are no abnormal samples in the training set, the optimization processes of unsupervised fairness-aware AD methods do not exploit any information directly from the abnormal data, and hence may not consider the fairness in abnormal data explicitly. To answer **[Q3]**, we visualize the ADPD of abnormal data from the test set in Figure 3. More results on Titanic and SP are provided in Appendix D.1. Compared to all baselines, our proposed methods achieve better or comparable group fairness in most cases. On the other hand, Deep Fair SVDD achieves better group fairness than Deep SVDD (fairness-unaware) on abnormal data of all four datasets. Two-stage pipelines (FarconVAE+LOF and FarconVAE+Deep SVDD) also achieve better group fairness than LOF and Deep SVDD (fairness-unaware) on abnormal data. On COMPAS and Credit, FairOD and CFAD both achieve better group fairness than Deep SVDD(fairness-unaware) on abnormal data. One possible reason for the success of these methods especially ours is that the abnormal data may have some similar latent structure as the normal data, or at least there exists a mapping (not too complex)[4] between the normal data distribution $\mathcal{D}_{\mathbf{x}}$ and the abnormal data distribution $\mathcal{D}_{\tilde{\mathbf{x}}}$, i.e., $\mathcal{D}_{\tilde{\mathbf{x}}} = \mathcal{Q}(\mathcal{D}_{\mathbf{x}})$. Thus, the AD methods can ensure fairness on abnormal data indirectly, where the intermediary is the normal data, which also indicates that the **Assumption** 2 is reasonable and can be accessed in real-world scenarios. Particularly, our methods are based on distribution transformation and the fairness of normal data can be transformed to abnormal data via the composition $\mathcal{P} \circ \mathcal{Q}$. That is also why our methods are more effective in terms of fairness than other methods on abnormal data.

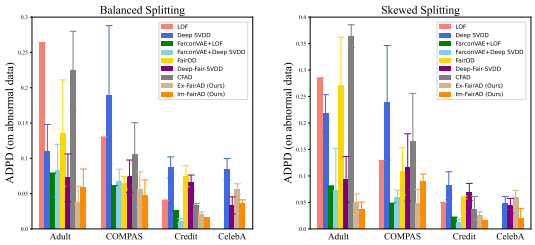

Figure 3: Fairness of all baselines on abnormal data from the test set.

4.2 Evaluation by Threshold-Dependent Metric

We also use *fairness ratio* (21) from previous works [Zhang and Davidson, 2021, Shekhar et al., 2021] and F1-score to evaluate fairness and detection accuracy. For calculating *fairness ratio* and F1-score, a threshold needs to be determined. We sort the anomaly scores of the training set in ascending order and set the threshold to pN-th smallest anomaly score, where we set $p = \{0.90, 0.95\}$ and N denotes the size of the training set. The results on COMPAS, Adult and Credit with $p = 0.90$ are reported in Table 1. More results are reported in Appendix D.2. Observing the Table 1, our methods still achieve better detection accuracy while maintaining a better or comparable *fairness ratio* in almost all cases **[Q1]**. Notably, the *fairness ratio* is highly sensitive to different thresholds (See all results with $p = \{0.90, 0.95\}$ in Appendix D.2). Consequently, a threshold-free fairness metric is essential for evaluating the performance from a holistic view.

Table 1: F1-score and *fairness ratio* on COMPAS, Adult and Credit with $p = 0.90$. The best two results are marked in **bold**.

Methods	COMPAS			Adult			Credit		
	F1(%) ↑	*fairness ratio* ↑		F1(%) ↑	*fairness ratio* ↑		F1(%) ↑	*fairness ratio* ↑	
		normal	all		normal	all		normal	all
LOF	29.85(0.00)	0.27(0.00)	0.35(0.00)	27.45(0.00)	0.34(0.00)	0.42(0.00)	20.11(0.00)	0.45(0.00)	0.55(0.00)
Deep SVDD	31.30(5.17)	0.49(0.18)	0.58(0.18)	44.73(4.72)	0.54(0.24)	0.63(0.11)	23.31(7.75)	0.76(0.12)	0.74(0.20)
FarconVAE+LOF	14.28(0.00)	0.69(0.00)	**0.77**(0.00)	22.87(0.00)	0.63(0.00)	0.62(0.00)	15.06(0.00)	**0.88**(0.00)	0.88(0.00)
FarconVAE+Deep SVDD	19.50(0.89)	**0.74**(0.07)	0.70(0.14)	22.08(8.87)	0.59(0.10)	0.73(0.18)	16.96(1.80)	**0.94**(0.02)	**0.95**(0.02)
FairOD	17.81(0.11)	0.70(0.04)	**0.79**(0.01)	36.87(3.99)	0.70(0.06)	0.32(0.04)	41.00(0.81)	0.70(0.04)	0.74(0.02)
Deep Fair SVDD	42.23(0.81)	0.46(0.04)	0.58(0.04)	45.03(1.56)	**0.86**(0.01)	0.83(0.05)	13.54(0.75)	0.81(0.06)	0.73(0.09)
CFAD	40.67(3.67)	0.39(0.13)	0.49(0.08)	48.01(5.51)	0.76(0.08)	0.55(0.14)	20.43(0.75)	0.69(0.05)	0.69(0.04)
Ex-FairAD (Ours)	**42.30**(5.88)	**0.71**(0.18)	0.74(0.15)	**52.75**(2.19)	**0.87**(0.11)	**0.92**(0.03)	**53.70**(2.53)	0.79(0.03)	**0.90**(0.04)
Im-FairAD (Ours)	**47.49**(4.32)	0.65(0.28)	0.68(0.09)	**56.04**(4.70)	0.72(0.06)	**0.84**(0.07)	**54.32**(3.43)	0.81(0.08)	**0.90**(0.04)

4.3 More Experimental Results and Analysis

Due to space limitations, the appendices contain more results and further experimental investigation. Appendix C: An extension of the proposed method to multiple protected attributes; Appendix D: More numerical results and visualization; Appendix E: Ablation study on the proposed methods (Im-FairD and Ex-FairAD); Appendix F: The selection of distance metric between distributions; Appendix G: Time complexity and implementation cost analysis for proposed method; Appendix H: The effectiveness analysis of the proposed methods on equal opportunity; Appendix I: Experimental

[4]This assumption is realistic because in many real scenarios, abnormality origins from normality.

investigation on contaminated training set (including unknown abnormal samples); Appendix J: Experimental investigation on text data.

5 Conclusion

In this paper, we focused on the group fairness of unsupervised anomaly detection, clearly discussing the necessary conditions of achieving group fairness, and proposed Im-FairAD and Ex-FairAD, two novel fairness-aware anomaly detection methods. Considering the limitations of existing fairness evaluation metrics used in previous works, we propose a novel threshold-free metrics ADPD, which provides a holistic view for evaluating the fairness of methods.

Empirical results on real-world datasets indicated that the proposed two methods achieve a better trade-off between detection accuracy and fairness than baselines. In most cases, Im-FairAD has better performance than Ex-FairAD. Moreover, we analyzed the reason why fairness can be ensured on the abnormal samples of the test set, although the model training process does not utilize any anomalous samples.

Acknowledgements

This work was supported by the General Program of Natural Science Foundation of Guangdong Province under Grant No.2024A1515011771, the National Natural Science Foundation of China under Grant No.62376236, the Guangdong Provincial Key Laboratory of Mathematical Foundations for Artificial Intelligence (2023B1212010001), Shenzhen Science and Technology Program ZDSYS20230626091302006, and Shenzhen Stability Science Program 2023.

References

Alekh Agarwal, Alina Beygelzimer, Miroslav Dudík, John Langford, and Hanna Wallach. A reductions approach to fair classification. In *International conference on machine learning*, pages 60–69. PMLR, 2018.

Alekh Agarwal, Miroslav Dudík, and Zhiwei Steven Wu. Fair regression: Quantitative definitions and reduction-based algorithms. In *International Conference on Machine Learning*, pages 120–129. PMLR, 2019.

Barry Becker and Ronny Kohavi. Adult. UCI Machine Learning Repository, 1996. DOI: https://doi.org/10.24432/C5XW20.

Roel Bouman, Zaharah Bukhsh, and Tom Heskes. Unsupervised anomaly detection algorithms on real-world data: how many do we need? *Journal of Machine Learning Research*, 25(105):1–34, 2024.

Markus M. Breunig, Hans-Peter Kriegel, Raymond T. Ng, and Jörg Sander. Lof: identifying density-based local outliers. *ACM SIGMOD Record*, page 93–104, Jun 2000. doi: 10.1145/335191.335388. URL `http://dx.doi.org/10.1145/335191.335388`.

Maarten Buyl and Tijl De Bie. Optimal transport of classifiers to fairness. *Advances in Neural Information Processing Systems*, 35:33728–33740, 2022.

Jinyu Cai and Jicong Fan. Perturbation learning based anomaly detection. *Advances in Neural Information Processing Systems*, 35, 2022.

Mattia Cerrato, Marius Köppel, Philipp Wolf, and Stefan Kramer. 10 years of fair representations: Challenges and opportunities. *arXiv preprint arXiv:2407.03834*, 2024.

Junyi Chai and Xiaoqian Wang. Self-supervised fair representation learning without demographics. *Advances in Neural Information Processing Systems*, 35:27100–27113, 2022.

Wenjing Chang, Kay Liu, Philip S Yu, and Jianjun Yu. Enhancing fairness in unsupervised graph anomaly detection through disentanglement. *arXiv preprint arXiv:2406.00987*, 2024.

Zhenpeng Chen, Jie M Zhang, Max Hort, Mark Harman, and Federica Sarro. Fairness testing: A comprehensive survey and analysis of trends. *ACM Transactions on Software Engineering and Methodology*, 33(5):1–59, 2024.

Alexandra Chouldechova. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big data*, 5(2):153–163, 2017.

Paulo Cortez. Student Performance. UCI Machine Learning Repository, 2008. DOI: https://doi.org/10.24432/C5TG7T.

Elliot Creager, David Madras, Jörn-Henrik Jacobsen, Marissa Weis, Kevin Swersky, Toniann Pitassi, and Richard Zemel. Flexibly fair representation learning by disentanglement. In *International conference on machine learning*, pages 1436–1445. PMLR, 2019.

Will Cukierski. Titanic - machine learning from disaster. `https://kaggle.com/competitions/titanic`, 2012. Kaggle.

Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. *Advances in neural information processing systems*, 26, 2013.

Jeffrey Dastin. Amazon scraps secret ai recruiting tool that showed bias against women. In *Ethics of data and analytics*, pages 296–299. Auerbach Publications, 2022.

Ian Davidson and Selvan Suntiha Ravi. A framework for determining the fairness of outlier detection. In *ECAI 2020*, pages 2465–2472. IOS Press, 2020.

P Deepak and Savitha Sam Abraham. Fair outlier detection. In *21th International Conference on Web Information Systems Engineering: WISE 2020*, pages 447–462, 2020.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pages 4171–4186, 2019.

Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. Fairness through awareness. In *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference*, Jan 2012a. doi: 10.1145/2090236.2090255. URL `http://dx.doi.org/10.1145/2090236.2090255`.

Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. Fairness through awareness. In *Proceedings of the 3rd innovations in theoretical computer science conference*, pages 214–226, 2012b.

Dazhi Fu, Zhao Zhang, and Jicong Fan. Dense projection for anomaly detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 8398–8408, 2024.

Pratik Gajane. On formalizing fairness in prediction with machine learning. *arXiv: Learning,arXiv: Learning*, Oct 2017.

Sachin Goyal, Aditi Raghunathan, Moksh Jain, Harsha Vardhan Simhadri, and Prateek Jain. Drocc: Deep robust one-class classification. In *Proceedings of the International Conference on Machine Learning*, pages 3711–3721. PMLR, 2020.

Arthur Gretton, Karsten M Borgwardt, Malte J Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *The Journal of Machine Learning Research*, 13(1):723–773, 2012.

Sara Hajian, Francesco Bonchi, and Carlos Castillo. Algorithmic bias: From discrimination discovery to fairness-aware data mining. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Aug 2016. doi: 10.1145/2939672.2945386. URL `http://dx.doi.org/10.1145/2939672.2945386`.

Songqiao Han, Xiyang Hu, Hailiang Huang, Minqi Jiang, and Yue Zhao. Adbench: Anomaly detection benchmark. *Advances in neural information processing systems*, 35:32142–32159, 2022.

Xiao Han, Lu Zhang, Yongkai Wu, and Shuhan Yuan. Achieving counterfactual fairness for anomaly detection. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pages 55–66. Springer, 2023.

Moritz Hardt, Eric Price, and Nathan Srebro. Equality of opportunity in supervised learning. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, NIPS'16, page 3323–3331, Red Hook, NY, USA, 2016. Curran Associates Inc. ISBN 9781510838819.

Lisa Anne Hendricks, Kaylee Burns, Kate Saenko, Trevor Darrell, and Anna Rohrbach. Women also snowboard: Overcoming bias in captioning models. In *Proceedings of the European conference on computer vision (ECCV)*, pages 771–787, 2018.

Geoffrey E Hinton and Ruslan R Salakhutdinov. Reducing the dimensionality of data with neural networks. *Science*, 313(5786):504–507, 2006.

Nikola Jovanović, Mislav Balunovic, Dimitar Iliev Dimitrov, and Martin Vechev. Fare: Provably fair representation learning with practical certificates. In *International Conference on Machine Learning*, pages 15401–15420. PMLR, 2023.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *Proceedings of the International Conference on Learning Representations*, 2015.

Preethi Lahoti, Alex Beutel, Jilin Chen, Kang Lee, Flavien Prost, Nithum Thain, Xuezhi Wang, and Ed Chi. Fairness without demographics through adversarially reweighted learning. *Advances in neural information processing systems*, 33:728–740, 2020.

Jeff Larson, Surya Mattu, Lauren Kirchner, and Julia Angwin. How we analyzed the compas recidivism algorithm. *ProPublica (5 2016)*, 9(1):3–3, 2016.

Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of International Conference on Computer Vision (ICCV)*, December 2015.

Christos Louizos, Kevin Swersky, Yujia Li, Max Welling, and Richard Zemel. The variational fair autoencoder. *arXiv preprint arXiv:1511.00830*, 2015.

David Madras, Elliot Creager, Toniann Pitassi, and Richard Zemel. Learning adversarially fair and transferable representations. In *International Conference on Machine Learning*, pages 3384–3393. PMLR, 2018.

Felix Meissen, Svenja Breuer, Moritz Knolle, Alena Buyx, Ruth Müller, Georgios Kaissis, Benedikt Wiestler, and Daniel Rückert. (predictable) performance bias in unsupervised anomaly detection. *arXiv preprint arXiv:2309.14198*, 2023.

Daniel Moyer, Shuyang Gao, Rob Brekelmans, Aram Galstyan, and Greg Ver Steeg. Invariant representations without adversarial training. *Advances in neural information processing systems*, 31, 2018.

Neng Kai Nigel Neo, Yeon-Chang Lee, Yiqiao Jin, Sang-Wook Kim, and Srijan Kumar. Towards fair graph anomaly detection: problem, benchmark datasets, and evaluation. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*, pages 1752–1762, 2024.

Changdae Oh, Heeji Won, Junhyuk So, Taero Kim, Yewon Kim, Hosik Choi, and Kyungwoo Song. Learning fair representation via distributional contrastive disentanglement. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 1295–1305, 2022.

Guansong Pang, Chunhua Shen, Longbing Cao, and Anton Van Den Hengel. Deep learning for anomaly detection: A review. *ACM computing surveys (CSUR)*, 54(2):1–38, 2021.

Alfréd Rényi. On measures of entropy and information. In *Proceedings of the fourth Berkeley symposium on mathematical statistics and probability, volume 1: contributions to the theory of statistics*, volume 4, pages 547–562. University of California Press, 1961.

Lukas Ruff, RobertA. Vandermeulen, Nico Goernitz, Lucas Deecke, ShoaibAhmed Siddiqui, Alexander Binder, Emmanuel Müller, and Marius Kloft. Deep one-class classification. *International Conference on Machine Learning,International Conference on Machine Learning*, Jul 2018.

Shubhranshu Shekhar, Neil Shah, and Leman Akoglu. Fairod: Fairness-aware outlier detection. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, Jul 2021. doi: 10.1145/3461702.3462517. URL http://dx.doi.org/10.1145/3461702.3462517.

Bernard W Silverman. *Density estimation for statistics and data analysis*. Routledge, 2018.

Eric Tzeng, Judy Hoffman, Ning Zhang, Kate Saenko, and Trevor Darrell. Deep domain confusion: Maximizing for domain invariance. *arXiv preprint arXiv:1412.3474*, 2014.

Ziwei Wu, Lecheng Zheng, Yuancheng Yu, Ruizhong Qiu, John Birge, and Jingrui He. Fair anomaly detection for imbalanced groups. *arXiv preprint arXiv:2409.10951*, 2024.

Feng Xiao, Jianfeng Zhou, Kunpeng Han, Haoyuan Hu, and Jicong Fan. Unsupervised anomaly detection using inverse generative adversarial networks. *Information Sciences*, 689:121435, 2025.

Qizhe Xie, Zihang Dai, Yulun Du, Eduard Hovy, and Graham Neubig. Controllable invariance through adversarial feature learning. *Advances in neural information processing systems*, 30, 2017.

I-Cheng Yeh. Default of Credit Card Clients. UCI Machine Learning Repository, 2009. DOI: https://doi.org/10.24432/C55S3H.

Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rodriguez, and Krishna P. Gummadi. Fairness beyond disparate treatment & disparate impact: Learning classification without disparate mistreatment. In *Proceedings of the 26th International Conference on World Wide Web*, Apr 2017a. doi: 10.1145/3038912.3052660. URL http://dx.doi.org/10.1145/3038912.3052660.

Muhammad Bilal Zafar, Isabel Valera, Manuel Rodriguez, Krishna Gummadi, and Adrian Weller. From parity to preference-based notions of fairness in classification. *Advances in neural information processing systems*, 30, 2017b.

Rich Zemel, Yu Wu, Kevin Swersky, Toni Pitassi, and Cynthia Dwork. Learning fair representations. In *International conference on machine learning*, pages 325–333. PMLR, 2013.

Brian Hu Zhang, Blake Lemoine, and Margaret Mitchell. Mitigating unwanted biases with adversarial learning. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, pages 335–340, 2018.

Hongjing Zhang and Ian Davidson. Towards fair deep anomaly detection. In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, pages 138–148, 2021.

Yue Zhao, Zain Nasrullah, and Zheng Li. Pyod: A python toolbox for scalable outlier detection. *Journal of Machine Learning Research*, 20(96):1–7, 2019. URL http://jmlr.org/papers/v20/19-011.html.

Bo Zong, Qi Song, Martin Renqiang Min, Wei Cheng, Cristian Lumezanu, Daeki Cho, and Haifeng Chen. Deep autoencoding gaussian mixture model for unsupervised anomaly detection. In *International Conference on Learning Representations*, 2018. URL https://openreview.net/forum?id=BJJLHbb0-.

NeurIPS Paper Checklist

1. **Claims**
   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?
   Answer: [Yes]
   Justification: The main contributions are concluded in Introduction.
   Guidelines:
   - The answer NA means that the abstract and introduction do not include the claims made in the paper.

- The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
- The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
- It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**
   Question: Does the paper discuss the limitations of the work performed by the authors?
   Answer: [Yes]
   Justification: The limitation discussion are provided in Appendix K.
   Guidelines:
   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.
3. **Theory assumptions and proofs**
   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?
   Answer: [Yes]
   Justification: The proofs of claim and propositions introduced in this work are provided in Appendix A.1, A.2, A.3.
   Guidelines:
   - The answer NA means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.
4. **Experimental result reproducibility**
   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]
Justification: The specific experimental settings are provided in Appendix B.2.
Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**
Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?
Answer: [Yes]
Justification: All the data used in experiments are publicly available and we provide the specific download URL in Appendix B.1. Our source code is provided in supplementary materials.
Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.

- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**
   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?
   Answer: [Yes]
   Justification: The detailed experimental settings are provided in Appendix B.2.
   Guidelines:
   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.
7. **Experiment statistical significance**
   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?
   Answer: [Yes]
   Justification: We report the standard variance for main results.
   Guidelines:
   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.
   - It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
   - For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
   - If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.
8. **Experiments compute resources**
   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?
   Answer: [Yes]
   Justification: We describe the hardware and software source used in our experiments in Appendix B.2.
   Guidelines:
   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).
9. **Code of ethics**

Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?
Answer: [Yes]
Justification:
Guidelines:
- The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
- If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**
Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?
Answer: [Yes]
Justification: Our work aims to improve fairness of unsupervised anomaly detection system, which is helpful to mitigate the social bias for different demographic groups to some extends.
Guidelines:
- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**
Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?
Answer: [NA]
Justification:
Guidelines:
- The answer NA means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?
Answer: [Yes]
Justification: We include the original URL links for all used data and software tool.
Guidelines:
- The answer NA means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**
Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?
Answer: [NA]
Justification:
Guidelines:
- The answer NA means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**
Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?
Answer: [NA]
Justification:
Guidelines:
- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**
Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?
Answer: [NA]
Justification:
Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**
Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.
Answer: [NA]
Justification:
Guidelines:
    - The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.

A Proof for Claim and Propositions

A.1 Proof for Claim 1

Proof. According to **Definition** 1, to achieve demographic parity, we need to guarantee $\mathbb{P}[f(\mathbf{x}) = \hat{y} \mid S = s] = \mathbb{P}[f(\mathbf{x}) = \hat{y}]$ for all s and $\hat{y}$. Therefore, we have

$$\begin{aligned}\mathbb{P}[f(\mathbf{x}) = \hat{y}] = &\ \mathbb{P}[f(\mathbf{x}) = \hat{y} \mid y = 0] \times \mathbb{P}[y = 0] \\ &+ \mathbb{P}[f(\mathbf{x}) = \hat{y} \mid y = 1] \times \mathbb{P}[y = 1]\end{aligned}$$

However, only normal data ($y = 0$) can be accessed in unsupervised anomaly detection, which means that we can only obtain $\mathbb{P}[f(\mathbf{x}) = \hat{y} \mid y = 0] \times \mathbb{P}[y = 0]$ in such scenario. Therefore, demographic parity cannot be guaranteed in unsupervised anomaly detection.

And according to **Definition** 2, for achieving equal opportunity, we need to guarantee $\mathbb{P}[\hat{y} = 1 \mid S = s, y = 1] = \mathbb{P}[\hat{y} = 1 \mid y = 1]$ for all s. Obviously, we cannot obtain $\mathbb{P}[\hat{y} = 1 \mid S = s, y = 1]$ and $\mathbb{P}[\hat{y} = 1 \mid y = 1]$ in unsupervised anomaly detection if without any additional assumptions. Therefore, equal opportunity cannot be guaranteed in unsupervised anomaly detection. □

A.2 Proof for Proposition 1

We split $\mathcal{X}$ into different protected groups $\mathcal{X}_{S=s_i}$ according to the values of a sensitive attribute S. Therefore, a fair learning on $\mathcal{X}$ is to ensure

$$\begin{aligned}&\mathbb{P}[\hat{y} \mid \mathbf{x} \in \mathcal{X}_{S=s_i}, \mathcal{T}^*(\mathbf{x}_{\pi_1}) \leq \tilde{y} \leq \mathcal{T}^*(\mathbf{x}_{\pi_n})] \\ =\ &\mathbb{P}[\hat{y} \mid \mathbf{x} \in \mathcal{X}_{S=s_j}, \mathcal{T}^*(\mathbf{x}_{\pi_1}) \leq \tilde{y} \leq \mathcal{T}^*(\mathbf{x}_{\pi_n})]\end{aligned} \tag{14}$$

we expect to find a h that map data from different demographic groups into a common target distribution where an effective anomaly score function ζ exist naturally.

Therefore, we obtain a detector $f = \zeta \circ h$ and $\hat{y} = f(\mathbf{x}) = \zeta \circ h(\mathbf{x})$. Indeed, if h is fair for S, we have

$$\begin{aligned}&\mathbb{P}[h(\mathbf{x}) \mid \mathbf{x} \in \mathcal{X}_{S=s_i}, \mathcal{T}^*(\mathbf{x}_{\pi_1}) \leq \tilde{y} \leq \mathcal{T}^*(\mathbf{x}_{\pi_n})] \\ =\ &\mathbb{P}[h(\mathbf{x}) \mid \mathbf{x} \in \mathcal{X}_{S=s_j}, \mathcal{T}^*(\mathbf{x}_{\pi_1}) \leq \tilde{y} \leq \mathcal{T}^*(\mathbf{x}_{\pi_n})].\end{aligned} \tag{15}$$

Proof. Equation (15) indicates that $h(\mathbf{x})$ is independent from S. Therefore, $\zeta(h(\mathbf{x}))$ is independent from S, that is

$$\begin{aligned}&\mathbb{P}[\zeta(h(\mathbf{x})) \mid \mathbf{x} \in \mathcal{X}_{S=s_i}, \mathcal{T}^*(\mathbf{x}_{\pi_1}) \leq \tilde{y} \leq \mathcal{T}^*(\mathbf{x}_{\pi_n})] \\ =\ &\mathbb{P}[\zeta(h(\mathbf{x})) \mid \mathbf{x} \in \mathcal{X}_{S=s_j}, \mathcal{T}^*(\mathbf{x}_{\pi_1}) \leq \tilde{y} \leq \mathcal{T}^*(\mathbf{x}_{\pi_n})]\end{aligned} \tag{16}$$

where $\zeta(h(\mathbf{x})) = \hat{y}$. □

A.3 Proof for Proposition 2

Proof. Since $\mathcal{M}(\cdot, \cdot)$ is a distance metric between distributions, we have $\mathcal{M}(\cdot, \cdot) \geq 0$. Therefore,

$$\begin{aligned}&\sum_{s \in S} \mathcal{M}(\mathcal{P}(\mathcal{D}_{\mathcal{X}_{S=s}}), \mathcal{D}_{\mathbf{z}}) = 0 \\ \Rightarrow\ &\mathcal{M}(\mathcal{P}(\mathcal{D}_{\mathcal{X}_{S=s_i}}), \mathcal{D}_{\mathbf{z}}) = 0,\ \forall i \\ \Rightarrow\ &\mathcal{P}(\mathcal{D}_{\mathcal{X}_{S=s_i}}) = \mathcal{D}_{\mathbf{z}},\ \forall i \\ \Rightarrow\ &\mathcal{P}(\mathcal{D}_{\mathcal{X}_{S=s_i}}) = \mathcal{P}(\mathcal{D}_{\mathcal{X}_{S=s_j}}),\ \forall i, j.\end{aligned} \tag{17}$$

It follows that

$$\mathbb{P}[h(\mathbf{x}) \in \mathcal{P}(\mathcal{D}_{\mathcal{X}_{S=s_i}})] = \mathbb{P}[h(\mathbf{x}) \in \mathcal{P}(\mathcal{D}_{\mathcal{X}_{S=s_j}})] \tag{18}$$

holds for any i, j. Then

$$\mathbb{P}[h(\mathbf{x}) | \mathbf{x} \in \mathcal{X}_{S=s_i}] = \mathbb{P}[h(\mathbf{x}) | \mathbf{x} \in \mathcal{X}_{S=s_j}], \tag{19}$$

for any i, j. This means $h(\mathbf{x})$ is independent from S. We obtain

$$\begin{aligned} & \mathbb{P}[h(\mathbf{x}) \mid \mathbf{x} \in \mathcal{X}_{S=s_i}, \mathcal{T}^*(\mathbf{x}_{\pi_1}) \leq \tilde{y} \leq \mathcal{T}^*(\mathbf{x}_{\pi_n})] \\ = \; & \mathbb{P}[h(\mathbf{x}) \mid \mathbf{x} \in \mathcal{X}_{S=s_j}, \mathcal{T}^*(\mathbf{x}_{\pi_1}) \leq \tilde{y} \leq \mathcal{T}^*(\mathbf{x}_{\pi_n})]. \end{aligned} \tag{20}$$

□

B Experimental Settings

B.1 Datasets and Baselines

The statistics of all datasets are provided in Table 2. The details of each dataset are as follows:

Table 2: Statistics of datasets.

Dataset	Type	Dimension	Sensitive Variable	Normal Set	Abnormal Set
Adult	tabular	14	gender	income ≤50K	income > 50K
COMPAS	tabular	8	race	no recidivism within 2 years	recidivism within 2 years
Credit	tabular	23	age	no default payment next month	default payment next month
Titanic	tabular	7	gender	no survived	survived
SP	tabular	32	gender	general final grades	extreme final grades
CelebA	image	64×64×3	gender	attractive face	plain face

- **Adult**[5] [Becker and Kohavi, 1996] The dataset is from the 1994 Census Income database contains 48,842 samples with 14 attributes, and gender (male or female) is selected as the sensitive attribute. Following the previous work [Han et al., 2023], we removed the samples with missing values.
- **Credit**[6] [Yeh, 2009] The dataset is about customers' default payments in Taiwan and contains 30,000 samples with 23 attributes. Age is selected as the sensitive attribute where one group includes people from 30 to 60 years of age and the other is other age groups.
- **Compas**[7] [Larson et al., 2016] The dataset contains 7,214 samples with 52 attributes. Following previous works [Larson et al., 2016, Han et al., 2023], we only selected African-American and Caucasian individuals, yielding 5278 clean samples with 8 attributes. The sensitive attribute is race (African-American and Caucasian).
- **CelebA**[8] [Liu et al., 2015] The dataset contains 202,599 color face images. Following [Zhang and Davidson, 2021], we resized all images to 64 × 64. Gender (male or female) is selected as the sensitive attribute.
- **Titanic**[9] [Cukierski, 2012] The dataset is from a kaggle competition to predicting survival on the Titanic. We removed the samples with missing values and obtained 712 clean samples with 8 attributes. Gender (male and female) is selected as the sensitive attribute.
- **SP**[10] [Cortez, 2008] The dataset is about student achievement in secondary education of two Portuguese schools and contains 649 samples with 32 attributes. The data attributes include student grades, demographic, social and school related features. Gender (male and female) is selected as the sensitive attribute.

We used LOF [Breunig et al., 2000], Deep SVDD [Ruff et al., 2018] as standard (fairness-unaware) UAD baselines, FairOD [Shekhar et al., 2021], Deep Fair SVDD [Zhang and Davidson, 2021], and CFAD [Han et al., 2023] as end-to-end fairness-aware UAD baselines. We utilized FRL technique FarconVAE [Oh et al., 2022] to construct "FarconVAE+LOF" and "FarconVAE+DeepSVDD" as two-stage pipelines.

[5]https://archive.ics.uci.edu/dataset/2/adult
[6]https://archive.ics.uci.edu/dataset/350/default+of+credit+card+clients
[7]https://github.com/propublica/compas-analysis
[8]https://mmlab.ie.cuhk.edu.hk/projects/CelebA.html
[9]https://www.kaggle.com/c/titanic/data
[10]https://archive.ics.uci.edu/dataset/320/student+performance

The implementations of LOF and Deep SVDD are built using the PyOD[11] library [Zhao et al., 2019]. FarconVAE[12], FairOD[13] and CFAD[14] are based on official codes and the hyperparameters are fine-tuned according to the suggestions from original papers. In light of the unavailable implementation for Deep Fair SVDD, we reproduce the code following the pseudo-code provided in the original paper [Zhang and Davidson, 2021].

B.2 Implementation Details

Neural Network Architectures For the tabular datasets Adult, COMPAS, Credit, Titanic and SP, the neural networks for all methods are Multi-Layer Perceptrons (MLP). For the image dataset CelebA, the neural networks used in all methods are Convolutional Neural Networks (CNN). We use Adam [Kingma and Ba, 2015] as the optimizer, and set coefficient α of entropy regularization term in the Sinkhorn distance to 0.1 in all experiments.

Data Splitting For fairness problems in unsupervised anomaly detection, the proportion of sample size across different demographic groups heavily influences the results [Meissen et al., 2023, Wu et al., 2024]. Therefore, in this work, we set balanced splitting and skewed splitting. The detailed splits for the training set and test set are provided in Tables 3 and 4. Note that there is no balanced splitting for Titanic and SP due to the limitations of extremely uneven samples size across different demographic groups for Titanic and insufficient sample size for SP.

Table 3: Balanced data splitting on Adult, COMPAS, CelebA and Credit. PV denotes the protected variable. 'Nor.' and 'Abnor.' denote normal sample and abnormal sample, respectively.

Dataset (PV)	Attribute Value	Training Set	Test Set	
		Size (Nor.)	Size (Nor.)	Size (Abnor.)
Adult (Gender)	Male	6000	1000	1000
	Female	6000	1000	1000
COMPAS (Race)	African-American	1000	280	280
	Female	1000	280	280
CelebA (Gender)	Male	8000	4000	4000
	Female	8000	4000	4000
Credit (Age)	[30, 60]	5000	2000	2000
	Others	5000	2000	2000

Table 4: Skewed data splitting on all datasets. PV denotes the protected variable. 'Nor.' and 'Abnor.' denote normal sample and abnormal sample, respectively.

Dataset (PV)	Attribute Value	Training Set	Test Set	
		Size (Nor.)	Size (Nor.)	Size (Abnor.)
Adult (Gender)	Male	8000	4000	4000
	Female	2000	1000	1000
COMPAS (Race)	African-American	800	400	400
	Female	200	100	100
CelebA (Gender)	Male	8000	4000	4000
	Female	2000	1000	1000
Credit (Age)	[30, 60]	8000	4000	4000
	Others	2000	1000	1000
Titanic (Gender)	Male	330	30	30
	Female	32	30	30
SP (Gender)	Male	135	26	30
	Female	148	26	30

Evaluation Metrics Following previous works such as [Lahoti et al., 2020, Buyl and De Bie, 2022], we use the AUC (Area Under the Receiver Operating Characteristic curve) score and F1 to

[11] https://github.com/yzhao062/pyod
[12] https://github.com/changdaeoh/FarconVAE
[13] https://github.com/Shubhranshu-Shekhar/fairOD
[14] https://github.com/hanxiao0607/CFAD

evaluate the detection accuracy of all compared methods. We use the proposed ADPD (threshold-free) and *fairness ratio* (threshold-dependent) [Zhang and Davidson, 2021] to evaluate the fairness of unsupervised anomaly detection. Note that the threshold is determined by percentile (p) of anomaly score on the training data and we set $p = \{0.9, 0.95\}$ for all baselines.

ALL experiments were conducted on 20 Cores Intel(R) Xeon(R) Gold 6248 CPU with one NVIDIA Tesla V100 GPU, CUDA 12.0. We run each experiment five times and report the average results with standard variance.

B.3 Threshold-Free Fairness Metrics

When evaluating the fairness of anomaly detection, existing work such as [Zhang and Davidson, 2021, Shekhar et al., 2021] usually uses a *fairness ratio*

$$r := \min\left(\frac{\mathbb{P}(\text{Score}(\mathcal{X})>t|S=s_i)}{\mathbb{P}(\text{Score}(\mathcal{X})>t|S=s_j)}, \frac{\mathbb{P}(\text{Score}(\mathcal{X})>t|S=s_j)}{\mathbb{P}(\text{Score}(\mathcal{X})>t|S=s_i)}\right) \tag{21}$$

where $\text{Score}(\cdot)$ denotes the anomaly score, t is a threshold and S denotes a protected variable. The metric has two limitations. First, (21) is sensitive to the selection of threshold t. Second, (21) will not work (undefined cases) when $\mathbb{P}(\text{Score}(\mathbf{x}) > t|S = s_j) = 0$ or $\mathbb{P}(\text{Score}(\mathbf{x}) > t|S = s_i) = 0$, where r is always zero. To overcome the two limitations, in this paper, we propose a new fairness metric called *Average Demographic Parity Difference* (ADPD):

$$\text{ADPD} := \frac{1}{n}\sum_{k=1}^{n}\Big|\mathbb{P}(\text{Score}(\mathcal{X}) > t_k|S = s_i) - \mathbb{P}(\text{Score}(\mathcal{X}) > t_k|S = s_j)\Big| \tag{22}$$

where $t_k \in \text{Score}(\mathcal{X})$ denotes the anomaly score of single sample. In our proposed methods, $t_k = \|h_{\phi^*}(\mathbf{x}_k)\|$. ADPD is a threshold-free metric (such as AUC) for measuring demographic parity. The range of ADPD is $[0, 1)$ and a smaller ADPD means a higher fairness. Overall, due to threshold-independent, ADPD provides a holistic view of the model's fairness by evaluating the performance of an AD method across all possible thresholds. Although we introduce a novel threshold-free metric for fairness measure, this is not to imply that the threshold-dependent metrics are useless.

The ADPD evaluates the performance of a model across all possible thresholds, providing a holistic view of the model's fairness. In contrast, the *fairness ratio* is calculated based on a specific threshold and exhibits high sensitivity to threshold selection (See Table 8 and Table 7). When prior knowledge (e.g., anomaly prevalence, operational constraints) is unavailable, ADPD is preferable as it mitigates bias introduced by arbitrary threshold choices, ensuring equitable performance comparisons. On the other hand, if domain-specific priors (e.g., training set contamination rate) are available, threshold-dependent metrics like the *fairness ratio* may align better with real requirements.

C Extension of Proposed Method

The original optimization objective of Im-FairAD is as follows:

$$\min_{\phi,\psi}\sum_{s\in S}\text{Sinkhorn}(h_\phi(\mathcal{X}_{S=s}), \mathcal{Z}) + \frac{\beta}{n}\sum_{i=1}^{n}\|\mathbf{x}_i - g_\psi(h_\phi(\mathbf{x}_i))\|^2, \tag{23}$$

where only single protected attribute is considered. When protecting multiple sensitive attributes (e.g., race & gender), the optimization objective of Im-FairAD can be naturally reformulated to the following form (There is the same reforming process on Ex-FairAD).

$$\min_{\phi,\psi}\sum_{S\in\Omega}\sum_{s\in S}\text{Sinkhorn}(h_\phi(\mathcal{X}_{S=s}), \mathcal{Z}) + \frac{\beta}{n}\sum_{i=1}^{n}\|\mathbf{x}_i - g_\psi(h_\phi(\mathbf{x}_i))\|^2, \tag{24}$$

where $\Omega = \{S^{(\text{race})}, S^{(\text{gender})}, \cdots\}$ denotes the set of sensitive attributes. Meanwhile, the proposed fairness metric, ADPD, becomes

$$\text{ADPD} := \frac{1}{n \cdot |\Omega|} \sum_{S \in \Omega} \sum_{k=1}^{n} \Big| \mathbb{P}(\text{Score}(\mathcal{X}) > t_k | S = s_i) - \mathbb{P}(\text{Score}(\mathcal{X}) > t_k | S = s_j) \Big|. \tag{25}$$

When the number of values of the protected attribute (e.g., race) exceeds two, i.e., $|S| > 2$, the ADPD becomes

$$\text{ADPD} := \frac{1}{n \cdot |\Omega|} \sum_{S \in \Omega} \sum_{k=1}^{n} \max \left(\left\{ \left| \mathbb{P}(s_i) - \mathbb{P}(s_j) \right| \right\}_{i \neq j}^{s_i, s_j \in S} \right), \tag{26}$$

where $\mathbb{P}(s_i) = \mathbb{P}(\text{Score}(\mathcal{X}) > t_k | S = s_i)$ and the range of ADPD still is [0, 1).

D Numerical Results and More Visualization

D.1 Results of ADPD and AUC on all datasets

Table 5: Results of ADPD and AUC on COMPAS, Adult, Credit and CelebA. Note that the baselines, FairOD and CFAD are tailored to tabular data. In the official code of FarconVAE, there is no support for image data.

Methods	Balanced Splitting			Skewed Splitting		
	AUC(%) ↑	ADPD(%) ↓		AUC(%) ↑	ADPD(%) ↓	
		normal	all		normal	all
COMPAS						
LOF	59.23(0.00)	12.27(0.00)	10.59(0.00)	57.25(0.00)	9.33(0.00)	9.53(0.00)
Deep SVDD	57.58(2.30)	24.33(7.87)	21.07(8.72)	60.24(3.08)	31.72(7.39)	30.18(5.32)
FarconVAE+LOF	48.72(0.00)	4.56(0.00)	6.56(0.00)	50.10(0.00)	7.25(0.00)	**4.12**(0.00)
FarconVAE+Deep SVDD	50.48(1.67)	**3.99**(1.51)	4.77(0.72)	48.83(0.75)	8.50(1.25)	7.03(1.10)
FairOD	61.05(1.05)	5.39(2.11)	5.53(1.50)	58.86(4.75)	**5.43**(0.60)	**5.32**(3.24)
Deep Fair SVDD	62.71(0.76)	10.25(2.40)	8.25(1.46)	61.05(1.07)	10.58(3.23)	11.04(3.13)
CFAD	62.27(0.68)	12.21(4.57)	10.94(4.40)	60.87(1.99)	23.81(3.71)	17.39(2.20)
Ex-FairAD (Ours)	**63.11**(2.67)	4.38 (0.43)	**4.54** (1.38)	**66.44**(3.62)	5.44 (1.05)	7.39 (1.18)
Im-FairAD (Ours)	**63.89**(3.45)	**4.16** (2.55)	**4.76** (2.37)	**66.58**(2.91)	**4.90** (1.00)	5.43 (0.49)
Adult						
LOF	58.09(0.00)	26.46(0.00)	25.58(0.00)	59.14(0.00)	25.19(0.00)	25.34(0.00)
Deep SVDD	60.47(3.59)	7.09(1.92)	7.35(1.62)	61.04(3.07)	19.43(5.40)	17.32(2.90)
FarconVAE+LOF	53.92(0.00)	5.00(0.00)	5.73(0.00)	50.56(0.00)	2.16(0.00)	3.36(0.00)
FarconVAE+Deep SVDD	55.05(2.81)	4.65(3.01)	5.76(3.24)	51.21(2.61)	5.65(5.84)	5.81(7.09)
FairOD	70.63(0.36)	11.80(4.62)	9.26(1.45)	58.30(4.18)	6.34(0.53)	14.35(5.97)
Deep Fair SVDD	62.45(0.65)	**2.18**(0.84)	3.01(2.27)	60.22(0.14)	6.86(1.85)	4.14(3.74)
CFAD	66.89(1.12)	6.07(2.95)	14.69(3.58)	60.28(1.09)	23.42(5.43)	30.41(3.64)
Ex-FairAD (Ours)	**73.49**(4.41)	2.92 (0.81)	**2.86** (1.09)	**69.56**(3.45)	**1.52** (0.44)	**1.79** (0.76)
Im-FairAD (Ours)	**72.05**(1.81)	**1.61** (0.61)	**2.13** (1.17)	**69.34**(1.82)	**1.92** (1.05)	**1.70** (0.86)
Credit						
LOF	50.09(0.00)	6.28(0.00)	3.19(0.00)	48.98(0.00)	2.14(0.00)	3.17(0.00)
Deep SVDD	52.34(1.98)	9.84(4.56)	8.64(3.26)	54.82(1.97)	7.49(1.33)	6.95(2.55)
FarconVAE+LOF	49.86(0.00)	2.21(0.00)	2.24(0.00)	52.93(0.00)	2.65(0.00)	**0.80**(0.00)
FarconVAE+Deep SVDD	51.07(0.58)	**1.03**(0.45)	**1.01**(0.32)	50.96(0.25)	**1.49**(0.70)	**0.88**(0.66)
FairOD	53.93(1.09)	7.00(1.33)	7.21(1.42)	60.39(0.55)	8.79(0.73)	7.36(0.50)
Deep Fair SVDD	56.01(0.69)	4.62(1.11)	4.54(1.22)	56.67(0.42)	6.24(1.88)	5.84(2.00)
CFAD	56.96(0.55)	5.89(0.75)	4.48(0.66)	57.21(0.47)	6.30(2.25)	5.76(1.94)
Ex-FairAD (Ours)	**64.96**(1.63)	2.95 (0.83)	**1.92** (0.39)	**63.85**(1.95)	2.71 (0.85)	1.57 (0.87)
Im-FairAD (Ours)	**63.70**(1.59)	**2.20** (0.86)	1.97 (0.66)	**65.13**(1.79)	**2.34** (0.59)	1.33 (0.24)
CelebA						
Deep SVDD	**62.77**(0.35)	10.44(1.66)	10.47(1.55)	60.80(0.68)	9.82(0.44)	8.24(0.93)
Deep Fair SVDD	57.97(0.72)	7.73(0.99)	4.79(0.35)	59.95(0.80)	**1.52**(1.21)	**3.24**(1.81)
Ex-FairAD (Ours)	**60.68**(0.72)	**1.06** (0.47)	**3.80** (0.83)	**63.07**(0.71)	**1.91** (0.38)	4.87 (0.60)
Im-FairAD (Ours)	58.07(0.87)	**2.37** (0.98)	**2.46** (0.45)	**61.72**(0.62)	2.73 (0.86)	**2.93** (0.54)

The results of ADPD and AUC on all datasets are provided in Table 5 and Table 6, where the best two results in each case are marked in **bold**, 'normal' means the score is computed only on normal samples of the test set, and 'all' means the score is computed on all samples from test set. From these two tables, we have the following observations:

- Im-FairAD and Ex-FairAD both achieve better detection accuracy (AUC) compared with all baselines not only on normal data but also on entire test set in most cases, while maintaining comparable or even better fairness (ADPD) **[Q1]**.
- In the transition from a balanced splitting to a skewed splitting, group fairness, as measured by ADPD, exhibits significantly larger values in most cases, which indicates the fairness of existing fairness-aware AD methods is easily affected by sample proportion from different sensitive groups and skewed data for different demographic groups poses a more intractable fairness problem than balanced data. In contrast, the fluctuation of ADPD is slight on our proposed methods **[Q2]**.

Table 6: Results of ADPD and AUC on Titanic and SP with gender as the sensitive attribute. Note that there is no balanced splitting for Titanic and SP due to the limitations of extremely uneven samples size across different demographic groups for Titanic and insufficient sample size for SP.

Methods	Titanic			SP		
	AUC(%) ↑	ADPD(%) ↓		AUC(%) ↑	ADPD(%) ↓	
		normal	all		normal	all
LOF	53.88(0.00)	23.16(0.00)	19.81(0.00)	63.55(0.00)	9.82(0.00)	6.03(0.00)
Deep SVDD	53.21(1.98)	26.02(11.29)	26.07(11.95)	68.37(1.11)	12.97(4.03)	12.52(2.44)
FarconVAE+LOF	48.52(0.00)	7.97(0.00)	6.58(0.00)	57.30(0.00)	**4.54**(0.00)	8.94(0.00)
FarconVAE+Deep SVDD	50.04(2.97)	7.60(1.89)	**3.99**(0.85)	55.00(3.17)	9.53(3.44)	5.98(1.13)
FairOD	46.04(0.20)	14.67(0.30)	6.38(0.16)	42.04(0.44)	9.70(0.51)	10.90(0.31)
Deep Fair SVDD	51.94(0.99)	20.67(3.08)	21.24(6.35)	61.90(1.20)	10.57(3.04)	7.24(2.12)
CFAD	55.63(0.59)	26.77(7.00)	23.02(2.10)	69.23(1.34)	9.26(1.24)	**4.53**(0.28)
Ex-FairAD (Ours)	**64.87**(1.48)	**6.35**(2.12)	5.75(1.00)	**74.16**(3.09)	6.00(1.32)	5.84(1.67)
Im-FairAD (Ours)	**64.02**(3.62)	**7.14**(2.61)	**3.82**(1.77)	**75.97**(4.32)	**5.84**(1.29)	**5.62**(1.95)

The visualization between detection accuracy (AUC) and fairness (ADPD) on skewed splitting is shown in Figure 4. The ADPD of abnormal data on Titanic and SP is shown in Figure 5.

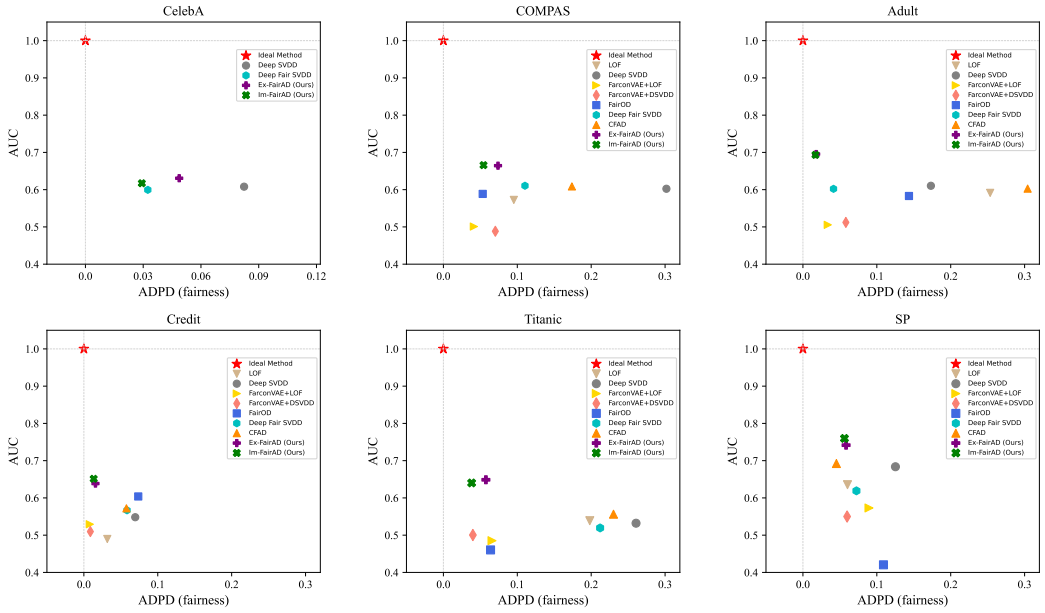

Figure 4: Accuracy-fairness trade-off on skewed splitting.

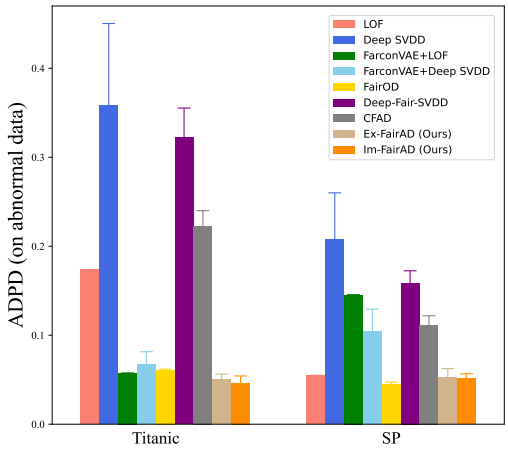

Figure 5: Fairness of all baselines on abnormal data from test set.

D.2 Results of *fairness ratio* and F1 on all datasets

Furthermore, the results of F1-score and *fairness ratio* on all datasets are provided in Table 8 and Table 7, where *fairness ratio* is highly sensitive to different thresholds $p = \{0.90, 0.95\}$.

Table 7: Results of F1-score and *fairness ratio* on Titanic and SP. Note that '0*' means that undefined case occurs when calculating *fairness ratio*.

Methods	Threshold=0.90			Threshold=0.95		
	F1(%) ↑	Fairness ratio ↑		F1(%) ↑	Fairness ratio ↑	
		normal	all		normal	all
Titanic						
LOF	35.55(0.00)	0.40(0.00)	0.57(0.00)	26.82(0.00)	0.22(0.00)	0.46(0.00)
Deep SVDD	31.90(10.57)	0.44(0.32)	0.44(0.30)	16.66(6.47)	0.68(0.36)	0.56(0.33)
FarconVAE+LOF	8.82(0.00)	0.66(0.00)	0.6(0.00)	6.15(0.00)	0.50(0.00)	0.25(0.00)
FarconVAE+Deep SVDD	17.06(5.66)	0.57(0.21)	0.69(0.19)	9.03(4.70)	0.16(0.21)	0.28(0.16)
FairOD	15.55(2.77)	0.57(0.09)	0.57(0.04)	11.42(0.00)	0.50(0.00)	**1.00**(0.00)
Deep Fair SVDD	14.31(2.26)	0.68(0.30)	0.66(0.23)	12.83(1.46)	0.80(0.24)	0.40(0.15)
CFAD	44.54(5.47)	0.37(0.13)	0.57(0.24)	27.73(1.02)	0.53(0.06)	0.48(0.05)
Ex-FairAD (Ours)	**66.66**(0.00)	**1.00**(0.00)	**1.00**(0.00)	**66.82**(0.00)	**1.00**(0.00)	**1.00**(0.00)
Im-FairAD (Ours)	**66.66**(0.00)	**1.00**(0.00)	**1.00**(0.00)	**66.66**(0.00)	**1.00**(0.00)	**1.00**(0.00)
SP						
LOF	44.44(0.00)	0.14(0.00)	0.57(0.00)	27.02(0.00)	0.12(0.00)	0.24(0.00)
Deep SVDD	22.45(6.56)	0.38(0.30)	0.47(0.19)	20.42(8.47)	0.59(0.35)	0.46(0.32)
FarconVAE+LOF	19.04(0.00)	0*	0.65(0.00)	6.89(0.00)	0*	0*
FarconVAE+Deep SVDD	14.41(2.98)	0.40(0.36)	0.59(0.26)	12.90(3.70)	0*	0*
FairOD	11.72(0.13)	**0.88**(0.02)	0.68(0.06)	13.65(1.41)	**0.86**(0.00)	0.26(0.01)
Deep Fair SVDD	19.76(4.81)	0.59(0.15)	0.58(0.22)	24.65(6.76)	0.42(0.12)	**0.71**(0.10)
CFAD	40.46(2.22)	**0.74**(0.11)	**0.78**(0.07)	32.14(6.57)	0.38(0.00)	**0.80**(0.10)
Ex-FairAD (Ours)	**41.77**(6.28)	0.54(0.15)	**0.77**(0.19)	**34.10**(5.08)	0.62(0.20)	0.65(0.10)
Im-FairAD (Ours)	**43.64**(9.16)	0.56(0.11)	0.55(0.26)	**35.51**(4.11)	**0.86**(0.00)	0.62(0.16)

E Ablation Study

In this section, we delve into the influence of different loss terms for our proposed method. More specifically, we investigate the reconstruction error term within the optimization objective of Im-FairAD and the fairness term within the optimization objective of Ex-FairAD.

Table 8: Results of F1-score and *fairness ratio* on COMPAS, Adult, Credit and CelebA. Note that the baselines, FairOD and CFAD are tailored to tabular data. In the official code of FarconVAE, there is no support for image data.

Methods	Threshold=0.90			Threshold=0.95		
	F1(%) ↑	Fairness ratio ↑		F1(%) ↑	Fairness ratio ↑	
		normal	all		normal	all
COMPAS						
LOF	29.85(0.00)	0.27(0.00)	0.35(0.00)	19.16(0.00)	0.34(0.00)	0.36(0.00)
Deep SVDD	31.30(5.17)	0.49(0.18)	0.58(0.18)	20.67(5.09)	0.39(0.18)	0.53(0.30)
FarconVAE+LOF	14.28(0.00)	0.69(0.00)	**0.77**(0.00)	8.11(0.00)	0.64(0.00)	0.55(0.00)
FarconVAE+Deep SVDD	19.50(0.89)	**0.74**(0.07)	0.70(0.14)	8.58(0.65)	**0.68**(0.22)	**0.75**(0.23)
FairOD	17.81(0.11)	0.70(0.04)	**0.79**(0.01)	14.54(0.66)	**0.87**(0.01)	**0.88**(0.06)
Deep Fair SVDD	42.23(0.81)	0.46(0.04)	0.58(0.04)	**31.54**(0.63)	0.53(0.09)	0.56(0.03)
CFAD	40.67(3.67)	0.39(0.13)	0.49(0.08)	26.95(1.49)	0.37(0.06)	0.37(0.03)
Ex-FairAD (Ours)	**42.30**(5.88)	**0.71**(0.18)	0.74(0.15)	28.07(5.06)	0.58(0.21)	0.51(0.07)
Im-FairAD (Ours)	**47.49**(4.32)	0.65(0.28)	0.68(0.09)	**33.88**(2.75)	0.57(0.21)	0.68(0.13)
Adult						
LOF	27.45(0.00)	0.34(0.00)	0.42(0.00)	24.58(0.00)	0.30(0.00)	0.34(0.00)
Deep SVDD	44.73(4.72)	0.54(0.24)	0.63(0.11)	26.01(13.02)	0.48(0.13)	0.54(0.07)
FarconVAE+LOF	22.87(0.00)	0.63(0.00)	0.62(0.00)	13.33(0.00)	0.73(0.00)	0.58(0.00)
FarconVAE+Deep SVDD	22.08(8.87)	0.59(0.10)	0.73(0.18)	10.33(4.14)	0.46(0.27)	0.48(0.22)
FairOD	36.87(3.99)	0.70(0.06)	0.32(0.04)	24.82(1.61)	0.67(0.04)	0.81(0.05)
Deep Fair SVDD	45.03(1.56)	**0.86**(0.01)	0.83(0.05)	39.43(1.35)	**0.83**(0.08)	**0.86**(0.06)
CFAD	48.01(5.51)	0.76(0.08)	0.55(0.14)	35.76(1.60)	0.77(0.06)	0.63(0.10)
Ex-FairAD (Ours)	**52.75**(2.19)	**0.87**(0.11)	**0.92**(0.03)	**48.46**(4.87)	0.80(0.12)	0.83(0.07)
Im-FairAD (Ours)	**56.04**(4.70)	0.72(0.06)	**0.84**(0.07)	**47.79**(1.84)	**0.90**(0.06)	**0.92**(0.04)
Credit						
LOF	20.11(0.00)	0.45(0.00)	0.55(0.00)	11.78(0.00)	0.24(0.00)	0.33(0.00)
Deep SVDD	23.31(7.75)	0.76(0.12)	0.74(0.20)	16.26(8.71)	0.72(0.11)	0.63(0.21)
FarconVAE+LOF	15.06(0.00)	0.88(0.00)	0.88(0.00)	8.06(0.00)	0.79(0.00)	0.83(0.00)
FarconVAE+Deep SVDD	16.96(1.80)	0.94(0.02)	0.95(0.02)	8.70(1.64)	0.85(0.12)	0.87(0.10)
FairOD	41.00(0.81)	0.70(0.04)	0.74(0.02)	35.18(1.02)	0.66(0.04)	0.70(0.07)
Deep Fair SVDD	13.54(1.79)	**0.81**(0.06)	0.73(0.09)	15.47(0.76)	0.71(0.03)	0.70(0.05)
CFAD	20.43(0.75)	0.69(0.05)	0.69(0.04)	10.53(0.22)	0.59(0.05)	0.58(0.02)
Ex-FairAD (Ours)	**53.70**(2.53)	0.79(0.03)	**0.90**(0.04)	**41.06**(3.79)	**0.77**(0.06)	**0.82**(0.10)
Im-FairAD (Ours)	**54.32**(3.43)	**0.81**(0.08)	**0.90**(0.04)	**44.30**(1.94)	**0.72**(0.05)	**0.86**(0.05)
CelebA						
Deep SVDD	30.86(1.38)	**0.91**(0.06)	0.81(0.05)	23.08(1.24)	0.82(0.09)	0.76(0.03)
Deep Fair SVDD	14.40(2.8)	0.58(0.18)	0.65(0.11)	16.95(1.32)	0.63(0.22)	0.69(0.12)
Ex-FairAD (Ours)	**50.29**(9.04)	0.86(0.10)	**0.84**(0.11)	**55.40**(10.57)	**0.83**(0.08)	**0.88**(0.08)
Im-FairAD (Ours)	**30.89**(2.34)	**0.89**(0.08)	**0.88**(0.07)	**19.83**(2.88)	**0.87**(0.04)	**0.84**(0.07)

E.1 Reconstruction Error Term in Im-FairAD

For the optimization objective of Im-FairAD, we adjust the hyperparameter β across the range $\{0.001, 0.01, 0.1, 1, 10, 100, 1000\}$ to observe the changing of performance, including detection accuracy and fairness. The experimental results are shown in Figure 6, where (a) and (b) depict the fluctuation of AUC with varied β on balanced and skewed data, respectively. And (c) and (d) depict the fluctuation of ADPD (all test set) with varied β on balanced and skewed data, respectively. From Figure 6, we observe that as β increases, the fluctuation of AUC on both balanced and skewed data gradually diminishes. Conversely, the fluctuation of ADPD becomes more pronounced and tends to increase. This observation aligns with expectations, as the dominance of the reconstruction term in the optimization objective makes it challenging to map different protected groups into the same target distribution.

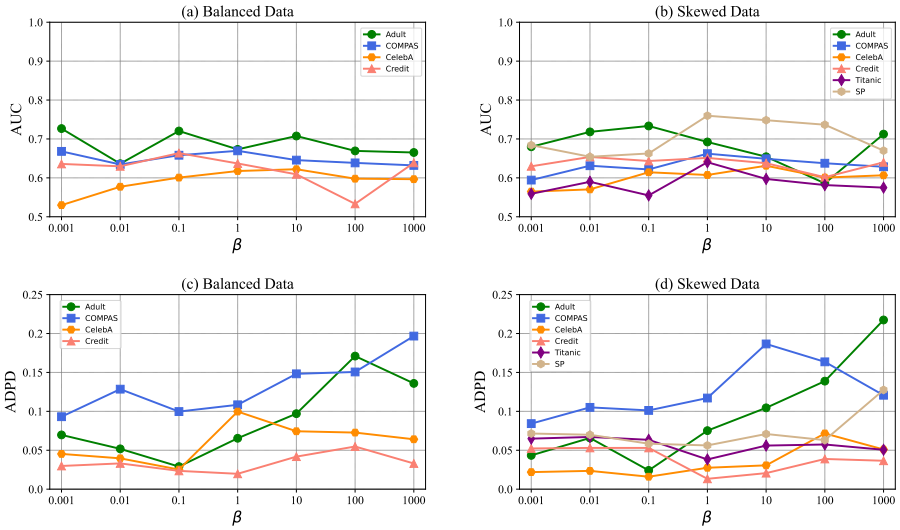

Figure 6: Average AUC and ADPD (all test set) with β varies in the range of $\{0.001, 0.01, 0.1, 1, 10, 100, 1000\}$.

E.2 Fairness Term in Ex-FairAD

For the optimization objective of Ex-FairAD, we remove the fairness regularization term and conduct experiments on all datasets. The results are reported in Table 9. Observing Table 9, the AUC (detection accuracy) of Ex-FairAD improved but fairness measured by ADPD suffers from adverse effects when without fairness term in optimization objective. Moreover, we adjust the hyperparameter λ across the range $\{0.001, 0.01, 0.1, 1, 10, 100, 1000\}$ to observe the changes of performance, including detection accuracy and fairness. The experimental results are visualized in Figure 7.

Table 9: Comparison between the objective with fairness regularization and the objective without fairness regularization in optimization problem of Ex-FairAD.

Datasets	Methods	Balanced Split			Skewed Split		
		AUC(%) ↑	ADPD		AUC(%) ↑	ADPD	
			normal	all		normal	all
COMPAS	Ex-FairAD	63.11	4.38	4.54	66.44	5.44	7.36
	W/O Fairness term	67.22	13.02	12.90	68.91	14.40	12.05
Adult	Ex-FairAD	73.49	2.92	2.86	69.56	1.52	1.79
	W/O Fairness term	75.32	5.56	5.39	74.94	9.29	8.07
CelebA	Ex-FairAD	60.68	1.06	3.80	63.07	1.91	4.87
	W/O Fairness term	62.56	3.44	5.30	64.29	3.29	3.55
Credit	Ex-FairAD	64.96	2.95	1.92	63.85	2.71	1.57
	W/O Fairness term	65.40	7.21	5.87	66.63	7.33	4.64
Titanic	Ex-FairAD	NA	NA	NA	64.87	6.35	5.75
	W/O Fairness term	NA	NA	NA	68.00	10.27	9.79
SP	Ex-FairAD	NA	NA	NA	74.16	6.00	5.84
	W/O Fairness term	NA	NA	NA	76.49	8.59	8.89

F Selection of Distance Metric $\mathcal{M}$ between Distributions

Based on the analysis on Module Formulation, we obtain the following optimization problem

$$\min_{\phi,\psi} \sum_{s\in S} \mathcal{M}(\mathcal{D}_{h_\phi(\mathcal{X}_{S=s})}, \mathcal{D}_{\mathbf{z}}) + \beta \mathcal{M}(\mathcal{D}_{g_\psi(h_\phi(\mathbf{x}))}, \mathcal{D}_{\mathbf{x}}). \tag{27}$$

However, the problem 27 is intractable as data distribution $\mathcal{D}_{\mathbf{x}}$ is unknown and $\mathcal{D}_{h_\phi(\mathcal{X}_{S=s})}, \mathcal{D}_{g_\psi(h_\phi(\mathbf{x}))}$ cannot be computed analytically, which leads to that we cannot use f-divergence [Rényi, 1961], such

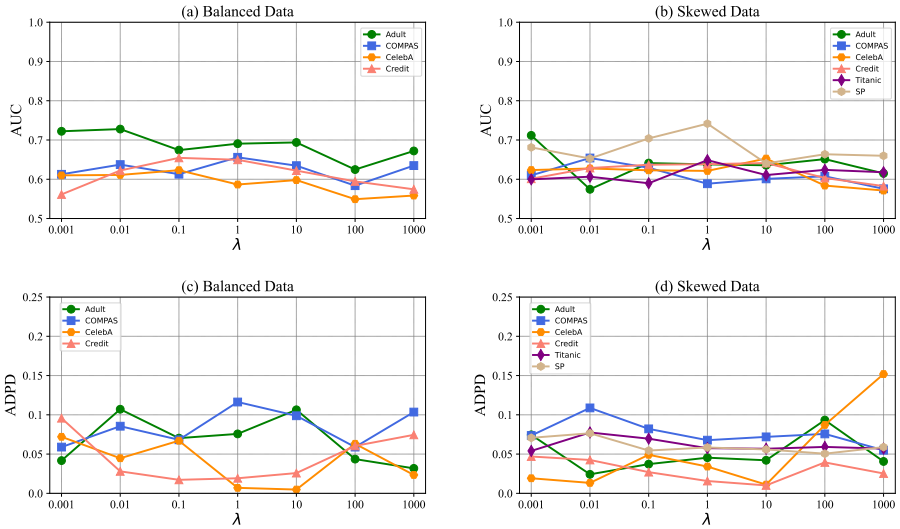

Figure 7: Average AUC and ADPD (all test set) with λ varies in the range of $\{0.001, 0.01, 0.1, 1, 10, 100, 1000\}$ on datasets Adult, COMPAS and CelebA.

as KL-divergence, Hellinger distance, to measure the difference between $\mathcal{D}_{h_\phi(\mathcal{X}_{S=s})}$ and $\mathcal{D}_{\mathbf{z}}$. Thus, we need to measure the divergence between $\mathcal{D}_{h_\phi(\mathcal{X}_{S=s})}$ and $\mathcal{D}_{\mathbf{z}}$ by using the known samples. In such situation, Wasserstein distance, Maximum Mean Discrepancy (MMD) [Gretton et al., 2012] and Sinkhorn distance [Cuturi, 2013] are all possible choices.

However, the computation cost of Wasserstein distance can quickly become prohibitive when the data dimension increases. Therefore, the options left are Sinkhorn distance and Maximum Mean Discrepancy.

Based on [Gretton et al., 2012], MMD is defined as

$$\mathrm{MMD}[\mathcal{F}, p, q] = \sup_{\|f\|_{\mathcal{H}} \leq 1} \left(\mathbb{E}_p[f(\mathbf{x})] - \mathbb{E}_q[f(\mathbf{y})]\right), \tag{28}$$

where p, q are probability distributions, $\mathcal{F}$ is a class of functions $f : \mathbb{R}^d \to \mathbb{R}$ and $\mathcal{H}$ denotes a reproducing kernel Hilbert space.

Its empirical estimate is

$$\begin{aligned} \mathrm{MMD}^2[\mathcal{F}, X, Y] = & \frac{1}{m(m-1)} \sum_{i=1}^{m} \sum_{j \neq i}^{m} k(\mathbf{x}_i, \mathbf{x}_j) \\ & + \frac{1}{n(n-1)} \sum_{i=1}^{n} \sum_{j \neq i}^{n} k(\mathbf{y}_i, \mathbf{y}_j) - \frac{2}{mn} \sum_{i=1}^{m} \sum_{j=1}^{n} k(\mathbf{x}_i, \mathbf{y}_j), \end{aligned} \tag{29}$$

where $X = \{\mathbf{x}_1, \ldots, \mathbf{x}_m\}$ and $Y = \{\mathbf{y}_1, \ldots, \mathbf{y}_n\}$ are samples consisting of i.i.d observations drawn from p and q, respectively. $k(\cdot, \cdot)$ denotes a kernel function.

We use Sinkhorn distance and MMD to replace the first term of the optimization problem 27, respectively, and use reconstruction error to replace the second term of the problem 27. The related experimental results are provided in Table 10, where we use Gaussian kernel $\exp(-\gamma\|\mathbf{x} - \mathbf{y}\|^2)$ as kernel function of MMD.

Based on the empirical results in COMPAS, we select Sinkhorn distance as distribution distance metric of the optimization problem 27.

Table 10: The comparison between Sinkhorn distance and MMD on COMPAS.

Methods	AUC(%) ↑	ADPD(%) ↓	
		normal	all
Im-FairAD (MMD)	60.78	6.75	6.81
Im-FairAD (Sinkhorn)	63.89	4.16	4.76
Ex-FairAD (MMD)	57.09	2.60	4.43
Ex-FairAD (Sinkhorn)	63.11	4.38	4.54

G Time Complexity and Implementation Cost Analysis

G.1 Time Complexity Analysis

As most baselines are based on deep learning techniques, the main time cost is from the training of neural networks. Thus, based on our proposed method, we analyze the time complexity of neural networks and other methods have the similar analysis process.

For convenience, the h_ϕ is specified as follows

$$h_\phi(\mathbf{x}) := \mathbf{W}^h_{L_h}(\sigma(\cdots(\mathbf{W}^h_2(\sigma(\mathbf{W}^h_1\mathbf{x})\cdots)), \tag{30}$$

where $\phi = \{\mathbf{W}^h_1, \mathbf{W}^h_2, \ldots, \mathbf{W}^h_{L_h}\}$, $\mathbf{W}^h_l \in \mathbb{R}^{d^h_l \times d^h_{l-1}}$, and L_h is the number of layers of the network. The definition of the layer width indicates that $d^h_0 = d$ and $d^h_{L_h} = m$. σ denotes the activation function such as ReLU, LeakyReLU, Sigmoid, or Tanh. The activation functions in different layers are not necessarily the same but here we use the same σ for convenience. Similarly, the g_ψ is specified as follows

$$g_\psi(\mathbf{z}) := \mathbf{W}^g_{L_g}(\sigma(\cdots(\mathbf{W}^g_2(\sigma(\mathbf{W}^g_1\mathbf{z})\cdots)), \tag{31}$$

where $\psi = \{\mathbf{W}^g_1, \mathbf{W}^g_2, \ldots, \mathbf{W}^g_{L_g}\}$, $\mathbf{W}^g_l \in \mathbb{R}^{d^g_l \times d^g_{l-1}}$, $d^g_1 = m$, and $d^g_{L_g} = d$. Note that we have omitted the bias terms of h_ϕ and g_ψ for simplicity.

In the training stage, suppose the batch size of optimization is data size n, then the time complexity of training neural network is $\mathcal{O}(T(n\sum_{l=1}^{L_h} d^h_l d^h_{l-1} + n\sum_{l=1}^{L_g} d^g_l d^g_{l-1}))$, where T is the total number of iterations of h_ϕ and g_ψ. If we further assume $\max(\max_l d^h_l, \max_l d^g_l) \leq \bar{d}$, and $L_h + L_g \leq \bar{L}$, the time complexity of neural network is at most $\mathcal{O}(Tn\bar{d}^2\bar{L})$. In addition, for the proposed Im-FairAD and Ex-FairAD, we utilize the Sinkhorn distance to compute the loss. Therefore, the overall time complexity is $\mathcal{O}(T(n\bar{d}^2\bar{L} + tn^2))$ where the time complexity of Sinkhorn algorithm is $\mathcal{O}(n^2)$ and t is the maximum iterations of Sinkhorn algorithm.

In the inference phase, for m new samples, the time complexity of computing the anomaly score is $\mathcal{O}(m\bar{d}^2\bar{L} + m)$ where $\mathcal{O}(m\bar{d}^2\bar{L})$ is from neural networks and $\mathcal{O}(m)$ is from the calculation of anomaly score. The detailed comparison of time complexity is provided in Table 11, where all the methods including ours have the same time complexity in inference phase.

Table 11: The time complexity of training and inference.

Methods	Time Complexity (Training)	Time Complexity (Inference)
Deep SVDD	$\mathcal{O}((T_{ae} + T_{oc})(n\bar{d}^2\bar{L} + n))$	$\mathcal{O}(m\bar{d}^2\bar{L} + m)$
FairOD	$\mathcal{O}(T(n\bar{d}^2\bar{L} + n))$	$\mathcal{O}(m\bar{d}^2\bar{L} + m)$
Deep Fair SVDD	$\mathcal{O}((T_{ae} + T_d + T_g)(n\bar{d}^2\bar{L} + n))$	$\mathcal{O}(m\bar{d}^2\bar{L} + m)$
CFAD	$\mathcal{O}((T_{gae} + T_d + T_{ae} + T_c)(n\bar{d}^2\bar{L} + n))$	$\mathcal{O}(m\bar{d}^2\bar{L} + m)$
Ex-FairAD	$\mathcal{O}(T(n\bar{d}^2\bar{L} + tn^2))$	$\mathcal{O}(m\bar{d}^2\bar{L} + m)$
Im-FairAD	$\mathcal{O}(T(n\bar{d}^2\bar{L} + tn^2))$	$\mathcal{O}(m\bar{d}^2\bar{L} + m)$

G.2 Implementation Cost Analysis

In Table 12, we report the training time and the time occupied by Sinkhorn distance in single epoch. We keep $\epsilon = 1e^{-4}$ (stop threshold on error) on the three tabular datasets (COMPAS, Adult and

Credit) and $\epsilon = 1e^{-3}$ for CelebA. The α (coefficient of entropic regularization term) is consistently set to 0.1 across all experiments.

Table 12: The training time (seconds) on the datasets with balanced splitting.

Dataset	Settings		Time			
	# samples	dimension	Ex-FairAD (all)	Ex-FairAD (Sinkhorn)	Im-FairAD (all)	Im-FairAD (Sinkhorn)
Compas	2000	4	0.3842	0.1888	0.6028	0.3414
Adult	12000	4	2.932	1.645	4.645	2.900
Credit	10000	8	2.922	1.710	4.770	3.063
CelebA	16000	512	243.51	150.18	264.12	159.58

H The Effectiveness of Equal Opportunity

According to the definition (2) of equal opportunity, we define EO as a fairness metric that can measure equal opportunity:

$$\begin{aligned}\text{EO} :=&|\mathbb{P}[\hat{y}=1 \mid S=s_i, y=1] - \mathbb{P}[\hat{y}=1 \mid S=s_j, y=1]| \\ =&|\mathbb{P}[\mathcal{T}^*(\mathbf{x})>t \mid \mathbf{x}\in\mathcal{X}_{S=s_i}, y=1] \\ &-\mathbb{P}[\mathcal{T}^*(\mathbf{x})>t \mid \mathbf{x}\in\mathcal{X}_{S=s_j}, y=1]|\end{aligned} \tag{32}$$

In this section, we explore the effectiveness of equal opportunity in unsupervised anomaly detection. According to definition 2 of equal opportunity, we need to measure the fairness on abnormal data. Therefore, we calculate the ADPD and EO (32) only on abnormal data. To determine threshold for calculating EO, we sort the anomaly scores of the training set in ascending order. The threshold t is then set to the pN-th smallest anomaly score, with p fixed at 0.95 for all methods, and N denoting the size of the training set. we report the experimental results in Table 13 and Table 14. From the Table 13 and Table 14, we have the following observations:

Table 13: Detection accuracy (AUC) and fairness (ADPD and EO) on COMPAS. Note that ADPD is measured only on abnormal data, which makes it comparable with EO.

Methods	Balanced Split			skewed Split		
	AUC ↑	Fairness ↓		AUC ↑	Fairness ↓	
		ADPD(%)	EO(%)		ADPD(%)	EO(%)
LOF	59.23	12.99	14.82	57.25	12.91	12.47
Deep SVDD	57.58	18.97	8.99	60.24	23.82	12.54
FarconVAE+LOF	48.72	6.13	3.37	50.10	**4.84**	**2.41**
FarconVAE+Deep SVDD	50.48	6.69	**1.54**	48.83	6.02	**3.23**
FairOD	61.05	6.51	**1.49**	58.86	10.87	5.82
Deep Fair SVDD	62.71	7.46	8.99	61.05	11.61	11.74
CFAD	62.27	10.55	14.63	60.87	16.50	6.69
Ex-FairAD (Ours)	**63.11**	**5.56**	4.35	**66.44**	**4.88**	5.34
Im-FairAD (Ours)	**63.89**	**4.70**	4.13	**66.58**	8.95	5.09

- In the transition from a balanced split to a skewed split, it can be observed that equal opportunity, as measured by ADPD and EO, exhibits significantly larger values in most cases which means a skewed split tends to introduce more unfairness compared to a balanced split, which is consistent with the observation from results on normal and overall test set. This indicates that a skewed split poses a more intractable problem for fairness than a balanced split.
- FairOD demonstrate superior equal opportunity on the COMPAS dataset compared to Deep SVDD, a fairness-unaware AD method. This observation suggests that unsupervised anomaly detection methods have the potential to ensure equal opportunity to some extents, especially when guided by reasonable assumptions (e.g., Assumption 2 proposed in this paper). However, on the Adult dataset, FairOD and CFAD exhibit poorer equal opportunity than Deep SVDD in both balanced and skewed splits, which indicates that existing fairness-aware unsupervised AD methods are unable to maintain equal opportunity effectively across different data domain.
- Compared to all baselines, our methods (Ex-FairAD and Im-FairAD) achieve better detection accuracy (AUC) while maintaining comparable or even better equal opportunity (ADPD and EO) in most cases. This observation supports the reasonability and practicality of Assumption 2 for our methods.

Table 14: Detection accuracy (AUC) and fairness metrics (ADPD and EO) on Adult. Note that ADPD is measured only on abnormal data, which makes it comparable with EO.

Methods	Balanced Split			skewed Split		
	AUC ↑	Fairness ↓		AUC ↑	Fairness ↓	
		ADPD(%)	EO(%)		ADPD(%)	EO(%)
LOF	58.09	26.38	2.89	59.14	28.50	27.05
Deep SVDD	60.47	10.95	6.85	61.04	21.81	5.92
FarconVAE+LOF	53.92	7.89	**2.20**	50.56	8.09	4.85
FarconVAE+Deep SVDD	55.05	8.28	6.68	51.21	7.26	9.64
FairOD	70.63	13.57	15.29	58.30	27.13	31.29
Deep Fair SVDD	62.45	7.24	8.63	60.22	9.35	**3.89**
CFAD	66.89	22.38	42.29	60.28	36.41	59.89
Ex-FairAD (Ours)	**73.49**	**3.78**	6.33	**69.56**	**4.97**	8.74
Im-FairAD (Ours)	**72.05**	**5.85**	**2.73**	**69.34**	**3.60**	**4.38**

I Experiments on Contaminated Training Set

In our main experiments, all methods including the baselines and ours focus on the standard setting of unsupervised anomaly detection [Ruff et al., 2018, Cai and Fan, 2022, Fu et al., 2024], that is the training set consists of only normal samples. However, in real scenarios, a small number of unknown abnormal samples may be mixed in the training set. Based on this consideration, we added $1\%(\text{abnormal}/\text{normal})$ anomalous samples to the balanced training set and keep the test set unchanged. The related results are provided in Table 15, where the detection accuracy of almost all methods has a slight decrease and our proposed methods still achieve better or comparable detection accuracy and fairness in comparison to all baselines.

Table 15: The results on contaminated training set of COMPAS and Adult.

Methods	COMPAS			Adult		
	AUC(%) ↑	APDP(%) ↓		AUC(%) ↑	APDP(%) ↓	
		normal	all		normal	all
LOF	54.70	5.28	5.79	49.83	**1.07**	2.01
Deep SVDD	**62.41**	26.91	18.23	62.70	5.69	7.53
FarconVAE+LOF	46.15	4.53	**1.78**	51.58	3.43	2.28
FarconVAE+Deep SVDD	52.82	2.73	**1.83**	51.36	2.16	2.54
FairOD	54.43	**2.20**	**3.07**	69.76	9.28	8.85
Deep Fair SVDD	60.09	9.50	8.65	62.02	**2.07**	3.74
CFAD	60.81	11.43	9.05	67.98	6.80	15.44
Ex-FairAD (Ours)	**61.67**	2.82	5.66	**73.10**	4.77	**1.76**
Im-FairAD (Ours)	61.10	**2.38**	4.95	**71.80**	5.20	**1.31**

Table 16: The performance changes on COMPAS with the increasing of contamination rate.

cr	Ex-FairAD			Im-FairAD		
	AUC (%)	normal (ADPD %)	all (ADPD %)	AUC (%)	normal (ADPD %)	all (ADPD %)
0.00	63.11	4.38	4.54	63.89	4.16	4.76
0.05	58.84	4.33	4.55	58.88	4.11	4.72
0.10	55.89	3.98	3.45	55.80	3.70	4.16
0.15	54.48	4.19	3.11	55.01	3.89	3.51
0.20	54.41	4.53	2.49	55.04	4.20	2.95

In addition, we conduct experiments on COMPAS (balanced splitting) to explore how fairness changes as the contamination rate $cr \in \{0.05, 0.10, 0.15, 0.20\}$ increases. The related results are reported in Table 16. In terms of fairness, we observe that the fairness (ADPD) of normal data exhibits minor fluctuations within a small range. However, the fairness across the entire test set (including normal and abnormal samples) shows the declining trends (better fairness) as cr increases. This phenomenon aligns with our optimization objective: as the number of abnormal samples in the training set grows, the model would achieve improved group fairness among the groups of abnormal data.

J Experiments on Text dataset

To explore the effectiveness of the proposed method on other data types, We conduct experiments on text data SST_sentiment_fairness_data (from HuggingFace[15]) with gender as protected attribute. In this experiment, we use BERT(bert-large-uncased) Devlin et al. [2019] to extract embeddings (dim=1024) and adopt balanced splitting for two groups. The related results are reported in Table 17.

Table 17: The results on SST with gender as protected attribute.

	AUC(%)	normal	all
FairOD	42.28	8.10	5.57
Deep Fair SVDD	62.23	12.57	9.48
Ex-FairAD (Ours)	62.79	6.99	5.17
Im-FairAD (Ours)	63.64	6.68	5.03

Existing studies on fairness-aware anomaly detection in graph data, such as FairGAD Neo et al. [2024] and DEFEND Chang et al. [2024], adopt a transductive learning paradigm where the training set and test set are identical. This setting differs from the learning paradigm (inductive learning) followed by the proposed method. We expect to extend the proposed framework to more data types in the future work.

K Limitations

In this work, we propose a compact distributional transformation for anomaly detection, where Sinkhorn distance [Cuturi, 2013] is used for measuring the distance between distributions by their finite samples. The time complexity of Sinkhorn distance $\mathcal{O}(n^2)$ causes a relative high computational cost for large-scale data in training stage. The calculation of Sinkhorn distance is not involved in the inference stage.

[15]https://huggingface.co/datasets/fatmaElsafoury2022/SST_sentiment_fairness_data