FairOD: Fairness-aware Outlier Detection

Shubhranshu Shekhar
shubhras@andrew.cmu.edu
Carnegie Mellon University
Pittsburgh, PA, USA

Neil Shah
nshah@snap.com
Snap Inc.
Seattle, WA, USA

Leman Akoglu
lakoglu@andrew.cmu.edu
Carnegie Mellon University
Pittsburgh, PA, USA

ABSTRACT

Fairness and Outlier Detection (OD) are closely related, as it is exactly the goal of OD to spot rare, minority samples in a given population. However, when being a minority (as defined by protected variables, such as race/ethnicity/sex/age) does not reflect positive-class membership (such as criminal/fraud), OD produces unjust outcomes. Surprisingly, fairness-aware OD has been almost untouched in prior work, as fair machine learning literature mainly focuses on supervised settings. Our work aims to bridge this gap. Specifically, we develop desiderata capturing well-motivated fairness criteria for OD, and systematically formalize the fair OD problem. Further, guided by our desiderata, we propose FairOD, a fairness-aware outlier detector that has the following desirable properties: FairOD (1) exhibits treatment parity at test time, (2) aims to flag equal proportions of samples from all groups (i.e. obtain group fairness, via statistical parity), and (3) strives to flag truly high-risk samples within each group. Extensive experiments on a diverse set of synthetic and real world datasets show that FairOD produces outcomes that are fair with respect to protected variables, while performing comparable to (and in some cases, even better than) fairness-agnostic detectors in terms of detection performance.

CCS CONCEPTS

• **Computing methodologies** → *Machine learning algorithms*; **Anomaly detection**.

KEYWORDS

fair outlier detection; outlier detection; anomaly detection; algorithmic fairness; end-to-end detector; deep learning

ACM Reference Format:
Shubhranshu Shekhar, Neil Shah, and Leman Akoglu. 2021. FairOD: Fairness-aware Outlier Detection. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society (AIES '21), May 19–21, 2021, Virtual Event, USA.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3461702.3462517

1 INTRODUCTION

Fairness in machine learning (ML) has received a surge of attention in the recent years. The community has largely focused on designing different notions of fairness [4, 14, 49] mainly tailored towards supervised ML problems [20, 23, 50]. However, perhaps surprisingly, fairness in the context of outlier detection (OD) is vastly understudied. OD is critical for numerous applications in security [21, 51, 55], finance [26, 33, 48], healthcare [8, 36] etc. and is widely used for detection of rare positive-class instances.

Outlier detection for "policing": In such critical systems, OD is often used to flag instances that reflect *riskiness*, which are then "policed" (or audited) by human experts. For example, law enforcement agencies might employ automated surveillance systems in public spaces to spot suspicious individuals based on visual characteristics, who are subsequently stopped and frisked. Alternatively, in the financial domain, analysts can police fraudulent-looking claims, and corporate trust and safety employees can police bad actors on social networks.

Group sample size disparity yields unfair OD: Importantly, outlier detectors are designed exactly to spot rare, *statistical minority* samples[1] with the hope that outlierness reflects riskiness, which prompts their bias against *societal minorities* (as defined by race/ethnicity/sex/age/etc.) as well, since minority group sample size is by definition small.

However, when minority status (e.g. Hispanic) does not reflect positive-class membership (e.g. fraud), OD produces ***unjust outcomes, by overly flagging the instances from the minority groups as outliers.*** This conflation of statistical and societal minorities can become an ethical matter.

Unfair OD leads to disparate impact: What would happen downstream if we did not strive for *fairness-aware* OD given the existence of societal minorities? OD models' inability to distinguish societal minorities (as induced by so-called *protected* variables (*PV*s)), from statistical minorities, contributes to the likelihood of minority group members being flagged as outliers (see Fig. 1). This is further exacerbated by proxy variables which partially-redundantly encode (i.e. correlate with) the *PV*(s), by increasing the number of subspaces in which minorities stand out. The result is *overpolicing* due to over-representation of minorities in OD outcomes. Note that overpolicing the minority group also implies underpolicing the majority group given limited policing capacity and constraints.

Overpolicing can also feed *back* into a system when the policed outliers are used as labels in downstream supervised tasks. Alarmingly, this initially skewed sample (due to unfair OD), may be amplified through a feedback loop via predicting policing where more outliers are identified in more heavily policed groups. Given that OD's use in societal applications has direct bearing on social well-being, ensuring that OD-based outcomes are non-discriminatory is pivotal. This demands the design of fairness-aware OD models, which our work aims to address.

Prior research and challenges: Abundant work on algorithm fairness has focused on supervised ML tasks [6, 23, 50]. Numerous

AIES '21, May 19–21, 2021, Virtual Event, USA

ACM ISBN 978-1-4503-8473-5/21/05…$15.00
https://doi.org/10.1145/3461702.3462517

[1] In this work, the words sample, instance, and observation are used interchangeably throughout text.

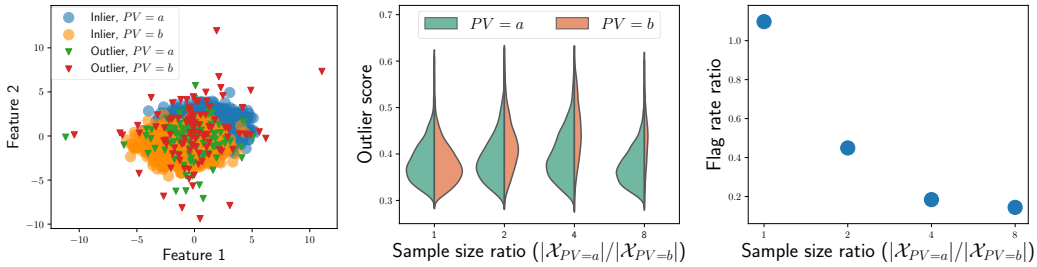


Figure 1: (left) Simulated 2-dim. data with equal sized groups i.e. $|\mathcal{X}_{PV=a}|=|\mathcal{X}_{PV=b}|$. (middle) Group score distributions induced by $PV = a$ and $PV = b$ are plotted by varying the simulated $|\mathcal{X}_{PV=a}|/|\mathcal{X}_{PV=b}|$ ratio. Notice that minority group ($PV = b$) receives larger outlier scores as the size ratio increases. (right) Flag rate ratio of the groups for the varying sample size ratio $|\mathcal{X}_{PV=a}|/|\mathcal{X}_{PV=b}|$. As we increase size disparity, the minority group is "policed" (i.e. flagged) comparatively more.

notions of fairness [4, 49] have been explored in such contexts, each with their own challenges in achieving equitable decisions [14]. In contrast, there is little to no work on addressing fairness in *unsupervised* OD. Incorporating fairness into OD is challenging, in the face of (1) many possibly-incompatible notions of fairness and, (2) the absence of ground-truth outlier labels.

The two works tackling[2] unfairness in the OD literature are by P and Abraham [41] which proposes an ad-hoc procedure to introduce fairness specifically to the LOF algorithm [9], and Zhang and Davidson [54] (concurrent to our work) which proposes an adversarial training based deep SVDD detector. Amongst other issues (see Sec. 5), the approach proposed in [41] invites disparate treatment, necessitating explicit use of *PV at decision time*, leading to taste-based discrimination [15] that is unlawful in several critical applications. On the other hand, the approach in [54] has several drawbacks (see Sec. 5), and in light of unavailable implementation, we include a similar baseline called ARL that we compare against our proposed method.

Alternatively, one could re-purpose existing fair representation learning techniques [7, 18, 52] as well as data preprocessing strategies [19, 27] for subsequent fair OD. However, as we show in Sec. 4 and discuss in Sec. 5, isolating representation learning from the detection task is suboptimal, largely (needlessly) sacrificing detection performance for fairness.

Our contributions: Our work strives to design a fairness-aware OD model to achieve equitable policing across groups and avoid an unjust conflation of statistical and societal minorities. We summarize our main contributions as follows:

(1) **Desiderata & Problem Definition for Fair Outlier Detection:** We identify 5 properties characterizing detection quality and fairness in OD as desiderata for fairness-aware detectors. We discuss their justifiability and achievability, based on which we formally define the (unsupervised) fairness-aware OD problem (Sec. 2).

(2) **Fairness Criteria & New, Fairness-Aware OD Model:** We introduce well-motivated fairness criteria and give mathematical objectives that can be optimized to obey the desiderata. These criteria are universal, in that they can be embedded into the objective of any end-to-end OD model. We propose FAIROD, a new detector which directly incorporates the prescribed criteria into its training. Notably, FAIROD (1) aims to equalize flag rates across groups, achieving group fairness via statistical parity, while (2) striving to flag truly high-risk samples within each group, and (3) avoiding disparate treatment. (Sec. 3)

(3) **Effectiveness on Real-world Data:** We apply FAIROD on several real-world and synthetic datasets with diverse applications such as credit risk assessment and hate speech detection. Experiments demonstrate FAIROD's effectiveness in achieving both fairness goals (Fig. 2) as well as accurate detection (Fig. 6, Sec. 4), significantly outperforming alternative solutions.

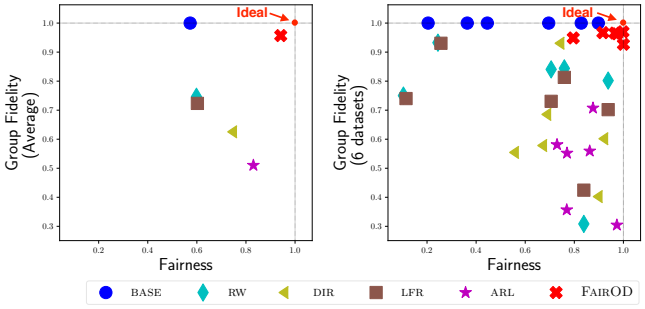


Figure 2: Fairness (statistical parity) vs. GroupFidelity (group-level rank preservation) of baselines and our proposed FAIROD (red cross), (left) averaged across 6 datasets, and (right) on individual datasets. FAIROD outperforms existing solutions (tending towards ideal), achieving fairness while preserving group fidelity from the BASE detector. See Sec. 4 for more details.

Reproducibility: All of our source code and datasets are shared publicly at https://tinyurl.com/fairOD.

2 DESIDERATA FOR FAIR OUTLIER DETECTION

Notation. We are given N samples (also, observations or instances) $\mathcal{X} = \{X_i\}_{i=1}^{N} \subseteq \mathbb{R}^d$ as the input for OD where $X_i \in \mathbb{R}^d$ denotes

[2] [16] aims to *quantify* fairness of OD model outcomes *post hoc*, which thus has a different scope.

Table 1: Frequently used symbols and definitions.

Symbol	Definition
X	d-dimensional feature representation of an observation
Y	true label of an observation, w/ values 0 (inlier), 1 (outlier)
PV	binary protected (or sensitive) variable, w/ groups a (majority), b (minority)
O	detector-assigned label to an observation, w/ value 1 (predicted/flagged outlier)
br_v	base rate of/fraction of ground-truth outliers in group v, i.e. $br_v = P(Y = 1 \mid PV = v)$
fr_v	flag rate of/fraction of flagged observations in group v, i.e $fr_v = P(O = 1 \mid PV = v)$

the feature representation for observation i. Each observation is additionally associated with a binary[3] protected (also, sensitive) variable, $\mathcal{PV} = \{PV_i\}_{i=1}^N$, where $PV_i \in \{a, b\}$ identifies two groups – the majority ($PV_i = a$) group and the minority ($PV_i = b$) group. We use $\mathcal{Y} = \{Y_i\}_{i=1}^N$, $Y_i \in \{0, 1\}$, to denote the *unobserved* ground-truth binary labels for the observations where, for exposition, $Y_i = 1$ denotes an outlier (positive outcome) and $Y_i = 0$ denotes an inlier (negative outcome). We use $O : X \mapsto \{0, 1\}$ to denote the predicted outcome of an outlier detector, and $s : X \mapsto \mathbb{R}$ to capture the corresponding numerical outlier score as the estimate of the outlierness. Thus, $O(X_i), s(X_i)$ respectively indicate predicted outlier label and outlier score for sample X_i. We use $\mathcal{O} = \{O(X_i)\}_{i=1}^N$ and $\mathcal{S} = \{s(X_i)\}_{i=1}^N$ to denote the set of all predicted labels and scores from a given model without loss of generality. Note that we can derive $O(X_i)$ from a simple thresholding of $s(X_i)$. We routinely drop i-subscripts to refer to properties of a single sample without loss of generality. We denote the group *base rate* (or prevalence) of outlierness as $br_a = P(Y = 1|PV = a)$ for the majority group. Finally, we let $fr_a = P(O = 1|PV = a)$ depict the *flag rate* of the detector for the majority group. Similar definitions extend to the minority group with $PV = b$. Table 1 gives a list of the notations frequently used throughout the paper.

Having presented the problem setup and notation, we state our fair OD problem (informally) as follows.

Informal Problem 1 (Fair Outlier Detection). *Given samples $\mathcal{X}$ and protected variable values $\mathcal{PV}$, estimate outlier scores $\mathcal{S}$ and assign outlier labels $\mathcal{O}$, such that*

(i) assigned labels and scores are "fair" w.r.t. the PV, and
(ii) higher scores correspond to higher riskiness encoded by the underlying (unobserved) $\mathcal{Y}$.

How can we design a fairness-aware OD model that is *not biased* against minority groups? What constitutes a "fair" outcome in OD, that is, what would characterize fairness-aware OD? What specific notions of fairness are most applicable to OD?

To approach the problem and address these motivating questions, we first propose a list of desired properties that an ideal fairness-aware detector should satisfy, and whether, in practice, the desired properties can be enforced followed by our proposed solution, FairOD.

2.1 Proposed Desiderata

D1. **Detection effectiveness:** We require an OD model to be accurate at detection, such that the scores assigned to the instances by OD are well-correlated with the ground-truth outlier labels. Specifically, OD benefits the policing effort only when the detection rate (also, precision) is strictly larger than the *base rate* (also, prevalence), that is,

$$P(Y = 1 \mid O = 1) > P(Y = 1) \ . \tag{1}$$

This condition ensures that any policing effort concerted through the employment of an OD model is able to achieve a *strictly larger precision* (LHS) *as compared to random sampling*, where policing via the latter would simply yield a precision that is equal to the prevalence of outliers in the population (RHS) in expectation. Note that our first condition in (1) is related to detection performance, and specifically, the usefulness of OD itself for policing applications.

How-to: We can indirectly control for detection effectiveness via careful feature engineering. Assuming domain experts assist in feature design, it would be reasonable to expect a better-than-random detector that satisfies Eq. (1).

Next, we present *fairness-related* conditions for OD.

D2. **Treatment parity:** OD should exhibit non-disparate treatment that explicitly avoid the use of PV for producing a decision. In particular, OD decisions should obey

$$P(O = 1 \mid X) \ = \ P(O = 1 \mid X, PV = v), \ \forall v \ . \tag{2}$$

In words, the probability that the detector outputs an outlier label O for a given feature vector X remains unchanged even upon observing the value of the PV. In many settings (e.g. employment), explicit PV use is unlawful at inference.

How-to: We can build an OD model using a disparate learning process [34] that uses PV only during the model training phase, but does not require access to PV for producing a decision, hence satisfying treatment parity.

Treatment parity ensures that OD decisions are effectively "blindfolded" to the PV. However, this notion of fairness alone is not sufficient to ensure equitable policing across groups; namely, removing the PV from scope may still allow discriminatory OD results for the minority group (e.g., African American) due to the presence of several other features (e.g., zipcode) that (partially-)redundantly encode the PV. Consequently, by default, OD will use the PV *indirectly*, through access to those correlated proxy features. Therefore, additional conditions follow.

D3. **Statistical parity (SP):** One would expect the OD outcomes to be independent of group membership, i.e. $O \perp\!\!\!\perp PV$. In the context of OD, this notion of fairness (also, demographic parity, group fairness, or independence) aims to enforce that the outlier flag rates are independent of PV and equal across the groups as induced by PV.

Formally, an OD model satisfies statistical parity under a distribution over (X, PV) where $PV \in \{a, b\}$ if

$$\begin{aligned} fr_a = fr_b \ \text{ or equivalently,} \\ P(O = 1|PV = a) \ = \ P(O = 1|PV = b) \ . \end{aligned} \tag{3}$$

SP implies that the fraction of minority (majority) members in the flagged set is the same as the fraction of minority (majority) in the overall population. Equivalently, one can show

$$\begin{aligned} fr_a = fr_b \ (\text{SP}) \iff P(PV = a|O = 1) = P(PV = a) \\ \text{and } P(PV = b|O = 1) = P(PV = b) \ . \end{aligned} \tag{4}$$

[3]For simplicity of presentation, we consider a single, binary protected variable (PV). We discuss extensions to multi-valued PV and multi-attribute PVs in Sec. 3.

The motivation for SP derives from luck egalitarianism [30] – a family of egalitarian theories of distributive justice that aim to counteract the distributive effects of "brute luck". By redistributing equality to those who suffer through no fault of their own choosing, mediated via race, gender, etc., it aims to counterbalance the manifestations of such "luck". Correspondingly, SP ensures equal flag rates across PV groups, eliminating such group-membership bias. Therefore, it merits incorporation in OD since OD results are used for policing or auditing by human experts in downstream applications.

How-to: We could enforce SP during OD model learning by comparing the distributions of the predicted outlier labels O amongst groups, and update the model to ensure that these output distributions match across groups.

SP, however, is not sufficient to ensure both equitable *and* accurate outcomes as it permits so-called "laziness" [4]. Being an unsupervised quantity that is agnostic to the ground-truth labels $\mathcal{Y}$, SP could be satisfied while producing decisions that are arbitrarily inaccurate for any or all of the groups. In fact, an extreme scenario would be random sampling; where we select a certain fraction of the given population uniformly at random and flag all the sampled instances as outliers. As evident via Eq. (4), this entirely random procedure would achieve SP (!). The outcomes could be worse – that is, not only inaccurate (put differently, as accurate as random) but also unfair for only *some* group(s) – when OD flags mostly the true outliers from one group while flagging randomly selected instances from the other group(s), leading to discrimination *despite* SP. Therefore, additional criteria is required to explicitly penalize "laziness," aiming to not only flag *equal fractions* of instances across groups but also those *true outlier* instances from both groups.

D4. **Group fidelity (also, Equality of Opportunity):** It is desirable that the *true* outliers are equally likely to be assigned higher scores, and in turn flagged, regardless of their membership to any group as induced by PV. We refer to this notion of fairness as group fidelity, which steers OD outcomes toward being faithful to the ground-truth outlier labels equally across groups, obeying the following condition

$$P(O = 1|Y = 1, PV = a) \;=\; P(O = 1|Y = 1, PV = b)\ . \tag{5}$$

Mathematically, this condition is equivalent to the so-called Equality of Opportunity[4] in the supervised fair ML literature, and is a special case of Separation [23, 49]. In either case, it requires that all PV-induced groups experience the same true positive rate. Consequently, it penalizes "laziness" by ensuring that the true-outlier instances are ranked above (i.e., receive higher outlier scores than) the inliers within each group.

The key caveat here is that (5) is a supervised quantity that requires access to the ground-truth labels $\mathcal{Y}$, which are explicitly unavailable for the *unsupervised* OD task. What is more, various impossibility results have shown that certain fairness criteria, including SP and Separation, are mutually exclusive or incompatible [4], implying that simultaneously satisfying both of these conditions (exactly) is not possible.

[4]Opportunity, because positive-class assignment by a supervised model in many fair ML problems is often associated with a positive outcome, such as being hired or approved a loan.

How-to: The unsupervised OD task does not have access to $\mathcal{Y}$, therefore, group fidelity cannot be enforced directly. Instead, we propose to enforce group-level rank preservation that maintains fidelity to within-group ranking from the BASE model, where BASE is a fairness-agnostic OD model. Our intuition is that rank preservation acts as a proxy for group fidelity, or more broadly Separation, via our assumption that within-group ranking in the BASE model is accurate and top-ranked instances within each group encode the highest risk samples within each group.

Specifically, let π^{BASE} represent the ranking of instances based on BASE OD scores, and let $\pi^{\text{BASE}}_{PV=a}$ and $\pi^{\text{BASE}}_{PV=b}$ denote the group-level ranked lists for majority and minority groups, respectively. Then, the rank preservation is satisfied when $\pi^{\text{BASE}}_{PV=v} = \pi_{PV=v}; \forall v \in \{a, b\}$ where $\pi_{PV=v}$ is the ranking of group-v instances based on outlier scores from our proposed OD model. Group rank preservation aims to address the "laziness" issue that can manifest while ensuring SP; we aim to not lose the within-group detection prowess of the original detector while maintaining fairness. Moreover, since we are using only a proxy for Separation, the mutual exclusiveness of SP and Separation may no longer hold, though we have not established this mathematically.

D5. **Base rate preservation:** The flagged outliers from OD results are often audited and then used as human-labeled data for supervised detection (as discussed in previous section) which can introduce bias through a feedback loop. Therefore, it is desirable that group-level base rates within the flagged population is reflective of the group-level base rates in the overall population, so as to not introduce group bias of outlier incidence downstream. In particular, we expect OD outcomes to ideally obey

$$P(Y = 1|O = 1, PV = a) = br_a\ , \text{ and} \tag{6}$$

$$P(Y = 1|O = 1, PV = b) = br_b\ . \tag{7}$$

Note that group-level base rate within the flagged population (LHS) is mathematically equivalent to group-level precision in OD outcomes, and as such, is also a supervised quantity which suffers the same caveat as in D4, regarding unavailability of $\mathcal{Y}$.

How-to: As noted, $\mathcal{Y}$ is not available to an unsupervised OD task. Importantly, provided an OD model satisfies D1 and D3, we show that it cannot simultaneously also satisfy D5, i.e. per-group equal base rate in OD results (flagged observations) and in the overall population.

Claim 1. *Detection effectiveness:* $P(Y = 1|O = 1) > P(Y = 1)$ *and SP:* $P(O = 1|PV = a) = P(O = 1|PV = b)$ *jointly imply that* $P(Y = 1|O = 1, PV = v) > P(Y = 1|PV = v), \exists v$.

Proof. We prove the claim in Appendix[5] A.1. □

Claim 1 shows an incompatibility and states that, provided D1 and D3 are satisfied, the base rate in the flagged population cannot be equal to (but rather, is an overestimate of) that in the overall population for *at least one of the groups*. As such, base rates in OD outcomes cannot be reflective of their true values. Instead, one may hope for the preservation of the *ratio* of the base rates (i.e. it

[5]https://tinyurl.com/fairOD

is not impossible). As such, a relaxed notion of D5 is to preserve proportional base rates across groups in the OD results, that is,

$$\frac{P(Y=1|O=1,PV=a)}{P(Y=1|O=1,PV=b)} = \frac{P(Y=1|PV=a)}{P(Y=1|PV=b)} \ . \tag{8}$$

Note that ratio preservation still cannot be explicitly enforced as (8) is also label-dependent. Finally we show in Claim 2 that, provided D1, D3 and Eq. (8) are all satisfied, then it entails that the base rate in OD outcomes is an overestimation of the true group-level base rate *for every group*.

Claim 2. *Detection effectiveness:* $P(Y=1|O=1) > P(Y=1)$, *SP:* $P(O=1|PV=a) = P(O=1|PV=b)$, *and Eq.* (8)*:* $\frac{P(Y=1|O=1,PV=a)}{P(Y=1|O=1,PV=b)} = \frac{P(Y=1|PV=a)}{P(Y=1|PV=b)}$ *jointly imply* $P(Y=1|PV=v, O=1) > P(Y=1|PV=v), \forall v$.

Proof. We prove the claim in Appendix A.2. □

Claim 1 and Claim 2 indicate that if we have both (*i*) better-than-random precision (D1) and (*ii*) SP (D3), interpreting the base rates in OD outcomes for downstream learning tasks would not be meaningful, as they would not be reflective of true population base rates. Due to both these incompatibility results, and also feasibility issues given the lack of $\mathcal{Y}$, we leave base rate preservation – despite it being a desirable property – out of consideration.

2.2 Problem Definition

Based on the definitions and enforceable desiderata, our fairness-aware OD problem is formally defined as follows:

Problem 1 (Fairness-Aware Outlier Detection). *Given samples* $\mathcal{X}$ *and protected variable values* $\mathcal{PV}$*, estimate outlier scores* $\mathcal{S}$ *and assign outlier labels* $\mathcal{O}$*, to achieve*

(i) $P(Y=1|O=1) > P(Y=1)$, *[Detection effectiveness]*

(ii) $P(O \mid X, PV=v) = P(O \mid X),\ \forall v \in \{a,b\}$, *[Treatment parity]*

(iii) $P(O=1|PV=a) = P(O=1|PV=b)$, *[Statistical parity]*

(iv) $\pi^{\text{BASE}}_{PV=v} = \pi_{PV=v}$, $\forall v \in \{a,b\}$*, where* BASE *is a fairness-agnostic detector.* *[Group fidelity proxy]*

Given a dataset along with *PV* values, the goal is to design an OD model that builds on an existing BASE OD model and satisfies the criteria (*i*)–(*iv*), following the proposed desiderata D1 – D4.

2.3 Caveats of a Simple Approach

A simple yet naïve fairness-aware OD approach to address Problem 1 can be designed as follows:

(1) Obtain ranked lists $\pi^{\text{BASE}}_{PV=a}$ and $\pi^{\text{BASE}}_{PV=b}$ from BASE, and
(2) Flag top instances as outliers from each ranked list at equal fraction such that
$P(O=1|PV=a) = P(O=1|PV=b), PV \in \{a,b\}$

This approach fully satisfies (*iii*) and (*iv*) in Problem 1 by design, as well as (*i*) given suitable features. However, it explicitly suffers from *disparate treatment*.

3 FAIRNESS-AWARE OUTLIER DETECTION

In this section, we describe our proposed FairOD – an unsupervised, fairness-aware, end-to-end OD model that embeds our proposed learnable (i.e. optimizable) fairness constraints into an existing BASE OD model. The key features of our model are that FairOD aims for equal flag rates across groups (statistical parity), and encourages correct top group ranking (group fidelity), while not requiring *PV* for decision-making on new samples (non-disparate treatment). As such, it aims to target the proposed desiderata D1 – D4 as described in Sec. 2.

3.1 Base Framework

Our proposed OD model instantiates a deep-autoencoder (AE) framework for the base outlier detection task. However, we remark that the fairness regularization criteria introduced by FairOD can be plugged into any end-to-end *optimizable* anomaly detector, such as one-class support vector machines [46], deep anomaly detector [11], variational AE for OD [3], and deep one-class classifiers [45]. Our choice of AE as the BASE OD model stems from the fact that AE-inspired methods have been shown to be state-of-the-art outlier detectors [13, 37, 56] and that our fairness-aware loss criteria can be optimized in conjunction with the objectives of such models. The main goal of FairOD is to incorporate our proposed notions of fairness into an end-to-end OD model, irrespective of the choice of the BASE model family.

AE consists of two main components: an encoder $G_E : X \in \mathbb{R}^d \mapsto Z \in \mathbb{R}^m$ and a decoder $G_D : Z \in \mathbb{R}^m \mapsto X \in \mathbb{R}^d$. $G_E(X)$ encodes the input X to a hidden vector (also, code) Z that preserves the important aspects of the input. Then, $G_D(Z)$ aims to generate X', a reconstruction of the input from the hidden vector Z. Overall, the AE can be written as $G = G_D \circ G_E$, such that $G(X) = G_D\,(G_E\,(X))$. For a given AE based framework, the outlier score for X is computed using the reconstruction error as

$$s(X) = \|X - G(X)\|_2^2 \ . \tag{9}$$

Outliers tend to exhibit large reconstruction errors because they do not conform to to the patterns in the data as coded by an auto-encoder, hence the use of reconstruction errors as outlier scores [2, 42, 47]. This scoring function is general in that it applies to many reconstruction-based OD models, which have different parameterizations of the reconstruction function G. We show in the following how FairOD regularizes the reconstruction loss from BASE through fairness constraints that are conjointly optimized during the training process. The BASE OD model optimizes the following

$$\mathcal{L}_{\text{BASE}} = \sum_{i=1}^{N} \|X_i - G(X_i)\|_2^2 \tag{10}$$

and we denote its outlier scoring function as $s^{\text{BASE}}(\cdot)$.

3.2 Fairness-aware Loss Function

We begin with designing a loss function for our OD model that optimizes for achieving SP and group fidelity by introducing regularization to the BASE objective criterion. Specifically, FairOD

minimizes the following loss:

$$\mathcal{L} = \alpha \underbrace{\mathcal{L}_{\text{BASE}}}_{\text{Reconstruction}} + (1-\alpha) \underbrace{\mathcal{L}_{SP}}_{\text{Statistical Parity}} + \gamma \underbrace{\mathcal{L}_{GF}}_{\text{Group Fidelity}} \tag{11}$$

where $\alpha \in (0,1)$ and $\gamma > 0$ are hyperparameters which govern the balance between different fairness criteria and reconstruction quality in the loss function.

The first term in Eq. (11) is the objective for learning the reconstruction (based on BASE model family) as given in Eq. (10), which quantifies the goodness of the encoding Z via the squared error between the original input and its reconstruction generated from Z. The second component in Eq. (11) corresponds to regularization introduced to enforce the fairness notion of independence, or statistical parity (SP) as given in Eq. (4). Specifically, the term seeks to minimize the absolute correlation between the outlier scores $\mathcal{S}$ (used for producing predicted labels O) and protected variable values $\mathcal{PV}$. $\mathcal{L}_{SP}$ is given as

$$\mathcal{L}_{SP} = \left| \frac{\left(\sum_{i=1}^{N} s(X_i) - \mu_s\right)\left(\sum_{i=1}^{N} PV_i - \mu_{PV}\right)}{\sigma_s \; \sigma_{PV}} \right| \tag{12}$$

where $\mu_s = \frac{1}{N}\sum_{i=1}^{N} s(X_i)$, $\sigma_s = \frac{1}{N}\sum_{i=1}^{N}(s(X_i) - \mu_s)^2$, $\mu_{PV} = \frac{1}{N}\sum_{i=1}^{N} PV_i$, and $\sigma_{PV} = \frac{1}{N}\sum_{i=1}^{N}(PV_i - \mu_{PV})^2$.

We adapt this absolute correlation loss from [6], which proposed its use in a supervised setting with the goal of enforcing statistical parity. As [6] mentions, while minimizing this loss does not guarantee independence, it performs empirically quite well and offers stable training. We observe the same in practice; it leads to minimal associations between OD outcomes and the protected variable (see details in Sec. 4).

Finally, the third component of Eq. (11) emphasizes that FAIROD should maintain fidelity to within-group rankings from the BASE model (penalizing "laziness"). We set up a listwise learning-to-rank objective in order to enforce group fidelity. Our goal is to train FAIROD such that it reflects the within-group rankings based on $s^{\text{BASE}}(\cdot)$ from BASE. To that end, we employ a listwise ranking loss criterion that is based on the well-known Discounted Cumulative Gain (DCG) [25] measure, often used to assess ranking quality in information retrieval tasks such as search. For a given ranked list, DCG is defined as

$$\text{DCG} = \sum_{r} \frac{2^{rel_r} - 1}{\log_2(1+r)}$$

where rel_r depicts the relevance of the item ranked at the r^{th} position. In our setting, we use the outlier score $s^{\text{BASE}}(X)$ of an instance X to reflect its relevance since we aim to mimic the group-level ranking by BASE. As such, DCG per group can be re-written as

$$\text{DCG}_{PV=v} = \sum_{X_i \in \mathcal{X}_{PV=v}} \frac{2^{s^{\text{BASE}}(X_i)} - 1}{\log_2\left(1 + \sum_{X_k \in \mathcal{X}_{PV=v}} \mathbb{1}[s(X_i) \le s(X_k)]\right)}$$

where $\mathcal{X}_{PV=a}$ and $\mathcal{X}_{PV=b}$ would respectively denote the set of observations from majority and minority groups, and $s(X)$ is the estimated outlier score from our FAIROD model under training.

A key challenge with DCG is that it is not differentiable, as it involves ranking (sorting). Specifically, the sum term in the denominator uses the (non-smooth) indicator function $\mathbb{1}(\cdot)$ to obtain the position of instance i as ranked by the estimated outlier scores. We circumvent this challenge by replacing the indicator function by the (smooth) sigmoid approximation, following [44]. Then, the group fidelity loss component $\mathcal{L}_{GF}$ is given as

$$\mathcal{L}_{GF} = \sum_{v \in \{a,b\}} \left(1 - \sum_{X_i \in \mathcal{X}_{PV=v}} \frac{2^{s^{\text{BASE}}(X_i)} - 1}{\text{DNM}}\right) \tag{13}$$

$$\text{DNM} = \log_2\left(1 + \sum_{X_k \in \mathcal{X}_{PV=v}} \text{sigm}(s(X_k) - s(X_i))\right) \cdot IDCG_{PV=v},$$

$\text{sigm}(x) = \frac{\exp(-cx)}{1+\exp(-cx)}$ is the sigmoid function where $c > 0$ is the scaling constant, and, $IDCG_{PV=v} = \sum_{j=1}^{|\mathcal{X}_{PV=v}|}((2^{s^{\text{BASE}}(X_j)} - 1)/\log_2(1+j))$ is the ideal (hence I), i.e. largest DCG value attainable for the respective group. Note that IDCG can be computed per group apriori to model training via BASE outlier scores alone, and serves as a normalizing constant in Eq. (13).

Note that having trained our model, scoring instances does not require access to the value of their PV, as PV is only used in Eq. (12) and (13) for training purposes. At test time, the anomaly score of a given instance X is computed simply via Eq. (9). Thus, FAIROD also fulfills the desiderata on treatment parity.

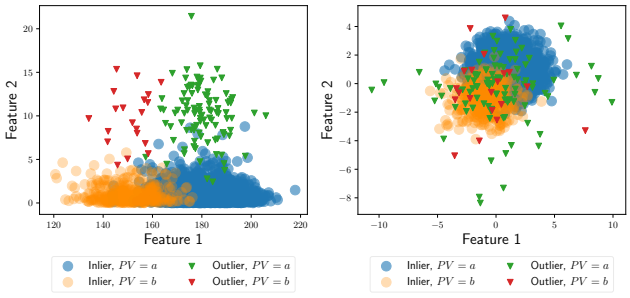
Optimization and Hyperparameter Tuning. We optimize the parameters of FAIROD by minimizing the loss function given in Eq. (11) by using the built-in Adam optimizer [29] implemented in PyTorch.

FAIROD comes with two tunable hyperparameters, α and γ. We define a grid for these and pick the configuration that achieves the best balance between SP and our proxy quantity for group fidelity (based on group-level ranking preservation). Note that both of these quantities are unsupervised (i.e., do not require access to ground-truth labels), therefore, FAIROD model selection can be done in a completely unsupervised fashion. We provide further details about hyperparameter selection in Sec. 4.

Generalizing to Multi-valued and Multiple Protected Attributes.
Multi-valued PV. FAIROD generalizes beyond binary PV, and easily applies to settings with multi-valued, specifically categorical PV such as race. Recall that $\mathcal{L}_{SP}$ and $\mathcal{L}_{GF}$ are the loss components that depend on PV. For a categorical PV, $\mathcal{L}_{GF}$ in Eq. (13) would simply remain the same, where the outer sum goes over all unique values of the PV. For $\mathcal{L}_{SP}$, one could one-hot-encode (OHE) the PV into multiple variables and minimize the correlation of outlier scores with each variable additively. That is, an outer sum would be added to Eq. (12) that goes over the new OHE variables encoding the categorical PV.
Multiple PVs. FAIROD can handle multiple different PVs simultaneously, such as race and gender, since the loss components Eq. (12) and Eq. (13) can be used additively for each PV. However, the caveat to additive loss is that it would only enforce fairness with respect to each individual PV, and yet may not exhibit fairness for the *joint* distribution of protected variables [28]. Even when additive extension may not be ideal, we avoid modeling multiple protected variables as a single PV that induces groups based on values from the cross-product of available values across all PVs. This is because partitioning of the data based on cross-product may yield many small groups, which could cause instability in learning and poor generalization.

Table 2: Summary statistics of real-world and synthetic datasets used for evaluation.

Dataset	N	d	PV	PV = b	$\lvert\mathcal{X}_{\text{PV=a}}\rvert/\lvert\mathcal{X}_{\text{PV=b}}\rvert$	% outliers	Labels
Adult	25262	11	gender	*female*	4	5	{income $\leq 50K$, income $> 50K$}
Credit	24593	1549	age	*age* ≤ 25	4	5	{paid, delinquent}
Tweets	3982	10000	racial dialect	*African-American*	4	5	{normal, abusive}
Ads	1682	1558	simulated	1	4	5	{non-ad, ad}
Synth1	2400	2	simulated	1	4	5	$\{0, 1\}$
Synth2	2400	2	simulated	1	4	5	$\{0, 1\}$



(a) Synth1 (b) Synth2

Figure 3: Synthetic datasets. See Appendix B.1 for the details of the data generating process.

4 EXPERIMENTS

Our proposed FairOD is evaluated through extensive experiments on a set of synthetic datasets as well as diverse real-world datasets. In this section, we present dataset description and the experimental setup, followed by key evaluation questions and results.

4.1 Dataset Description

Table 2 gives an overview of the datasets used in evaluation. A brief summary follows, with details on generative process of synthetic data and detailed descriptions in Appendix B.1.

4.1.1 Synthetic. We illustrate the efficacy of FairOD on two synthetic datasets, Synth1 and Synth2. These datasets present scenarios that mimic real-world settings, where we may have features that are uncorrelated with the outcome labels but partially correlated with the *PV* (see Fig. 3a), or features which are correlated both to outcome labels and *PV* (see Fig. 3b).

4.1.2 Real-world. We experiment on 4 real-world datasets from diverse domains that have various types of PV: specifically gender, age, and race (see Table 2).

4.2 Baselines

We compare FairOD to two classes of baselines: (*i*) a fairness-agnostic base detector that aims to solely optimize for detection performance, and (*ii*) preprocessing methods that aim to correct for bias in the underlying distribution and generate a dataset obfuscating the *PV*.

Base detector model:

- base: A deep anomaly detector that employs an autoencoder neural network. The reconstruction error of the autoencoder is used as the anomaly score. base omits the protected variable from model training.

Preprocessing based methods:

- rw [27]: A preprocessing approach that assigns weights to observations in each group differently to counterbalance the under-representation of minority samples.
- dir [19] A preprocessing approach that edits feature values such that protected variables can not be predicted based on other features in order to increase group fairness. It uses *repair_level* as a hyperparameter, where 0 indicates no repair, and the larger the value gets, the more obfuscation is enforced.
- lfr: This baseline is based on [52] that aims to find a latent representation of the data while obfuscating information about protected variables. In our implementation, we omit the classification loss component during representation learning. It uses two hyperparameters – A_z to control for SP, and A_x to control for the quality of representation.
- arl: This is based on [7] that finds new latent representations by employing an adversarial training process to remove information about the protected variables. In our implementation, we use reconstruction error in place of the classification loss. arl uses λ to control for the trade-off between accuracy (in our implementation, reconstruction quality) and obfuscating protected variable. This baseline optimizes an objective similar to that proposed in [54] which substitutes SVDD loss for reconstruction loss.

The OD task proceeds the preprocessing, where we employ the base detector on the modified data transformed or learned by each of the preprocessing based baselines. We do not compare to the LOF-based fair detector in [41] as it exhibits disparate treatment and is inapplicable in settings that we consider.

Hyperparameters The hyperparameter settings for the competing methods are detailed in Appendix C.

4.3 Evaluation

We design experiments to answer the following questions:

- **[Q1] Fairness:** How well does FairOD (a) achieve fairness as compared to the baselines, and (b) retain the within-group ranking from base?
- **[Q2] Fairness-accuracy trade-off:** How accurately are the outliers detected by FairOD as compared to fairness-agnostic base detector?

- **[Q3] Ablation study:** How do different elements of FairOD influence group fidelity and detector fairness?

4.3.1 *Evaluation Measures.*

Fairness. Fairness is measured in terms of statistical parity. We use flag-rate ratio $r = \frac{P(O=1|PV=a)}{P(O=1|PV=b)}$ which measures the statistical fairness of a detector based on the predicted outcome where $P(O=1|PV=a)$ is the flag-rate of the *majority* group and $P(O=1|PV=b)$ is the flag-rate of the *minority* group. We define Fairness $= \min(r, 1/r) \in [0, 1]$. For a maximally fair detector, Fairness $= 1$ as $r = 1$.

GroupFidelity. We use the Harmonic Mean (HM) of per-group NDCG to measure how well the group ranking of BASE detector is preserved in the fairness-aware detectors. HM between two scalars p and q is defined as $1/(\frac{1}{p}+\frac{1}{q})$. We use HM to report GroupFidelity since it is (more) sensitive to lower values (than e.g. arithmetic mean); as such, it takes large values when *both* of its arguments have large values. We define GroupFidelity $=$ HM($NDCG_{PV=a}$, $NDCG_{PV=b}$), where

$$NDCG_{PV=a} = \sum_{i=1}^{|\mathcal{X}_{PV=a}|} \frac{2^{s^{\text{BASE}}(X_i)} - 1}{\log_2(1 + \sum_{k=1}^{|\mathcal{X}_{PV=a}|} \mathbb{1}(s(X_i) \le s(X_k))) \cdot IDCG},$$

$|\mathcal{X}_{PV=a}|$ is the number of instances in group with $PV = a$, $\mathbb{1}(cond)$ is the indicator function that evaluates to 1 if *cond* is true and 0 otherwise, $s(X_i)$ is the predicted score of the fairness-aware detector, $s^{\text{BASE}}(X_i)$ is the outlier score from BASE detector and $IDCG = \sum_{j=1}^{|\mathcal{X}_{PV=a}|} \frac{2^{s^{\text{BASE}}(X_j)}-1}{\log_2(j+1)}$. GroupFidelity ≈ 1 indicates that group ranking from the BASE detector is well preserved.

Top-k Rank Agreement. We also measure how well the final ranking of the method aligns with the purely performance-driven BASE detector, as BASE optimizes only for reconstruction error. We compute top-k rank agreement as the Jaccard set similarity between the top-k observations as ranked by two methods. Let $\pi^{\text{BASE}}_{[1:k]}$ denote the top-k of the ranked list based on outlier scores $s^{\text{BASE}}(X_i)$'s, and $\pi^{detector}_{[1:k]}$ be the top-k of the ranked list for competing methods where $detector \in$ {RW, DIR, LFR, ARL, FairOD }. Then the measure is given as Top-k Rank Agreement $= |\pi^{\text{BASE}}_{[1:k]} \cap \pi^{detector}_{[1:k]}| / |\pi^{\text{BASE}}_{[1:k]} \cup \pi^{detector}_{[1:k]}|$.

AUC-ratio and AP-ratio. Finally, we consider supervised parity measures based on ground-truth labels, defined as the ratio of ROC AUC and Average Precision (AP) performances across groups; AUC-ratio $= \text{AUC}_{PV=a}/\text{AUC}_{PV=b}$ and AP-ratio $= \text{AP}_{PV=a}/\text{AP}_{PV=b}$.

[Q1] Fairness

In Fig. 2 (presented in Introduction), FairOD is compared against BASE, as well as all the preprocessing baselines across datasets. The methods are evaluated using the best configuration of each method[6] on each dataset. The best hyperparameters for FairOD are the ones for which GroupFidelity and Fairness[7] are closest to the "ideal" point as indicated in Fig. 2.

[6]In Appendix D, for all methods and all datasets, we report detailed values for different metrics for each PV induced group.

[7]Note that we can do model selection in this manner without access to any labels, since both are unsupervised measures.

In Fig. 2 (left), the average of Fairness and GroupFidelity for each method over datasets is reported. FairOD achieves 9× and 5× improvement in Fairness as compared to BASE method and the nearest competitor, respectively. For FairOD, Fairness is very close to 1, while at the same time the group ranking from the BASE detector is well preserved where GroupFidelity also approaches 1. FairOD dominates the baselines (see Fig. 2 (right)) as it is on the Pareto frontier of GroupFidelity and Fairness. Here, each point on the plot represents an evaluated dataset. Notice that FairOD preserves the group ranking while achieving SP consistently across datasets. Fig. 4 reports Top-k

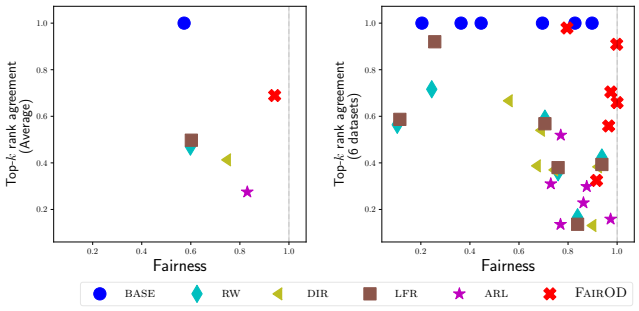


Figure 4: (left) FairOD achieves the best Top-k Rank Agreement compared to the competitors (BASE is shown for reference) in addition to the best overall Fairness, across datasets on average, and (right) measures are shown on individual datasets.

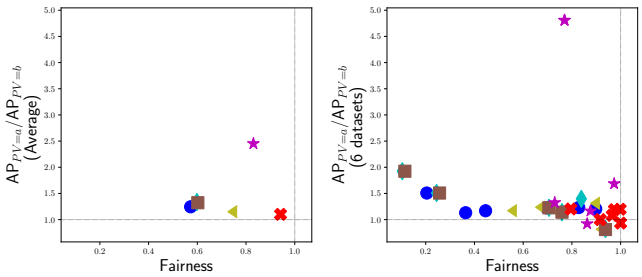
Rank Agreement (computed at top-5% of ranked lists) of each method evaluated across datasets. The agreement measures the degree of alignment of the ranked results by a method with the fairness-agnostic BASE detector. In Fig. 4 (left), as averaged over datasets, FairOD achieves better rank agreement as compared to the competitors. In Fig. 4 (right), FairOD approaches ideal statistical parity across datasets while achieving better rank agreement with the BASE detector. Note that FairOD does not strive for a perfect Top-k Rank Agreement (=1) with BASE, since BASE is shown to fall short with respect to our desired fairness criteria. Our purpose in illustrating it is to show that the ranked list by FairOD is not drastically different from BASE, which simply aims for detection performance.

Next we evaluate the competing methods against supervised (label-aware) fairness metrics. Note that FairOD does not (by design) optimize for these supervised fairness measures. Fig. 5a evaluates the methods against Fairness and label-aware parity criterion – specifically, group AP-ratio (ideal AP-ratio is 1). FairOD approaches ideal Fairness as well as ideal AP-ratio across all datasets. FairOD outperforms the competitors on the averaged metrics over datasets (Fig. 5a (left)) and across individual datasets (Fig. 5a (right)). In contrast, the preprocessing baselines are up to ~5× worse than FairOD over AP-ratio measure across datasets. Fig. 5b reports evaluation of methods against Fairness and another label-aware parity measure – specifically, group AUC-ratio (ideal AUC-ratio = 1). As shown in Fig. 5b (left), FairOD outperforms all the baselines in expectation as averaged over all datasets. Further, in Fig. 5b (right), FairOD consistently approaches ideal AUC-ratio across datasets, while the preprocessing baselines are up to ~1.9× worse comparatively.

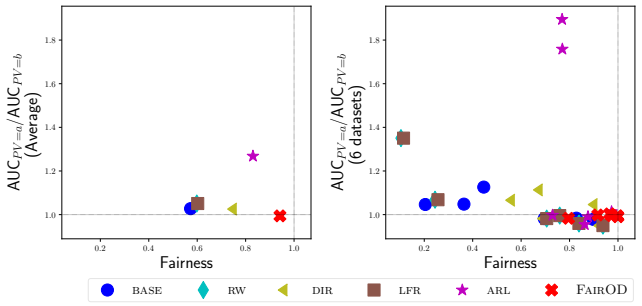
We note that impressively, FairOD approaches parity across different supervised fairness measures despite not being able to optimize for label-aware criteria explicitly.

[Q2] Fairness-accuracy trade-off

In the presence of ground-truth outlier labels, the performance of a detector could be measured using a ranking accuracy metric such as area under the ROC curve (ROC AUC).



(a) Fairness vs. AP-ratio



(b) Fairness vs. AUC-ratio

Figure 5: FairOD outperforms all competitors on averaged label-aware parity metrics over datasets (left) and for individual datasets (right): we report Fairness against (a) Group AP-ratio and (b) Group AUC-ratio.

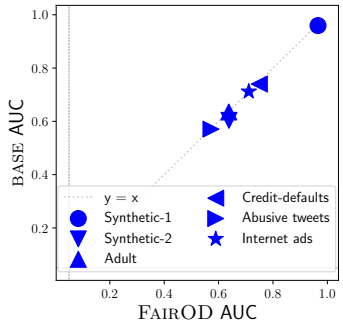


Figure 6: ROCAUC of FairOD vs. BASE: FairOD matches the performance of BASE detector, while enforcing fairness criteria (maintaining good performance *with* fairness).

In Fig. 6, we compare the AUC performance of FairOD to that of BASE detector for all datasets. Notice that each of the symbols (i.e. datasets) is slightly below the diagonal line indicating that FairOD achieves equal or sometimes even better (!) detection performance as compared to BASE. The explanation is that since FairOD enforces SP and does not allow "laziness", it addresses the issue of falsely or unjustly flagged minority samples by BASE, thereby, improving detection performance.

From Fig. 6, we conclude that FairOD does not trade-off detection performance much, and in some cases it even improves performance by eliminating false positives from the minority group, as compared to the performance-driven, fairness-agnostic BASE.

[Q3] Ablation study

Finally, we evaluate the effect of various components in the design of FairOD's objective. Specifically, we compare to the results of two relaxed variants of FairOD, namely FairOD-L and FairOD-C, described as follows.

- FairOD-L: We retain only the SP-based regularization term from FairOD objective along with the reconstruction error. This relaxation of FairOD is partially based on the method proposed in [6], which minimizes the correlation between model prediction and group membership to the PV. In FairOD-L, the reconstruction error term substitutes the classification loss used in the optimization criteria in [6]. Note that FairOD-L concerns itself with only group fairness to attain SP which may suffer from "laziness" (hence, FairOD-L) (see Sec. 2).
- FairOD-C: Instead of training with NDCG-based group fidelity regularization, FairOD-C utilizes a simpler regularization, aiming to minimize the correlation (hence, FairOD-C) of the outlier scores per-group with the corresponding scores from BASE detector. Thus, FairOD-C attempts to maintain group fidelity over the entire ranking within a group, in contrast to FairOD's NDCG-based regularization which emphasizes the quality of the ranking at the top. Specifically, FairOD-C substitutes $\mathcal{L}_{GF}$ in Eq. (11) with the following.

$$\mathcal{L}_{GF} = -\sum_{v\in\{a,b\}} \left| \frac{\left(\sum_{X_i\in X_{PV=v}} s(X_i)-\mu_s\right)\left(\sum_{X_i\in X_{PV=v}} s^{\text{BASE}}(X_i)-\mu_{s^{\text{BASE}}}\right)}{\sigma_s\ \sigma_{s^{\text{BASE}}}} \right|$$

where $v \in \{a, b\}$, and $\mu_{s^{\text{BASE}}}$, $\sigma_{s^{\text{BASE}}}$ are defined similar to μ_s, σ_s respectively.

Fig. 7 presents the comparison of FairOD and its variants. In Fig. 7 (left), we report the evaluation against GroupFidelity and Fairness averaged over datasets, and in Fig. 7 (right), the metrics are reported for each individual dataset. FairOD-L approaches SP and achieves comparable Fairness to FairOD except on one dataset as shown in Fig. 7 (right). This results in lower Fairness compared to FairOD when averaged over datasets as shown in Fig. 7 (left). However, FairOD-L suffers with respect to GroupFidelity as compared to FairOD. This is because FairOD-L may randomly flag instances to achieve SP since it does not include any group ranking criterion in its objective. On the other hand, FairOD-C improves Fairness when compared to BASE, while under-performing on the majority of datasets compared to FairOD across metrics. Since FairOD-C tries to preserve group-level ranking, it trades-off on Fairness as

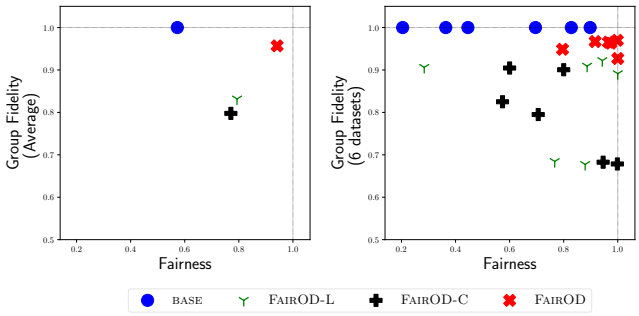


Figure 7: FAIROD compared to its variants FAIROD-L and FAIROD-C across datasets, to evaluate the effect of different regularization components. FAIROD-L achieves comparable Fairness to FAIROD while compromising GroupFidelity. FAIROD-C improves Fairness as compared to BASE, but is ill-suited to optimizing for GroupFidelity.

measured against FAIROD-L. We also observe that FAIROD outperforms FAIROD-C on all datasets, which suggests that preserving the entire group-level rankings may be a harder task than preserving top of the rankings; it is also a needlessly ill-suited one since what matters for outlier detection is the top of the ranking.

5 RELATED WORK

A majority of work on algorithmic fairness focuses on supervised learning problems. We refer to [5, 39] for an excellent overview. We organize related work in three sub-areas related to fairness in outlier detection, fairness-aware representation learning, and data de-biasing strategies.

Outlier Detection and Fairness Outlier detection (OD) is a well-studied problem in the literature [2, 12, 22], and finds numerous applications in high-stakes domains like health-care [36], security [21], and finance [43]. However, only a few studies focus on OD's fairness aspects. P and Sam Abraham [41] propose a detector called FairLOF that applies an ad-hoc procedure to introduce fairness specifically to the LOF algorithm [9]. This approach suffers from several drawbacks: (i) it mandates disparate treatment, which may be at times infeasible/unlawful, e.g. in domains like housing or employment, (ii) only prioritizes SP, which as we discussed in Sec. 2, can permit "laziness," (iii) it is heuristic, and cannot be concretely optimized end-to-end. Concurrent to our work, Zhang and Davidson [54] introduce a deep SVDD based detector employing adversarial training to obfuscate protected group membership, similar to our ARL baseline. This approach also has issues: (i) it only considers SP, and (ii) it suffers from well-known instability due to adversarial training [10, 31, 38]. A related work by Davidson and Ravi [17] focuses on quantifying the fairness of an OD model's outcomes after detection, which thus has a different scope.

Fairness-aware Representation Learning Several works aim to map input samples to an embedding space, where the representations are indistinguishable across groups [35, 52]. Most recently, adversarial training has been used to obfuscate PV association in representations while preserving accurate classification [1, 7, 18, 38, 53]. Most of these methods are supervised. Substituting classification or label-aware loss terms with unsupervised reconstruction loss can plausibly extend such methods to OD (by using masked representations as inputs to a detector). However, a common shortcoming is that statistical parity (SP) is employed as the primary fairness criterion in these methods, e.g. in fair principal component analysis [40] and fair variational autoencoder [35]. To summarize, fair representation learning techniques exhibit two key drawbacks for unsupervised OD: (i) they only employ SP, which may be prone to "laziness", and (ii) isolating embedding from detection makes embedding oblivious to the task itself, and therefore can yield poor detection performance (as shown in experiments in Sec. 4).

Strategies for Data De-Biasing Some of the popular de-biasing methods [27, 32] draw from topics in learning with imbalanced data [24] that employ under- or over-sampling or point-wise weighting of the instances based on the class label proportions to obtain balanced data. These methods apply preprocessing to the data in a manner that is agnostic to the subsequent or downstream task and consider only the fairness notion of SP, which is prone to "laziness."

6 CONCLUSIONS

Although fairness in machine learning has become increasingly prominent in recent years, fairness in the context of unsupervised outlier detection (OD) has received comparatively little study. OD is an integral data-driven task in a variety of domains including finance, healthcare and security, where it is used to inform and prioritize auditing measures. Without careful attention, OD as-is can cause unjust flagging of *societal minorities* (w.r.t. race, sex, etc.) because of their standing as *statistical minorities*, when minority status does not indicate positive-class membership (crime, fraud, etc.). This unjust flagging can propagate to downstream supervised classifiers and further exacerbate the issues. Our work tackles the problem of fairness-aware outlier detection. Specifically, we first introduce guiding desiderata for, and concrete formalization of the fair OD problem. We next present FAIROD, a fairness-aware, principled end-to-end detector which addresses the problem, and satisfies several appealing properties: (i) *detection effectiveness:* it is effective, and maintains high detection accuracy, (ii) *treatment parity:* it does not suffer disparate treatment at decision time, (iii) *statistical parity:* it maintains group fairness across minority and majority groups, and (iv) *group fidelity:* it emphasizing flagging of truly high-risk samples within each group, aiming to curb detector "laziness". Finally, we show empirical results across diverse real and synthetic datasets, demonstrating that our approach achieves fairness goals while providing accurate detection, significantly outperforming unsupervised fair representation learning and data de-biasing based baselines. We hope that our expository work yields further studies in this area.

ACKNOWLEDGMENTS

This research is sponsored by NSF CAREER 1452425. In addition, we thank Dimitris Berberidis for helping with the early development of the ideas and the preliminary code base. Conclusions expressed in this material are those of the authors and do not necessarily reflect the views, expressed or implied, of the funding parties.

REFERENCES

[1] Tameem Adel, Isabel Valera, Zoubin Ghahramani, and Adrian Weller. 2019. One-network adversarial fairness. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 33. 2412–2420.

[2] Charu C Aggarwal. 2015. Outlier analysis. In *Data mining*. Springer, 237–263.

[3] Jinwon An and Sungzoon Cho. 2015. Variational autoencoder based anomaly detection using reconstruction probability. *Special Lecture on IE* 2, 1 (2015), 1–18.

[4] Solon Barocas, Moritz Hardt, and Arvind Narayanan. 2017. Fairness in machine learning. *NIPS Tutorial* 1 (2017).

[5] Solon Barocas, Moritz Hardt, and Arvind Narayanan. 2019. *Fairness and Machine Learning*. fairmlbook.org. http://www.fairmlbook.org.

[6] Alex Beutel, Jilin Chen, Tulsee Doshi, Hai Qian, Allison Woodruff, Christine Luu, Pierre Kreitmann, Jonathan Bischof, and Ed H Chi. 2019. Putting fairness principles into practice: Challenges, metrics, and improvements. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*. 453–459.

[7] Alex Beutel, Jilin Chen, Zhe Zhao, and Ed H Chi. 2017. Data decisions and theoretical implications when adversarially learning fair representations. *arXiv preprint arXiv:1707.00075* (2017).

[8] Marcel Bosc, Fabrice Heitz, Jean-Paul Armspach, Izzie Namer, Daniel Gounot, and Lucien Rumbach. 2003. Automatic change detection in multimodal serial MRI: application to multiple sclerosis lesion evolution. *NeuroImage* 20, 2 (2003), 643–656.

[9] Markus M Breunig, Hans-Peter Kriegel, Raymond T Ng, and Jörg Sander. 2000. LOF: identifying density-based local outliers. In *Proceedings of the 2000 ACM SIGMOD international conference on Management of data*. 93–104.

[10] George Cevora. 2020. Fair Adversarial Networks. *arXiv preprint arXiv:2002.12144* (2020).

[11] Raghavendra Chalapathy, Aditya Krishna Menon, and Sanjay Chawla. 2018. Anomaly Detection using One-Class Neural Networks. *arXiv preprint arXiv:1802.06360* (2018).

[12] Varun Chandola, Arindam Banerjee, and Vipin Kumar. 2009. Anomaly detection: A survey. *ACM computing surveys (CSUR)* 41, 3 (2009), 1–58.

[13] Jinghui Chen, Saket Sathe, Charu Aggarwal, and Deepak Turaga. 2017. Outlier detection with autoencoder ensembles. In *Proceedings of the 2017 SIAM international conference on data mining*. SIAM, 90–98.

[14] Sam Corbett-Davies and Sharad Goel. 2018. The Measure and Mismeasure of Fairness: A Critical Review of Fair Machine Learning. *CoRR* abs/1808.00023 (2018). http://dblp.uni-trier.de/db/journals/corr/corr1808.html#abs-1808-00023

[15] Sam Corbett-Davies and Sharad Goel. 2018. The measure and mismeasure of fairness: A critical review of fair machine learning. *arXiv:1808.00023* (2018).

[16] Ian Davidson and Selvan Suntiha Ravi. 2020. A framework for determining the fairness of outlier detection. In *Proceedings of the 24th European Conference on Artificial Intelligence (ECAI2020)*, Vol. 2029.

[17] Ian Davidson and Selvan Suntiha Ravi. 2020. A framework for determining the fairness of outlier detection. In *Proceedings of the 24th European Conference on Artificial Intelligence (ECAI2020)*, Vol. 2029.

[18] Harrison Edwards and Amos Storkey. 2015. Censoring representations with an adversary. *arXiv preprint arXiv:1511.05897* (2015).

[19] Michael Feldman, Sorelle A Friedler, John Moeller, Carlos Scheidegger, and Suresh Venkatasubramanian. 2015. Certifying and removing disparate impact. In *proceedings of the 21th ACM SIGKDD*. 259–268.

[20] Naman Goel, Mohammad Yaghini, and Boi Faltings. 2018. Non-discriminatory machine learning through convex fairness criteria. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*. 116–116.

[21] Prasanta Gogoi, DK Bhattacharyya, Bhogeswar Borah, and Jugal K Kalita. 2011. A survey of outlier detection methods in network anomaly identification. *Comput. J.* 54, 4 (2011), 570–588.

[22] Manish Gupta, Jing Gao, Charu C Aggarwal, and Jiawei Han. 2013. Outlier detection for temporal data: A survey. *IEEE TKDE* 26, 9 (2013), 2250–2267.

[23] Moritz Hardt, Eric Price, and Nati Srebro. 2016. Equality of opportunity in supervised learning. In *Advances in neural information processing systems*. 3315–3323.

[24] Haibo He and Edwardo A Garcia. 2009. Learning from imbalanced data. *IEEE Transactions on knowledge and data engineering* 21, 9 (2009), 1263–1284.

[25] K. Järvelin and J. Kekäläinen. 2002. Cumulated gain-based evaluation of IR techniques. *ACM Transactions on Information Systems (TOIS)* 20 (2002), 422–446.

[26] Justin M Johnson and Taghi M Khoshgoftaar. 2019. Medicare fraud detection using neural networks. *Journal of Big Data* 6, 1 (2019), 63.

[27] Faisal Kamiran and Toon Calders. 2012. Data preprocessing techniques for classification without discrimination. *Knowledge and Information Systems* 33, 1 (2012), 1–33.

[28] Michael Kearns, Seth Neel, Aaron Roth, and Zhiwei Steven Wu. 2018. Preventing fairness gerrymandering: Auditing and learning for subgroup fairness. In *International Conference on Machine Learning*. PMLR, 2564–2572.

[29] Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014).

[30] Carl Knight. 2009. *Luck Egalitarianism: Equality, Responsibility, and Justice*. Edinburgh University Press. http://www.jstor.org/stable/10.3366/j.ctt1r2483

[31] Naveen Kodali, Jacob Abernethy, James Hays, and Zsolt Kira. 2017. On convergence and stability of gans. *arXiv preprint arXiv:1705.07215* (2017).

[32] Emmanouil Krasanakis, Eleftherios Spyromitros-Xioufis, Symeon Papadopoulos, and Yiannis Kompatsiaris. 2018. Adaptive sensitive reweighting to mitigate bias in fairness-aware classification. In *Proceedings of the 2018 World Wide Web Conference*. 853–862.

[33] Meng-Chieh Lee, Yue Zhao, Aluna Wang, Pierre Jinghong Liang, Leman Akoglu, Vincent S Tseng, and Christos Faloutsos. 2020. AutoAudit: Mining Accounting and Time-Evolving Graphs. *arXiv preprint arXiv:2011.00447* (2020).

[34] Zachary Lipton, Julian McAuley, and Alexandra Chouldechova. 2018. Does mitigating ML's impact disparity require treatment disparity?. In *Advances in Neural Information Processing Systems*. 8125–8135.

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

[36] Wei Luo and Marcus Gallagher. 2010. Unsupervised DRG upcoding detection in healthcare databases. In *2010 IEEE ICDM Workshops*. IEEE, 600–605.

[37] Yunlong Ma, Peng Zhang, Yanan Cao, and Li Guo. 2013. Parallel auto-encoder for efficient outlier detection. In *2013 IEEE International Conference on Big Data*. IEEE, 15–17.

[38] David Madras, Elliot Creager, Toniann Pitassi, and Richard Zemel. 2018. Learning adversarially fair and transferable representations. *arXiv preprint arXiv:1802.06309* (2018).

[39] Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. 2019. A survey on bias and fairness in machine learning. *arXiv preprint arXiv:1908.09635* (2019).

[40] Matt Olfat and Anil Aswani. 2019. Convex formulations for fair principal component analysis. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 33. 663–670.

[41] Deepak P and Savitha Sam Abraham. 2020. Fair Outlier Detection. arXiv:2005.09900 [cs.LG]

[42] Guansong Pang, Chunhua Shen, Longbing Cao, and Anton van den Hengel. 2020. Deep learning for anomaly detection: A review. *arXiv preprint arXiv:2007.02500* (2020).

[43] Clifton Phua, Vincent Lee, Kate Smith, and Ross Gayler. 2010. A comprehensive survey of data mining-based fraud detection research. *arXiv preprint arXiv:1009.6119* (2010).

[44] Tao Qin, Tie-Yan Liu, and Hang Li. 2010. A general approximation framework for direct optimization of information retrieval measures. *Information retrieval* 13, 4 (2010), 375–397.

[45] Lukas Ruff, Robert A. Vandermeulen, Nico Görnitz, Lucas Deecke, Shoaib A. Siddiqui, Alexander Binder, Emmanuel Müller, and Marius Kloft. 2018. Deep One-Class Classification. In *Proceedings of the 35th International Conference on Machine Learning*, Vol. 80. 4393–4402.

[46] Bernhard Schölkopf, John C Platt, John Shawe-Taylor, Alex J Smola, and Robert C Williamson. 2001. Estimating the support of a high-dimensional distribution. *Neural computation* 13, 7 (2001), 1443–1471.

[47] Neil Shah, Alex Beutel, Brian Gallagher, and Christos Faloutsos. 2014. Spotting suspicious link behavior with fbox: An adversarial perspective. In *2014 IEEE International Conference on Data Mining*. IEEE, 959–964.

[48] Véronique Van Vlasselaer, Cristián Bravo, Olivier Caelen, Tina Eliassi-Rad, Leman Akoglu, Monique Snoeck, and Bart Baesens. 2015. APATE: A novel approach for automated credit card transaction fraud detection using network-based extensions. *Decision Support Systems* 75 (2015), 38–48.

[49] Sahil Verma and Julia Rubin. 2018. Fairness definitions explained. In *2018 IEEE/ACM International Workshop on Software Fairness (FairWare)*. IEEE, 1–7.

[50] Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rodriguez, and Krishna P Gummadi. 2017. Fairness beyond disparate treatment & disparate impact: Learning classification without disparate mistreatment. In *Proceedings of the 26th international conference on world wide web*. 1171–1180.

[51] Sultan Zavrak and Murat İskefiyeli. 2020. Anomaly-based intrusion detection from network flow features using variational autoencoder. *IEEE Access* 8 (2020), 108346–108358.

[52] Rich Zemel, Yu Wu, Kevin Swersky, Toni Pitassi, and Cynthia Dwork. 2013. Learning fair representations. In *International Conference on Machine Learning*. 325–333.

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

[54] Hongjing Zhang and Ian Davidson. 2020. Towards Fair Deep Anomaly Detection. *arXiv preprint arXiv:2012.14961* (2020).

[55] Jiong Zhang and Mohammad Zulkernine. 2006. Anomaly based network intrusion detection with unsupervised outlier detection. In *2006 IEEE International Conference on Communications*, Vol. 5. IEEE, 2388–2393.

[56] Chong Zhou and Randy C Paffenroth. 2017. Anomaly detection with robust deep autoencoders. In *Proceedings of the 23rd ACM SIGKDD*. 665–674.