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

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, 15, 51] mainly tailored towards supervised ML problems [21, 24, 52]. However, surprisingly, fairness in the context of outlier detection (OD) is vastly understudied. OD is critical for numerous applications in security [22, 53, 57], finance [27, 34, 50], healthcare [9, 38] 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 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, 24, 52]. Numerous
notions of fairness [4, 51] have been explored in such contexts, each with their own challenges in achieving equitable decisions [15]. 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 unfairness in the OD literature are by P and Abraham [43] which proposes an ad-hoc procedure to introduce fairness specifically to the LOF algorithm [10], and Zhang and Davidson [56] (concurrent to our work) which proposes an adversarial training based deep SVDD detector. Amongst other issues (see Sec. 5), the approach proposed in [43] invites disparate treatment, necessitating explicit use of PV at decision time, leading to taste-based discrimination [16] that is unlawful in several critical applications. On the other hand, the approach in [56] 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, 19, 54] as well as data preprocessing strategies [20, 28] 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

Figure 1: (left) Simulated 2-dim. data with equal sized groups i.e. $|X_{PV=a}| = |X_{PV=b}|$. (middle) Group score distributions induced by $PV = a$ and $PV = b$ are plotted by varying the simulated $|X_{PV=a}|/|X_{PV=b}|$ 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 $|X_{PV=a}|/|X_{PV=b}|$. As we increase size disparity, the minority group is “policed” (i.e. flagged) comparatively more.

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.

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.

Reproducibility: All of our source code and datasets are shared publicly at https://tinyurl.com/fairOD.
the feature representation for observation \( i \). Each observation is additionally associated with a binary\(^3\) protected (also, sensitive) variable, \( \mathcal{P}V \), where \( \mathcal{P}V \in \{ a, b \} \) identifies two groups - the majority (\( \mathcal{P}V = a \)) group and the minority (\( \mathcal{P}V = b \)) group. We use \( \mathcal{Y} = \{ \mathcal{Y}_i \}_{i=1}^{N} \), \( \mathcal{Y}_i \in \{ 0, 1 \} \), to denote the unobserved ground-truth binary labels for the observations where, for exposition, \( \mathcal{Y}_i = 1 \) denotes an outlier (positive outcome) and \( \mathcal{Y}_i = 0 \) 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 outlierliness. Thus, \( O(X_i) \) respectively indicate outlier label and outlier score for sample \( X_i \). We use \( O = \{ O(X_i) \}_{i=1}^{N} \) and \( S = \{ s(X_i) \}_{i=1}^{N} \) 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 \( b_{ra} = P(\mathcal{Y} = 1 | \mathcal{P}V = a) \) for the majority group. Finally, we let \( f_{ra} = P(O = 1 | \mathcal{P}V = a) \) depict the flag rate of the detector for the majority group. Similar definitions extend to the minority group with \( \mathcal{P}V = b \). 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 \( X \) and protected variable values \( \mathcal{P}V \), estimate outlier scores \( S \) and assign outlier labels \( O \), such that

(i) assigned labels and scores are "fair" w.r.t. the \( \mathcal{P}V \), 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,

\[
P(Y = 1 | O = 1) > P(Y = 1).
\] (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 \( \mathcal{P}V \) for producing a decision. In particular, OD decisions should obey

\[
P(O = 1 | X) = P(O = 1 | X, \mathcal{P}V = a), \forall \mathcal{P}V .
\] (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 \( \mathcal{P}V \). In many settings (e.g., employment), explicit \( \mathcal{P}V \) use is unlawful at inference.

**How-to:** We can build an OD model using a disparate learning process [36] that uses \( \mathcal{P}V \) only during the model training phase, but does not require access to \( \mathcal{P}V \) for producing a decision, hence satisfying treatment parity.

Treatment parity ensures that OD decisions are effectively “blindfolded” to the \( \mathcal{P}V \). However, this notion of fairness alone is not sufficient to ensure equitable policing across groups; namely, removing the \( \mathcal{P}V \) 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 \( \mathcal{P}V \). Consequently, by default, OD will use the \( \mathcal{P}V \) 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. \( O \perp \mathcal{P}V \). 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 \( \mathcal{P}V \) and equal across the groups as induced by \( \mathcal{P}V \).

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

\[
f_{ra} = f_{rb} \quad \text{or equivalently,} \quad P(O = 1 | \mathcal{P}V = a) = P(O = 1 | \mathcal{P}V = b). \quad (3)
\]

**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

\[
f_{ra} = f_{rb} \quad (\text{SP}) \iff P(\mathcal{P}V = a | O = 1) = P(\mathcal{P}V = a) \quad \text{and} \quad P(\mathcal{P}V = b | O = 1) = P(\mathcal{P}V = b). \quad (4)
\]
The motivation for SP derives from luck egalitarianism [31] – 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 inliers within each group.

\[ \text{instances are ranked above (i.e., receive higher outlier scores than)} \]

instances from the other group(s), leading to discrimination despite SP. Therefore, additional criteria is required to explicitly penalize the true outliers from one group while flagging randomly selected some instances that is, not only inaccurate (put differently, as accurate as 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

\[ P(O = 1|Y = 1, PV = a) = P(O = 1|Y = 1, PV = b). \]  

(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 [24, 51]. 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 \( \pi_{\text{BASE}}^{o} \) represent the ranking of instances based on base OD scores, and let \( \pi_{PV=b}^{o} \) and \( \pi_{PV=a}^{o} \) denote the group-level ranked lists for majority and minority groups, respectively. Then, the rank preservation is satisfied when \( \pi_{PV=a}^{o} = \pi_{PV=b}^{o}, \forall o \in \{a, b\} \) where \( \pi_{PV=a}^{o} \) is the ranking of group-a 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

\[ P(Y = 1|O = 1, PV = a) = br_a, \text{ and } \]  

(6)

\[ P(Y = 1|O = 1, PV = b) = br_b. \]  

(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 \( P(Y = 1|O = 1, PV = a) > P(Y = 1|PV = a, \exists o) \).

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,
\[
P(Y = 1|O = 1, PV = a) = P(Y = 1|PV = a) \quad \text{and} \quad P(Y = 1|O = 1, PV = b) = P(Y = 1|PV = b)\, . \tag{8}
\]
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: \(P(Y = 1|O = 1) > P(Y = 1, SP; P(O = 1|PV = a) = P(O = 1|PV = b), \text{ and Eq.\;}(8))\), \(P(Y = 1|O = 1, PV = a) = P(Y = 1|PV = a)\) and Eq. (8): \(P(Y = 1|O = 1, PV = b) = P(Y = 1|PV = b)\) jointly imply \(P(Y = 1|PV = a, O = 1) > P(Y = 1|PV = b, O = 0)\).

**Proof.** We prove the claim in Appendix A.2. \(\square\)

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 \(Y\), 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 \(X\) and protected variable values \(PV\), estimate outlier scores \(s\) and assign outlier labels \(O\), to achieve

(i) \(P(Y = 1|O = 1) > P(Y = 1)\) \quad [Detection effectiveness]

(ii) \(P(O = 1|X, PV = a) = P(O = 1|X), \forall O \in \{a, b\}\) \quad [Treatment parity]

(iii) \(P(O = 1|PV = a) = P(O = 1|PV = b)\) \quad [Statistical parity]

(iv) \(s_{\text{BASE}} = s_{\text{PV}}\) \quad \forall O \in \{a, b\}\), where \(BASE\) is a fairness-agnostic detector.\quad [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 \(s_{\text{BASE}}\) and \(s_{\text{PV}}\) 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.

### 3.1 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 [48], deep anomaly detector [12], variational AE for OD [3], and deep one-class classifiers [47]. 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 [14, 39, 58] 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 the patterns in the data as coded by an auto-encoder, hence the use of reconstruction errors as outlier scores [2, 44, 49]. 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 jointly optimized during the training process. The base OD model optimizes the following

\[
L_{\text{BASE}} = \sum_{i=1}^{N} ||X_i - G(X_i)||^2_2 \quad \tag{10}
\]

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:

$$L = \alpha \frac{L_{\text{base}}}{\text{Reconstruction}} + (1 - \alpha) \frac{L_{SP}}{\text{Statistical Parity}} + \gamma \frac{L_{GF}}{\text{Group Fidelity}}$$  \hspace{1cm} (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 $S$ (used for predicting produced labels $O$) and protected variable values $P_{PV}$. $L_{SP}$ is given as

$$L_{SP} = \frac{\left( \sum_{i=1}^{N} s(X_i) - \mu_s \right) \left( \sum_{i=1}^{N} P_{Vi} - \mu_{PV} \right)}{\sigma_s \sigma_{PV}}$$  \hspace{1cm} (12)$$

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

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 $\delta_{\text{base}}$ from base. To that end, we employ a listwise ranking loss criterion that is based on the well-known Discounted Cumulative Gain (DCG) [26] measure, often used to assess ranking quality in information retrieval tasks such as search. For a given ranked list, DCG is defined as

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

where $\text{rel}_r$ depicts the relevance of the item ranked at the $r^{th}$ position. In our setting, we use the outlier score $\delta_{\text{base}}(X)$ 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

$$\text{DCG}_{PV=v} = \sum_{X_v \in X_{PV=v}} \log_2 (1 + \sum_{X_h \in X_{PV=v}} \mathbb{1}[s(X_h) \leq s(X_v)]) - 1$$

where $X_{PV=v}$ and $X_{PV=d}$ 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 $\mathbb{1}()$ 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 [46]. Then, the group fidelity loss component $L_{GF}$ is given as

$$L_{GF} = \sum_{v \in \{a,b\}} \left( 1 - \sum_{X_i \in X_{PV=v}} \frac{2^{\delta_{\text{base}}(X_i)} - 1}{\text{dnm}} \right)$$  \hspace{1cm} (13)$$

$$\text{dnm} = \log_2 (1 + \sum_{X_h \in X_{PV=v}} \text{sigm}(s(X_h) - s(X_v))) \cdot \text{DCG}_{PV=v},$$

where $\text{sigm}(x) = \frac{\exp(+cx)}{\exp(+cx) + \exp(-cx)}$ is the sigmoid function where $c > 0$ is the scaling constant, and, $\text{DCG}_{PV=v} = \frac{1}{N} \sum_{j=1}^{N} (2^{\delta_{\text{base}}(X_j)} - 1) / \log_2(1 + j))$ is the ideal (hence $I$), i.e. largest DCG value attainable for the respective group. Note that DCG 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 [30] implemented in PyTorch. FairOD comes with two tunable hyperparameters, $\alpha$ and $\gamma$. 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 $L_{SP}$ and $L_{GF}$ are the loss components that depend on $PV$. For a categorical $PV$, $L_{GF}$ in Eq. (13) would simply remain the same, where the outer sum goes over all unique values of the $PV$. For $L_{SP}$, 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 $PV$s.** FairOD can handle multiple different $PV$s 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 [29]. 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 $PV$s. 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     | $PV = a$ | % outliers | Labels                                |
|-----------|------|-----|--------|----------|------------|---------------------------------------|
| Adult     | 25262| 11  | gender | female   | 4          | 5          | {income ≤ 50K, income > 50K}          |
| Credit    | 24593| 1549| age    | $\leq 25$| 4          | 5          | {paid, delinquent}                    |
| Tweets    | 3982 | 10000| racial dialect | African-American | 4          | 5          | {normal, abusive}                     |
| Ads       | 1682 | 1558| simulated |        | 1          | 4          | {non-ad, ad}                          |
| Synth1    | 2400 | 2   | simulated |        | 1          | 4          | {0, 1}                                |
| Synth2    | 2400 | 2   | simulated |        | 1          | 4          | {0, 1}                                |

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** [28]: A preprocessing approach that assigns weights to observations in each group differently to counterbalance the under-representation of minority samples.
- **DIR** [20] A preprocessing approach that edits feature values such that protected variables cannot 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 [54] 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 – $A_{x}$ to control for $SP$, and $A_{e}$ 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 $\lambda$ to control for the trade-off between accuracy (in our implementation, reconstruction quality) and obfuscation of protected variable. This baseline optimizes an objective similar to that proposed in [56] 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 [43] 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 $\text{Fairness} = \min(r, 1/r) \in [0, 1]$. For a maximally fair detector, $\text{Fairness} = 1$ 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 $\text{GroupFidelity} = \text{HM}(\text{NDCGPV}_{a=0}, \text{NDCGPV}_{b=1})$, where

$$\text{NDCGPV}_{a=0} = \frac{1}{|\mathcal{X}_{PV=0}|} \sum_{i=1}^{\infty} \frac{2^{\text{bas}(X_i)} - 1}{\log_2(1 + e^{\frac{|\mathcal{X}_{PV=0}|}{2}}(s(X_i) \leq s(X_k))))} \cdot \text{IDCG},$$

$|\mathcal{X}_{PV=0}|$ is the number of instances in group with $PV = a$, $\mathcal{X}_{PV=0}$ is the indicator function that evaluates to 1 if $PV$ 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 $\text{IDCG} = \sum_{i=1}^{\text{Top-k}} \frac{2^{\text{bas}(X_i)} - 1}{\log_2(|\mathcal{X}_{PV=0}|)}$. GroupFidelity $\approx 1$ 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 $\pi_{base}[i:k]$ denote the top-k of the ranked list based on outlier scores $s_{\text{BASE}}(X_i)$’s, and $\pi_{\text{detector}}[i:k]$ be the top-k of the ranked list for competing methods where $\text{detector} \in \{\text{rw, dir, lfr, arl, FairOD}\}$. Then the measure is given as Top-k Rank Agreement $= |\pi_{\text{BASE}}[i:k] \cap \pi_{\text{detector}}[i:k]| / |\pi_{\text{BASE}}[i:k] \cup \pi_{\text{detector}}[i:k]|$.

**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;

$\text{AUC-ratio} = \text{AUC}_{PV=a}/\text{AUC}_{PV=b}$ and $\text{AP-ratio} = \text{AP}_{PV=a}/\text{AP}_{PV=b}$.

**[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 on each dataset. The best hyperparameters for FairOD are the ones for which GroupFidelity and Fairness are closest to the “ideal” point as indicated in Fig. 2.

In Fig. 2 (left), the average of Fairness and GroupFidelity for each method over datasets is reported. FairOD achieves 9x and 5x 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 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 ~5x worse than FairOD over AP-ratio measure across datasets. Fig. 5b is plots 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 ~1.9x 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).

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 "laizziness", 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 "laizziness" (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.

$$\mathcal{L}_{GF} = - \sum_{v \in \{a, b\}} \left( \frac{\sum_{X_i \in X_v} \mathbb{1}_{y_i=v} (X_i - \mu_v)}{\sigma_v} \right) ^{2}$$

where $v \in \{a, b\}$, and $\mu_v^{base}$, $\sigma_v^{base}$ are defined similar to $\mu_s$, $\sigma_s$ 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
5 RELATED WORK

A majority of work on algorithmic fairness focuses on supervised learning problems. We refer to [5, 41] 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, 13, 23], and finds numerous applications in high-stakes domains like health-care [38], security [22], and finance [45]. However, only a few studies focus on OD’s fairness aspects. P and Sam Abraham [43] propose a detector called FairLOF that applies an ad-hoc procedure to introduce fairness specifically to the LOF algorithm [10]. 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 [56] introduce a deep SVDD based detector employing adversarial training to obfuscate protected group membership, similar to our ART baseline. This approach also has issues: (i) it only considers SP, and (ii) it suffers from well-known instability due to adversarial training [11, 32, 40]. A related work by Davidson and Ravi [18] 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 [37, 54]. Most recently, adversarial training has been used to obfuscate PV association in representations while preserving accurate classification [1, 7, 19, 40, 55]. 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 [42] and fair variational autoencoder [37]. 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 [28, 33] draw from topics in learning with imbalanced data [25] 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 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.
A PROOFS

A.1 Proof of Claim 1

Proof. We want OD to exhibit detection effectiveness i.e. \( P(Y = 1 | O = 1) > P(Y = 1) \).

Now, \( P(Y = 1 | O = 1) = P(PV = a | O = 1) \).
\[
P(Y = 1 | PV = a, O = 1) + \]
\[
P(PV = b | O = 1) \].
\[
P(Y = 1 | PV = b, O = 1) \]

Given SP, we have
\[
P(O = 1 | PV = a) = P(O = 1 | PV = b) \]
\[
\implies P(PV = a | O = 1) = P(PV = b), \text{ and} \]
\[
P(PV = b | O = 1) = P(PV = b) \]
Therefore, we have
Now, \( P(Y = 1 | O = 1) = P(PV = a) \).
\[
P(Y = 1 | PV = a, O = 1) + \]
\[
P(PV = b). \] \hspace{1cm} (14)
\[
P(Y = 1 | PV = b, O = 1) \]

Now,
\[
P(Y = 1) = P(PV = a) \cdot P(Y = 1 | PV = a) + \]
\[
P(PV = b) \cdot P(Y = 1 | PV = b) \]
Therefore, if we want \( P(Y = 1 | O = 1) > P(Y = 1) \), then
\[
P(PV = a) \cdot P(Y = 1 | PV = a, O = 1) + \]
\[
P(PV = b) \cdot P(Y = 1 | PV = b, O = 1) \]
\[
> 0 \] \hspace{1cm} (15)
\[
P(PV = a) \cdot P(Y = 1 | PV = a) + \]
\[
P(PV = b) \cdot P(Y = 1 | PV = b) \]
\[
\implies \exists a \in \{a, b\} \text{ s.t. } P(Y = 1 | PV = a, O = 1) \]
\[
> 0 \]
\[
P(Y = 1 | PV = a) \]

\( \square \)

A.2 Proof of Claim 2

Proof. Without loss of generality, assume that \( P(Y = 1 | PV = a, O = 1) > P(Y = 1 | PV = a) \) i.e. (i.e. \( P(Y = 1 | PV = a, O = 1) = K \cdot P(Y = 1 | PV = a) \)); \( K > 1 \), and let \( P(Y = 1 | PV = a) = \frac{P(Y = 1 | PV = a, O = 1)}{P(Y = 1 | PV = a)} = \frac{1}{2} \) then
Case 1: When \( P(Y = 1 | PV = b, O = 1) < P(Y = 1 | PV = b) \)
\[
P(Y = 1 | PV = b, O = 1) < P(Y = 1 | PV = b) \]
\[
\implies P(Y = 1 | PV = b, O = 1) < r \cdot P(Y = 1 | PV = a) \]
\[
\implies P(Y = 1 | PV = b, O = 1) < r \cdot P(Y = 1 | PV = a, O = 1) \]
\[
\implies P(Y = 1 | PV = a, O = 1) > P(Y = 1 | PV = a) \]
This contradicts our assumption that \( P(Y = 1 | PV = b, O = 1) = r \cdot P(Y = 1 | PV = a, O = 1) \), therefore it must be that
\[
P(Y = 1 | PV = b, O = 1) \geq P(Y = 1 | PV = b) \]

Case 2: When \( P(Y = 1 | PV = b, O = 1) = P(Y = 1 | PV = b) \)
\[
P(Y = 1 | PV = b, O = 1) = P(Y = 1 | PV = b) \]
\[
\implies P(Y = 1 | PV = b, O = 1) = r \cdot P(Y = 1 | PV = a) \]
\[
\implies P(Y = 1 | PV = b, O = 1) < r \cdot P(Y = 1 | PV = a, O = 1) \]
\[
\implies P(Y = 1 | PV = a, O = 1) > P(Y = 1 | PV = a) \]
This contradicts our assumption that \( P(Y = 1 | PV = b, O = 1) = r \cdot P(Y = 1 | PV = a, O = 1) \), therefore it must be that
\[
P(Y = 1 | PV = b, O = 1) \geq P(Y = 1 | PV = b) \]

Case 3: When \( P(Y = 1 | PV = b, O = 1) > P(Y = 1 | PV = b) \) i.e.
\[
(P(Y = 1 | PV = b, O = 1) = L \cdot P(Y = 1 | PV = b); L > 1) \]
Now, we know that,
\[
P(Y = 1 | PV = a) \cdot P(Y = 1 | PV = b, O = 1) \]
\[
= P(Y = 1 | PV = b) \cdot P(Y = 1 | PV = a, O = 1) \]
\[
\implies P(Y = 1 | PV = a) \cdot P(Y = 1 | PV = b, O = 1) \]
\[
= P(Y = 1 | PV = b) \cdot K \cdot P(Y = 1 | PV = a) \]
\[
\implies P(Y = 1 | PV = b, O = 1) = K \cdot P(Y = 1 | PV = b) \]
\[
\implies P(Y = 1 | PV = b, O = 1) > P(Y = 1 | PV = b) \]
And, for ratio to be preserved, it must be that \( L = K \).

Hence, enforcing preservation of ratios implies base-rates in flagged observations are larger than their counterparts in the population. \( \square \)

B DATA DESCRIPTION

B.1 Synthetic data

We illustrate the effectiveness of FairOD on two synthetic datasets, namely Synth1 and Synth2 (as illustrated in Fig. 3). These datasets are constructed to present scenarios that mimic real-world settings, where we may have features which are uncorrelated with respect to outcome labels but partially correlated with \( PV \), or features which are correlated both to outcome labels and \( PV \).

- Synth1: In Synth1, we simulate a 2-dimensional dataset comprised of samples \( X = [x_1, x_2] \) where \( x_1 \) is correlated with the protected variable \( PV \), but does not offer any predictive value with respect to ground-truth outlier labels \( Y \), while \( x_2 \) is correlated with these labels \( Y \) (see Fig. 3a). We draw 2400 samples, of which \( PV = a \) (majority) for 2000 points, and \( PV = b \) (minority) for 400 points. 120 (5%) of these points are outliers. \( x_1 \) differs in terms of shifted means, but equal variances, for both majority and minority groups, drawn from a normal distribution for outliers, and an exponential for inliers. The detailed generative process for the data is below, and Fig. 3a shows a visual.
Synth1

Simulate samples $X = \{x_1, x_2\}$ by...

\[ PV \sim \text{Bernoulli}(4/5) \]

\[ Y \sim \text{Bernoulli}(1/20) \]

\[ \begin{align*}
&\text{Normal}(-1, 1.44) \quad \text{if} \quad Y = 0, PV = 1 \\
&\text{Normal}(1, 1.44) \quad \text{if} \quad Y = 0, PV = 0 \\
&2 \times \text{Exponential}(1) (1 - 2 \times \text{Bernoulli}(1/2)) \quad \text{if} \quad Y = 1
\end{align*} \]

\[ x_1 \sim \begin{cases}
\text{Normal}(180, 10) & \text{if} \quad PV = 1 \\
\text{Normal}(150, 10) & \text{if} \quad PV = 0
\end{cases} \quad \text{if} \quad Y = 1, \text{[majority]} \\
\text{Exponential}(1) & \quad \text{if} \quad Y = 0, \text{[outlier]}
\]

\[ x_2 \sim \begin{cases}
\text{Normal}(3, 2) & \text{if} \quad Y = 1, \text{[minority; inlier]} \\
\text{Exponential}(1) & \text{if} \quad Y = 0, \text{[inlier]}
\end{cases} \]

Synth2

Simulate samples $X = \{x_1, x_2\}$ by...

\[ PV \sim \text{Bernoulli}(4/5) \]

\[ Y \sim \text{Bernoulli}(1/20) \]

\[ \begin{align*}
&\text{Normal}(-1, 1.44) \quad \text{if} \quad Y = 0, PV = 1 \\
&\text{Normal}(1, 1.44) \quad \text{if} \quad Y = 0, PV = 0 \\
&2 \times \text{Exponential}(1) (1 - 2 \times \text{Bernoulli}(1/2)) \quad \text{if} \quad Y = 1
\end{align*} \]

\[ x_1 \sim \begin{cases}
\text{Normal}(180, 10) & \text{if} \quad PV = 1 \\
\text{Normal}(150, 10) & \text{if} \quad PV = 0
\end{cases} \quad \text{if} \quad Y = 1, \text{[majority]} \\
\text{Exponential}(1) & \quad \text{if} \quad Y = 0, \text{[inlier]}
\]

\[ x_2 \sim \begin{cases}
\text{Normal}(10, 3) & \text{if} \quad Y = 1, \text{[minority; inlier]} \\
\text{Exponential}(1) & \text{if} \quad Y = 0
\end{cases} \]

B.1.1 Real-world data. We conduct experiments on 4 real-world datasets and select them from diverse domains that have different types of (binary) protected variables, specifically gender, age, and race. Detailed descriptions are as follows.

• Adult [35] (Adult). The dataset is extracted from the 1994 Census database where each data point represents a person. The dataset records income level of an individual along with features encoding personal information on education, profession, investment and family. In our experiments, gender $\in \{\text{male, female}\}$ is used as the protected variable where female represents minority group and high earning individuals who exceed an annual income of 50,000 i.e. annual income $> 50,000$ are assigned as outliers ($Y = 1$). We further downsample female to achieve a male to female sample size ratio of 4:1 and ensure that percentage of outliers remains the same (at 5%) across groups induced by the protected variable.

• Credit-defaults [35] (Credit). This is a risk management dataset from the financial domain that is based on Taiwan’s credit card clients’ default cases. The data records information of credit card customers including their payment status, demographic factors, credit data, historical bill and payments. Customer age is used as the protected variable where age $> 25$ indicates the majority group and age $\leq 25$ indicates the minority group. We assign individuals with delinquent payment status as outliers ($Y = 1$). The age $> 25$ to age $\leq 25$ imbalance ratio is 4:1 and contains 5% outliers across groups induced by the protected variable.

• Abusive Tweets [8] (Tweets). The dataset is a collection of Tweets along with annotations indicating whether a tweet is abusive or not. The data are not annotated with any protected variable by default; therefore, to assign protected variable to each Tweet, we employ the following process: We predict the racial dialect — African-American or Mainstream — of the tweets in the corpus using the language model proposed by [8]. The dialect is assigned to a Tweet only when the prediction probability is greater than 0.7, and then the predicted racial dialect is used as protected variable where African-American dialect represents the minority group. In this setting, abusive tweets are labeled as outliers ($Y = 1$) for the task of flagging abusive content on Twitter. The group sample size ratio of racial dialect = African-American to racial dialect = Mainstream is set to 4:1. We further sample data points to ensure equal percentage (5%) of outliers across dialect groups.

• Internet ads [35] (Ads). This is a collection of possible advertisements on web-pages. The features characterize each ad by encoding phrases occurring in the ad URL, anchor text, alt text, and encoding geometry of the ad image. We assign observations with class label ad as outliers ($Y = 1$) and downsample the data to get an outlier rate of 5%. There exists no demographic information available, therefore we simulate a binary protected variable by randomly assigning each observation to one of two values (i.e. groups) $\in \{0, 1\}$ such that the group sample size ratio is 4:1.

C HYPERPARAMETERS

We choose the hyperparameters of FairOD from $\alpha \in \{0.01, 0.5, 0.9\} \times \gamma \in \{0.01, 0.1, 1.0\}$ by evaluating the Pareto curve for fairness and group fidelity criteria. The base and FairOD methods both use an auto-encoder with two hidden layers. We fix the number of hidden nodes in each layer to 2 if $d \leq 100$, and 8 otherwise. The representation learning methods LFR and ARL use the model configurations as proposed by their authors. The hyperparameter grid for the preprocessing baselines are set as follows: repair_level $\in \{0.0001, 0.01, 0.1, 1.0\}$ for dir, $A_2 \in \{0.0001, 0.001, 0.01, 0.1, 0.9\}$ and $A_1 = 1 - A_2$ for LFR, and $\lambda \in \{0.0001, 0.001, 0.01, 0.1, 0.9\}$ for ARL. We pick the best model for the preprocessing baselines using Fairness as they only optimize for statistical parity. The best BASE model is selected based on reconstruction error through cross validation upon multiple runs with different random seeds.

D SUPPLEMENTAL RESULTS

In this section, we report Flag-rate, GroupFidelity, AUC and AP (see Table 3) for the competing methods on a set of datasets (see Appendix B.1.1) w.r.t groups induced by $PV = v; v \in \{a, b\}$ to supplement the experimental results presented in Sec. 4. Notice that in most cases (see Table 3a through Table 3f), FairOD outperforms the BASE model on label-aware parity metrics (AUC-ratio, AP-ratio) and, furthermore, outperforms BASE on at least one of the performance metrics (e.g. AUC, AP). Fairness need not imply worse OD performance.
Table 3: Evaluation measures are reported for the competing methods on the datasets presented in Appendix B.1.

(a) Synth1

| Method          | Flag-rate | Group Fidelity | AUC   | AP   |
|-----------------|-----------|----------------|-------|------|
|                 | PV = a    | PV = b         | PV = a| PV = b|
| BASE            | 0.0262    | 0.1282         | 1.0   | 1.0  |
| RW              | 0.033     | 0.135          | 0.9299| 0.9309|
| DIR             | 0.0445    | 0.0775         | 0.3953| 0.9281|
| LFR             | 0.0350    | 0.1350         | 0.9299| 0.9309|
| ARL             | 0.0520    | 0.0400         | 0.9136| 0.3955|
| FairOD-L        | 0.0500    | 0.0500         | 0.9639| 0.9671|
| FairOD-C        | 0.0495    | 0.0525         | 0.9149| 0.9295|
|                  | 0.0480    | 0.0600         | 0.8929| 0.9082|
|                  | 0.0499    | 0.0928         | 0.7542| 0.6501|

(b) Synth2

| Method          | Flag-rate | Group Fidelity | AUC   | AP   |
|-----------------|-----------|----------------|-------|------|
|                 | PV = a    | PV = b         | PV = a| PV = b|
| BASE            | 0.0361    | 0.0811         | 1.0   | 1.0  |
| RW              | 0.0205    | 0.1975         | 0.9242| 0.6313|
| DIR             | 0.0465    | 0.0675         | 0.4224| 0.9164|
| LFR             | 0.0205    | 0.1975         | 0.9242| 0.6313|
| ARL             | 0.0520    | 0.0400         | 1.0   | 1.0  |
| FairOD-L        | 0.0500    | 0.0500         | 0.9339| 0.9201|
| FairOD-C        | 0.0430    | 0.0750         | 0.8997| 0.9095|

(c) Adult

| Method          | Flag-rate | Group Fidelity | AUC   | AP   |
|-----------------|-----------|----------------|-------|------|
|                 | PV = a    | PV = b         | PV = a| PV = b|
| BASE            | 0.0358    | 0.0433         | 1.0   | 1.0  |
| RW              | 0.0515    | 0.0391         | 0.8399| 0.8479|
| DIR             | 0.0515    | 0.0391         | 0.9299| 0.9309|
| LFR             | 0.0515    | 0.0391         | 0.8099| 0.8099|
| ARL             | 0.0507    | 0.0444         | 0.9147| 0.5765|
| FairOD-L        | 0.0497    | 0.0511         | 0.9646| 0.9616|
| FairOD-C        | 0.0513    | 0.0403         | 0.9178| 0.9085|
|                  | 0.0527    | 0.0302         | 0.8119| 0.7877|
|                  | 0.6533    | 0.6229         | 0.1872| 0.1435|

(d) Credit

| Method          | Flag-rate | Group Fidelity | AUC   | AP   |
|-----------------|-----------|----------------|-------|------|
|                 | PV = a    | PV = b         | PV = a| PV = b|
| BASE            | 0.0445    | 0.064          | 1.0   | 1.0  |
| RW              | 0.0467    | 0.06627        | 0.8399| 0.8409|
| DIR             | 0.0467    | 0.06627        | 0.8899| 0.8809|
| LFR             | 0.0467    | 0.06627        | 0.7299| 0.7309|
| ARL             | 0.0471    | 0.0645         | 0.5533| 0.6118|
| FairOD          | 0.0468    | 0.0866         | 0.9235| 0.9421|
| FairOD-L        | 0.0475    | 0.0662         | 0.7147| 0.6564|
| FairOD-C        | 0.0467    | 0.0662         | 0.7871| 0.8029|
|                  | 0.7327    | 0.7484         | 0.1333| 0.1091|

(e) Tweets

| Method          | Flag-rate | Group Fidelity | AUC   | AP   |
|-----------------|-----------|----------------|-------|------|
|                 | PV = a    | PV = b         | PV = a| PV = b|
| BASE            | 0.0369    | 0.0105         | 1.0   | 1.0  |
| RW              | 0.0494    | 0.0507         | 0.388 | 0.4178|
| DIR             | 0.0497    | 0.0571         | 0.4082| 0.5312|
| LFR             | 0.0494    | 0.0507         | 0.5583| 0.582 |
| ARL             | 0.0482    | 0.0538         | 0.5432| 0.5762|
| FairOD-L        | 0.0331    | 0.1167         | 0.9337| 0.8986|
| FairOD-C        | 0.0501    | 0.0488         | 0.6753| 0.6903|

(f) Ads

| Method          | Flag-rate | Group Fidelity | AUC   | AP   |
|-----------------|-----------|----------------|-------|------|
|                 | PV = a    | PV = b         | PV = a| PV = b|
| BASE            | 0.0236    | 0.0318         | 1.0   | 1.0  |
| RW              | 0.0491    | 0.0523         | 0.8236| 0.7813|
| DIR             | 0.0491    | 0.0523         | 0.6236| 0.5813|
| LFR             | 0.0491    | 0.0523         | 0.7236| 0.6813|
| ARL             | 0.0499    | 0.0500         | 0.5062| 0.2181|
| FairOD-L        | 0.0683    | 0.0588         | 0.5551| 0.8684|
| FairOD-C        | 0.0499    | 0.0500         | 0.6611| 0.6966|
|                  | 0.7007    | 0.7251         | 0.2636| 0.2455|