FairOD: Fairness-aware Outlier Detection

SHUBHRANSHU SHEKHAR, Heinz College, Carnegie Mellon University
NEIL SHAH, Snap Inc.
LEMAN AKOGLU, Heinz College, Carnegie Mellon University

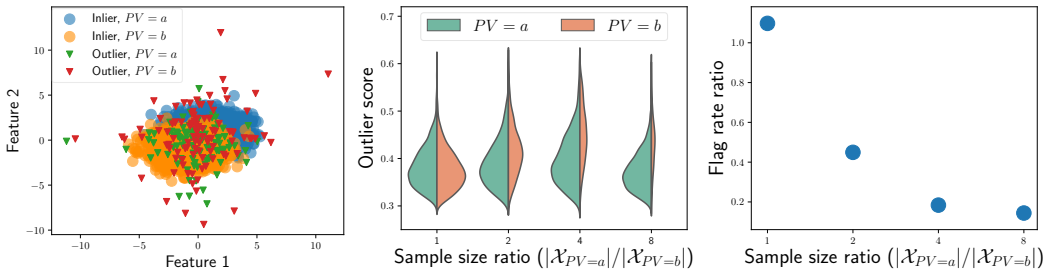
Fairness and Outlier Detection (OD) are closely related, as it is exactly the goal of OD to spot rare, minority samples in a given population. When being a minority (as defined by protected variables, e.g. race/ethnicity/sex/age) does not reflect positive-class membership (e.g. criminal/fraud), however, OD produces unjust outcomes. Surprisingly, fairness-aware OD has been almost untouched in prior work, as fair machine learning literature mainly focus 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, which has the following, desirable properties: FairOD (1) does not employ disparate treatment 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 fraction of 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.

1 INTRODUCTION

Fairness in machine learning (ML) has received a surge of attention in the recent years. The fairness community has largely focused on designing different notions of fairness [5, 16, 67] mainly tailored towards supervised ML problems [24, 27, 69]. However, perhaps surprisingly, fairness in the context of outlier detection (OD) is vastly understudied. OD is critical for numerous applications in security [18, 22, 25, 71, 74], finance [20, 32, 40, 55, 66, 68], healthcare [11, 28, 46, 56, 64] etc. and is widely used for detection of rare positive-class instances such as network intrusion, crime hot-spots, fraudulent transactions, drug abuse, and so on.

Outlier detection for "policing": In high-stakes, critical systems, OD often serves to flag instances that reflect a tendency towards *riskiness* which are then "policed" (or audited) by human experts. The flagged instances typically describe a sparse, low-density or otherwise unusual region in the data, indicating exogeneity of these instances. For example, law enforcement agencies might employ automated surveillance systems in public spaces, such as railway stations, to spot suspicious individuals (based on their e.g. outfit/behavior/demographic characteristics as captured by surveillance cameras) who could pose risk to thousands of commuters. In this scenario, the outliers flagged by such automated systems are those who may be policed (stop and frisk) by law enforcement agents. Notably, policing need not be limited to law enforcement, and rather represents a general auditing process. For example, in the financial domain, analysts can "police" individual claims that are flagged as suspected fraudulent by a detector. The same can be said of trust and safety-focused employees who aim to discover and suspend bad actors and spammy users on social networks.

This research is sponsored by NSF CAREER 1452425. 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.
Authors' addresses: Shubhranshu Shekhar, shubhras@andrew.cmu.edu, Heinz College, Carnegie Mellon University, 4800 Forbes Avenue, Pittsburgh, PA, 15213; Neil Shah, nshah@snap.com, Snap Inc., 2025 1st Ave, Seattle, WA, 98121; Leman Akoglu, lakoglu@andrew.cmu.edu, Heinz College, Carnegie Mellon University.



(a) (b) (c)

Fig. 1. (a) A simulated 2-dimensional data with equal sized groups i.e. $|\mathcal{X}_{PV=a}| = |\mathcal{X}_{PV=b}|$. (b) Score distributions of groups as induced by $PV = a$ and $PV = b$ are plotted by varying the $|\mathcal{X}_{PV=a}|/|\mathcal{X}_{PV=b}|$ ratio on the simulated data. Notice that minority group ($PV = b$) starts receiving larger scores as sample size ratio increases. (c) Flag rate ratio of the groups for the varying sample size ratio $|\mathcal{X}_{PV=a}|/|\mathcal{X}_{PV=b}|$. As we increase sample size disparity, minority group is "policed" (i.e. flagged) comparatively more.

Group sample size disparity yields unfair OD: Importantly, outlier detectors may be biased against *societal minority groups* (as defined by race/ethnicity/sex/age/etc.) since the sample size of a minority group, by definition, is smaller than that of the majority group, prompting minority group instances to stand out as rare, *statistical minorities* as well. Outlier detectors are designed exactly to spot such rare, minority samples[1] – with the hope that outlierness reflects true riskiness: for example, users with unusually large engagement volumes and short inter-action durations on social networks, insurance claims for amounts much larger than expected, etc.

However, when being a minority (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.*** While for OD, the outcome is simply a matter of the statistical properties of the underlying data, from the perspective of societal values, and particularly fairness, it becomes an ethical matter.

Unfair OD leads to disparate impact: What would happen if we did not strive for *fairness-aware* OD in the existence of societal minorities in the data? What effect would the aforementioned group disparity in OD outcomes (i.e. disparate group flag rates) have downstream?

OD models treating the minority values in what-is-called *protected* variables (PV) (such as race/sex/etc.) as mere statistically rare observations contributes to the likelihood of minority group members being flagged as outliers (see Fig. 1). This issue is further exacerbated by potentially many other variables, called proxies, that partially-redundantly encode (i.e. correlate with) the PV(s), by increasing the number of subspaces in which the minority samples stand out. This results in over-representation of minority groups in OD outcomes, leading to the *over-policing* of minority group members. While this may enable obtaining a more complete picture of the minority group (e.g. by auditors, law-enforcement agents), it also implies under-policing the majority group given a limited policing capacity, wherein only a small fraction of all the data samples with the largest outlierness (based on the model) can be investigated given human, time and cost constraints.

Over-policing via OD can also feed *back* into a system when the policed outliers are used as labeled data in downstream supervised learning tasks. Alarmingly, this initially skewed sample (due to unfair OD), may be compounded through a feedback loop that amplifies the sample skewness over

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

time. For example, in predictive policing, once the initial police allocations to city neighborhoods are made using the predictions from a classifier, the subsequent classifier is updated from the crime observations from those allocated neighborhoods. Thereby, it creates a feedback loop wherein more crime is likely identified in more heavily policed neighborhoods, leading to police being allocated more often to the same neighborhoods irrespective of potentially increasing (but unobserved) true crime rates in other, under-policed neighborhoods.

As such, use of OD in societal applications has a direct bearing on individuals' social well-being; therefore, it is pivotal to ensure that the decisions based on OD outcomes are not discriminatory against any societal groups. This demands design of fairness-aware outlier detection models, which our work aims to address.

Prior research and challenges: The vast majority of work on algorithmic fairness has focused on supervised ML tasks [8, 16, 27, 69]. Numerous notions (definitions and corresponding mathematical formalisms) of fairness [5, 16, 67] have been explored in the context of supervised classification and regression problems. Each fairness notion has its own challenges in achieving equitable decisions in those supervised settings [16]. 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 for learning. The only work tackling[2] unfairness in the OD literature is an LOF [12] based detector [51] by P and Sam Abraham that proposes an ad-hoc procedure to introduce fairness specifically to the LOF algorithm. A key issue with this approach, however, is that it invites disparate treatment at decision time (among a few other shortcomings, see Sec. 5). Disparate treatment necessitates the use of protected variables *at decision time*, leading to taste-based discrimination [17]. Moreover, in many critical application domains where OD is employed, as discussed earlier, use of protected variables for decision-making explicitly is unlawful.

On the other hand, one could potentially re-purpose existing fair representation learning techniques [9, 21, 72] as well as data pre-processing strategies [23, 33] for fairness-aware OD. These approaches respectively learn new embeddings of the input samples such that the membership of samples to protected groups is obfuscated/masked, or readjust the data distributions in effort to equalize representation for the protected groups. Subsequently, the transformed data can be fed into any off-the-shelf outlier detector to achieve fair outcomes. A key issue with these types of approaches (among other shortcomings, see Sec. 5) is that data pre-processing as an isolated step prior to detection is oblivious to the detection task itself, which in turn, as we show in our experiments (Sec. 4) yields suboptimal detection performance, characteristic of largely (needlessly) sacrificing detection performance for fairness.

Our contributions: In this work, our goal is to design a fairness-aware OD model that aims to achieve equitable policing across groups (as induced by protected variables). Through fair OD, under-represented minority groups in the data would receive justifiable treatments in outlier determination, and avoid unjust policing simply because they constitute statistically rare/minority samples. To that end, we first motivate characterizing properties of fair OD based on which we define the fairness-aware OD problem, propose well-motivated fairness criteria for unsupervised OD, and introduce an end-to-end fairness-aware OD model, called FairOD, which incorporates the proposed fairness criteria directly into OD during model training. We summarize our main contributions as follows:

[2]There exists an earlier piece of work by Davidson and Ravi [19] with the aim to *quantify* or measure the fairness of OD model outcomes *post hoc* (i.e. proceeding detection), which thus has a different scope.

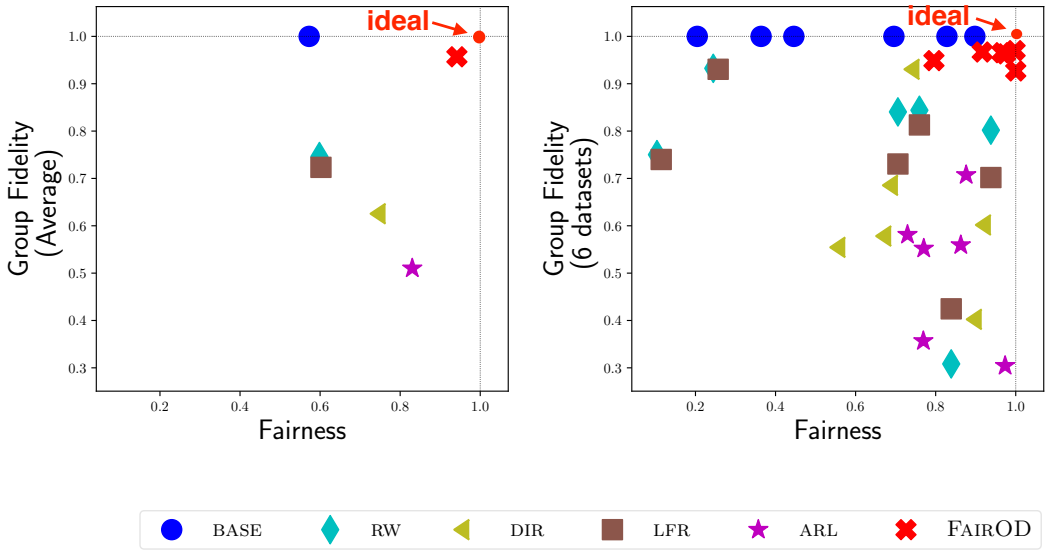


Fig. 2. `Fairness` (as quantified by statistical parity) vs. `GroupFidelity` (as quantified by group-level rank preservation) of baseline methods and our proposed FAIROD, (left) averaged across datasets, and (right) on individual datasets (depicted by separate points per method). Note that FAIROD outperforms existing solutions and achieves Fairness while preserving group-level ranking (Group fidelity) from the BASE detector. See Sec. 4 for more details.

(1) **Desiderata and Problem Definition for Fair Outlier Detection:** We identify five properties that characterize detection quality and fairness in OD. These properties dictate the design of detectors that are fairness-aware. We present justification for each of the identified properties and outline what properties can be realized in an unsupervised detector, based on which we formally define the (unsupervised) fairness-aware OD problem (Sec. 2).

(2) **Fairness Criteria and New, Fairness-Aware OD Model:** We introduce well-motivated fairness criteria and give mathematical objectives which can be optimized to obey desiderata for the defined fairness-aware OD problem. The criteria are universal, in that they can be embedded into the objective function of any end-to-end outlier detector. We propose FAIROD, a fairness-aware detector, which incorporates the prescribed criteria directly into its training. Notably, FAIROD (1) does not employ disparate treatment at test time, (2) aims to flag equal proportion of samples from all groups (i.e. obtain group fairness, via statistical parity), while (3) striving to flag truly high-risk fraction of samples within each group. (Sec. 3.1)

(3) **Effectiveness on Real-world Data:** We apply FAIROD on a number of both real-world and synthetic datasets, including use cases such as credit risk assessment and hate speech detection. Experiments demonstrate the effectiveness of FAIROD in achieving the fairness goals (Fig. 2) as well as providing accurate detection (Fig. 6, Sec. 4), significantly outperforming state-of-the-art unsupervised fairness techniques utilized for representation learning and data pre-processing prior to the OD task.

Reproducibility: The source code for FAIROD and all datasets used in our evaluation are released at https://tinyurl.com/fairOD.

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 $fr_v = P(O = 1 \mid PV = v)$

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 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 $\mathcal{Y} = \{Y_i\}_{i=1}^N$, where $Y_i \in \{0, 1\}$, 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 $O : \mathcal{X} \mapsto \{0, 1\}$ to denote the predicted outcomes of an outlier detector, and $s : \mathcal{X} \mapsto \mathbb{R}$ to capture the corresponding numerical outlier scores 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 $\mathcal{S} = \{s(i)\}_{i=1}^N$ 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 $br_a = P(Y = 1 \mid PV = a)$ for the majority group. Finally, we let $fr_a = P(O = 1 \mid PV = a)$ depict the *flag rate* of the detector for the majority group. Similar definitions extend to the minority group 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" with respect to 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, followed by the formal problem definition and our proposed solution, FairOD.

2.1 Proposed Desiderata

D1. **Detection effectiveness:** We require an OD model to be accurate at detection, that is when the scores assigned to the instances by OD well-correlate with the ground-truth outlier labels.

[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.

Specifically, OD benefits the policing effort only when the detection rate (a.k.a. positive predictive value, or precision) is strictly larger than the *base rate* (a.k.a. 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. We present *fairness-related* conditions for OD next.

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

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

In words, the probability that the OD outputs an outlier label O for a given feature vector X remains unchanged even upon observing the value of the PV.
Non-disparate treatment 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.

D3. **Statistical parity (SP):** One would expect the OD outcomes to be independent of group membership, i.e. $O \perp\!\!\!\perp PV$. To this end, this notion of fairness (a.k.a. demographic parity, group fairness, or independence), in the context of OD, 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

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

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

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

The motivation for SP derives from luck egalitarianism [38] – a family of egalitarian theories of distributive justice that aim to counteract the distributive effects of "brute luck". It aims to counterbalance the manifestations of "brute luck," by redistributing equality to those who suffer through no fault of their own choosing, mediated via race, gender, etc. Importantly, SP ensures proportional flag rates across PV groups, eliminating group-level bias. Therefore, it merits incorporation in OD since the results of OD are used for policing or auditing by human experts in downstream applications.
SP, however, is not sufficient to ensure both equitable *and* accurate outcomes as it permits so-called "laziness" [5]. Being an unsupervised quantity that is agnostic to the ground-truth labels $\boldsymbol{y}$, SP could be satisfied while producing decisions that are arbitrarily inaccurate for any or all of the groups. In fact, perhaps one extreme 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 (unusefully, 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 but those from the *true-outlier* portion of each group.

D4. **Group fidelity (a.k.a. Equality of Opportunity):** It is desirable that true outliers are equally likely to be assigned higher scores, in turn flagged, regardless of their membership to any groups as induced by PV. We call 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 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 [27, 67]. 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.
A strong caveat here is that (5) is a supervised quantity that requires access to the ground-truth labels $\boldsymbol{\mathcal{Y}}$, 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 [5], implying that simultaneously satisfying both of these conditions (exactly) is not possible.

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 caveats as in D4, regarding unavailability of $\boldsymbol{\mathcal{Y}}$.

2.2 Enforceable Desiderata

The above listed desiderata outline the properties that an ideal OD model would exhibit. In practice, however, all the desired properties are not actually enforceable. Next, we reassess the properties with respect to whether they can be enforced in a fairness-aware unsupervised OD framework given problem constraints.

D1. **Detection effectiveness:** We can indirectly control for detection effectiveness via careful feature engineering. We assume that, with domain experts assisting in feature design, the features would be reflective of the outcome labels, and enable a suitable detector achieving $P(Y = 1|O = 1) > P(Y = 1)$ – that is, a detection rate better than one would achieve by

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

flagging instances as outliers through random sampling. Given carefully-designed features and expressive models, it would be reasonable to expect a better-than-random detector.

D2. **Treatment parity:** We can build an OD model using a disparate learning process [43] that uses PV only during the model training phase, but does not require access to PV for producing a decision. In this case, treatment parity is inherently satisfied since OD decisions do not rely on PV.

D3. **Statistical parity (SP):** To achieve SP, the PV should be statistically independent of the OD decision, as indicated by Eq. (3). This implies that the output label distributions across the groups are equal. We could achieve 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. In other words, we can consider SP as an enforceable constraint during model training.

D4. **Group fidelity:** 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 ranked list 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 group ranked list based on outlier scores from our proposed OD model. Group rank preservation addresses the "laziness" issue which 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 and defer a formal analysis to future work.

D5. **Base rate preservation:** As noted in the previous subsection, population base rate preservation also involves the ground-truth labels $\mathcal{Y}$, which are not available to an unsupervised OD task. Importantly, provided an OD model satisfies D1 (detection effectiveness) and D3 (SP), 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 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 can hope for the preservation of the *ratio* of the base rates (i.e. it is not impossible). As such, a relaxed notion of D5 is to preserve proportional base rates across groups in the OD results, i.e.

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

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 rates *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 that* $P(Y=1|PV=v, O=1) > P(Y=1|PV=v), \forall v$.

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

Claim 1 and Claim 2 indicate that if we have both (a) better-than-random precision (D1) and (b) 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.

2.3 Problem Definition

Based on the above 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* $\mathcal{O}$*, such that*

(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) $P(O=1|PV=a) = P(O=1|PV=b)$ *, and* *[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.

2.4 Caveats of a Simple Approach

A simple yet naïve approach to address Problem 1 and obtain a fairness-aware OD based on a BASE model can be designed as follows:

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

As such, outlier instances are selected from each group while ensuring equal flag rates across groups. Note that this approach would fully satisfy (iii) and (iv) in Problem 1 by design, as well as (i) given suitable features. However, although easy to implement, it suffers from *disparate treatment* since it would require access to the value of PV for new incoming instances in order to identify their group-specific ranked list, hence violating (ii).

Next, we present our proposed approach, called FairOD, which satisfies all of the enforceable desiderata (i)–(iv) in Problem 1, thereby providing the benefits of the direct approach without exhibiting disparate treatment, which is unacceptable in a variety of application domains.

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 [61], deep anomaly detector [13], variational AE for OD [3], and deep one-class classifiers [60]. 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 [15, 47, 75] 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 called 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 $G(X) = G_D\left(G_E\left(X\right)\right)$. 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, 52, 62]. 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. Specifically, 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 function:

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

where $\alpha \in [0, 1]$ and $\gamma \geq 0$ are hyperparameters which govern the balance between different components 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,

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

We adapt this absolute correlation loss from [8], which proposed its use in a supervised setting with the goal of enforcing statistical parity. As [8] 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 quite low correlation 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. 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) [31] 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 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 [57]. Then, the group fidelity loss component $\mathcal{L}_{GF}$ is given as

$$\mathcal{L}_{GF} = \sum_{v \in \{a,b\}} \left(1 - \sum_{X_i \in \mathcal{X}_{PV=v}} \frac{2^{s^{\text{BASE}}(X_i)} - 1}{\log_2\left(1 + \sum_{X_k \in \mathcal{X}_{PV=v}} \text{sigm}(s(X_k) - s(X_i))\right) \cdot IDCG_{PV=v}}\right) \tag{13}$$

where $\text{sigm}(x) = \frac{\exp(-cx)}{1+\exp(-cx)}$ is the sigmoid function where $c > 0$ is the scaling constant, and,

$$IDCG_{PV=v} = \sum_{j=1}^{|\mathcal{X}_{PV=v}|} \frac{2^{s^{\text{BASE}}(X_j)} - 1}{\log_2(1+j)}$$

is the ideal (hence I), i.e. largest possible DCG value attainable for the respective group. Note that IDCG can be computed per group apriori to model training based on the 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 [37] implemented in PyTorch, thus we do not elaborate further on model optimization.

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 statistical parity 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.

3.3 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 [35]. 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.

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 details and the experimental setup, followed by key evaluation questions and results.

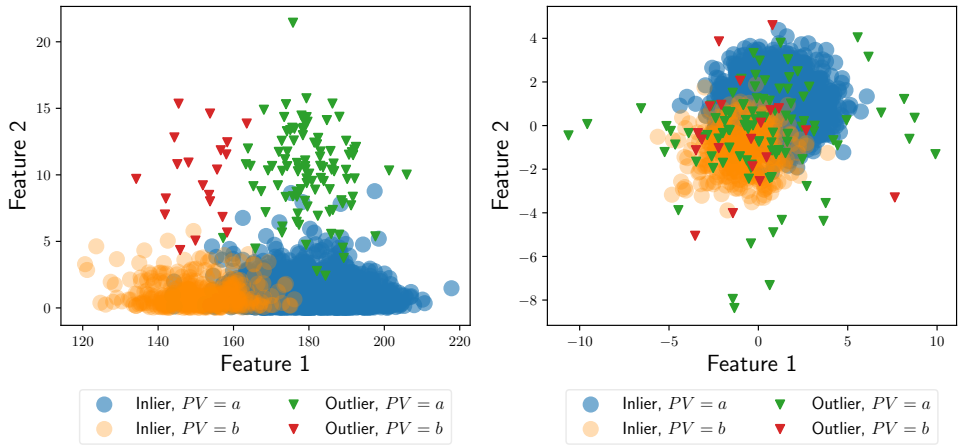
4.1 Dataset Description

Table 2 gives an overview of the datasets used in evaluation. We elaborate on details as follows.

4.1.1 Synthetic data. We illustrate the effectiveness of FAIROD on two synthetic datasets, namely `Synth1` and `Synth2` (as illustrated in Fig. 3). These datasets 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.

Table 2. Summary of datasets.

Dataset	N	d	PV	$PV = b$	$\lvert X_{PV=a}\rvert / \lvert X_{PV=b}\rvert$	% 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

Fig. 3. Synthetic datasets. See Sec. 4.1 for the details of the data generating process.

- 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 $\mathcal{Y}$, while x_2 is correlated with these labels $\mathcal{Y}$ (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. x_2 is distributed similarly 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 (left), and Fig. 3a shows a visual.

- Synth2: In Synth2, we again simulate a 2-dimensional dataset comprised of samples $X = [x_1, x_2]$ where x_1, x_2 are partially correlated with both the protected variable PV as well as ground-truth outlier labels $\mathcal{Y}$ (see Fig. 3b). 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. For inliers, both x_1, x_2 are normally distributed, and differ across majority and minority groups only in terms of shifted means, but equal variances. Outliers are drawn from a product distribution of an exponential and linearly transformed Bernoulli distribution (product taken for symmetry). The detailed generative process for the data is below (right), and Fig. 3b shows a visual.

Synth1

Simulate samples $X = [x_1, x_2]$ by...

$$PV \sim \text{Bernoulli}(4/5)$$
$$Y \sim \text{Bernoulli}(1/20)$$
$$x_1 \sim \begin{cases} \text{Normal}(180, 10) & \text{if} \quad PV = 1 \quad \text{[a, majority]} \\ \text{Normal}(150, 10) & \text{if} \quad PV = 0 \quad \text{[b, minority]} \end{cases}$$
$$x_2 \sim \begin{cases} \text{Normal}(10, 3) & \text{if} \quad Y = 1 \quad \text{[outlier]} \\ \text{Exponential}(1) & \text{if} \quad Y = 0 \quad \text{[inlier]} \end{cases}$$

Synth2

Simulate samples $X = [x_1, x_2]$ by...

$$PV \sim \text{Bernoulli}(4/5)$$
$$Y \sim \text{Bernoulli}(1/20)$$
$$x_1 \sim \begin{cases} \text{Normal}(-1, 1.44) & \text{if} \quad Y = 0,\ PV = 1 \quad \text{[a, majority; inlier]} \\ \text{Normal}(1, 1.44) & \text{if} \quad Y = 0,\ PV = 0 \quad \text{[b, minority; inlier]} \\ 2 \times \text{Exponential}(1)(1 - 2 \times \text{Bernoulli}(1/2)) & \text{if} \quad Y = 1 \quad \text{[outlier]} \end{cases}$$
$$x_2 \sim \begin{cases} \text{Normal}(-1, 1) & \text{if} \quad Y = 0,\ PV = 1 \quad \text{[a, majority; inlier]} \\ \text{Normal}(1, 1) & \text{if} \quad Y = 0,\ PV = 0 \quad \text{[b, minority; inlier]} \\ 2 \times \text{Exponential}(1)(1 - 2 \times \text{Bernoulli}(1/2)) & \text{if} \quad Y = 1 \quad \text{[outlier]} \end{cases}$$

4.1.2 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** [42] (`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* $\in$ {*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** [42] (`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* ≤ 25 indicates the minority group. We assign individuals with delinquent *payment status* as outliers ($Y = 1$). The *age* > 25 to *age* ≤ 25 imbalance ratio is 4:1 and contains 5% outliers across groups induced by the protected variable.

• **Abusive Tweets** [10] (`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 [10]. 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** [42] (`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.

We avoid specific references to each dataset in the main results discussion for clarity of presentation, and focus on observed trends between methodological choices across datasets.

4.2 Baselines

We compare FairOD to two classes of baselines: (*i*) a fairness-agnostic base detector that employs a procedure to 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:

(1) 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:

(1) RW [33]: A preprocessing approach that assigns weights to observations in each group differently to counterbalance the under-representation of minority samples.
(2) DIR [23] 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.
(3) LFR: This baseline is based on [72] 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.
(4) ARL: This is based on [9] that finds latent representation for the underlying data 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.

    The OD task proceeds the preprocessing, where we employ the BASE detector on the data representation learned by each of the preprocessing based baselines.

We do not compare against the LOF-based fair detector introduced by [51] as it explicitly relies on disparate treatment and is hence inapplicable in settings we consider.

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* $\in \{0.0001, 0.001, 0.01, 0.1, 1.0\}$ for DIR, $A_z \in \{0.0001, 0.001, 0.01, 0.1, 0.9\}$ and $A_x = 1 - A_z$ 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` (see Eq.(14)) as they only optimize for statistical parity. The best BASE model is selected based on reconstruction error through cross validation upon multiple runs with random seeds.

4.3 Evaluation

We design our experiments to answer the following questions:

- **[Q1] Fairness:** How well does FairOD achieve fairness w.r.t. the fairness metrics as compared to the baselines? How well does FairOD retain the within-group ranking from BASE?

- **[Q2] Fairness-accuracy trade-off:** How accurately are the outliers detected by FairOD compared to fairness-agnostic base detector?
- **[Q3] Ablation study:** How do different design elements of FairOD influence group fidelity and fairness of the detector?

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

$$\texttt{Fairness} = \min(r, \frac{1}{r}) \in [0, 1] \ . \tag{14}$$

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.

$$\texttt{GroupFidelity} = \text{HM}(NDCG_{PV=a}, NDCG_{PV=b}) \tag{15}$$

where

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

where $|\mathcal{X}_{PV=a}|$ 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 such that $detector \in$ {rw, dir, lfr, arl, FairOD }. Then the measure is given as

$$\texttt{Top-}k\texttt{ Rank Agreement} = \frac{|\pi^{\text{BASE}}_{[1:k]} \cap \pi^{detector}_{[1:k]}|}{|\pi^{\text{BASE}}_{[1:k]} \cup \pi^{detector}_{[1:k]}|} \ . \tag{16}$$

Supervised parity measures. We next introduce supervised measures of parity — `AP-ratio` and `P@`k`-ratio` — when the ground-truth labels $\mathcal{Y}$ are available for evaluation. The former is the ratio of Average Precision `AP-ratio` across groups, defined as

$$\texttt{AP-ratio} = \frac{\text{AP}_{PV=a}}{\text{AP}_{PV=b}} \ . \tag{17}$$

The latter is the ratio of Precision@k across groups is given as

$$\texttt{P@}k\texttt{-ratio} = \frac{\text{Precision}_{PV=a}}{\text{Precision}_{PV=b}} \ . \tag{18}$$

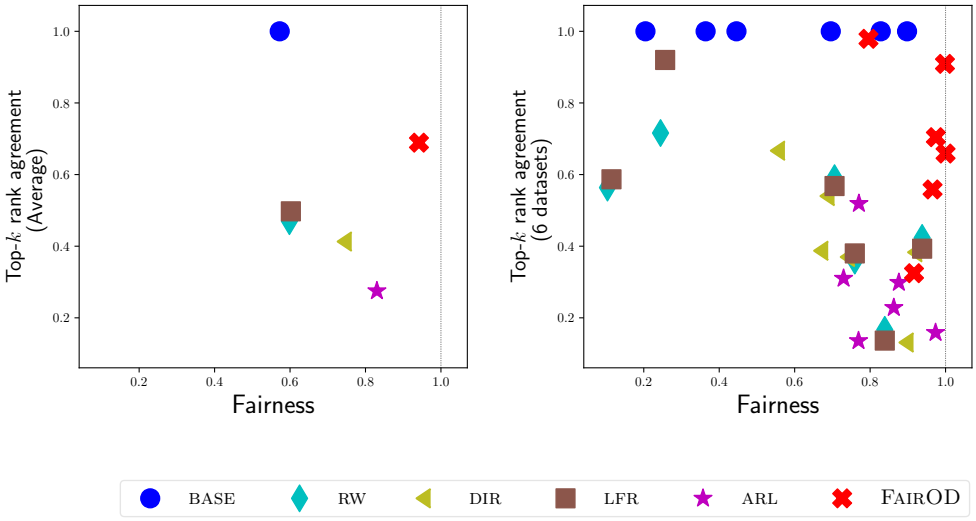


Fig. 4. FAIROD achieves better Top-k Rank Agreement compared to the competitors as averaged over datasets (left). FAIROD is on the Pareto front of Top-k Rank Agreement and Fairness across datasets (right). Each point on the right plot represents an evaluation for a dataset.

[Q1] Fairness

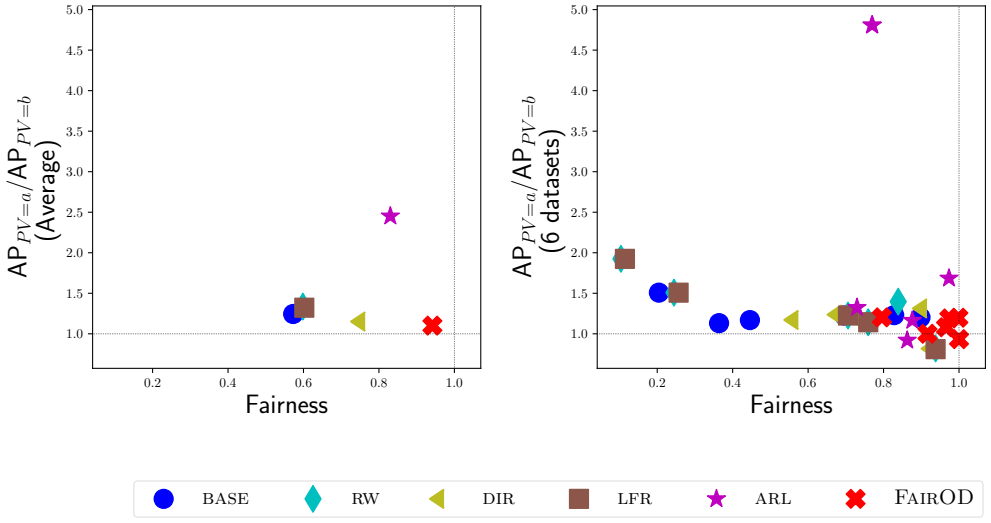
In Fig. 2 (presented in Introduction), FAIROD is compared against BASE, as well as all the preprocessing baselines across datasets. The methods are evaluated across datasets using the best configuration of each method. The best hyperparameters for FAIROD are the ones for which GroupFidelity and Fairness[5] 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 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.

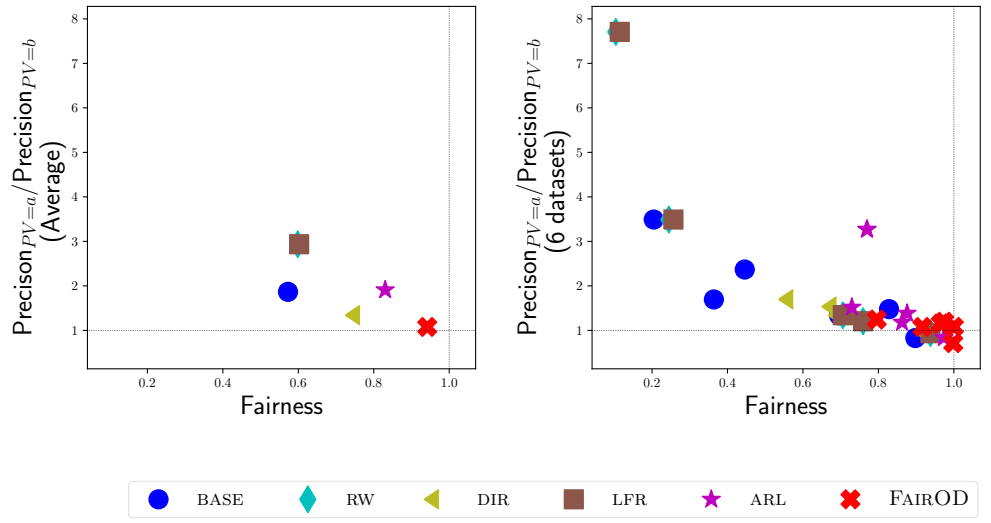
Fig. 4 reports Top-k Rank Agreement (computed at top-5% of ranked lists) of each method evaluated across datasets. The agreement measures the degree of alignment of the ranked results by a method with the fairness-agnostic BASE detector. In Fig. 4 (left), as averaged over datasets, FAIROD achieves better 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 (in fact, cannot) optimize for label-aware fairness measures. Fig. 5a evaluates

[5]Note that we can do model selection in this manner without access to any labels, since both are unsupervised measures. See Eq. (14) and (15).



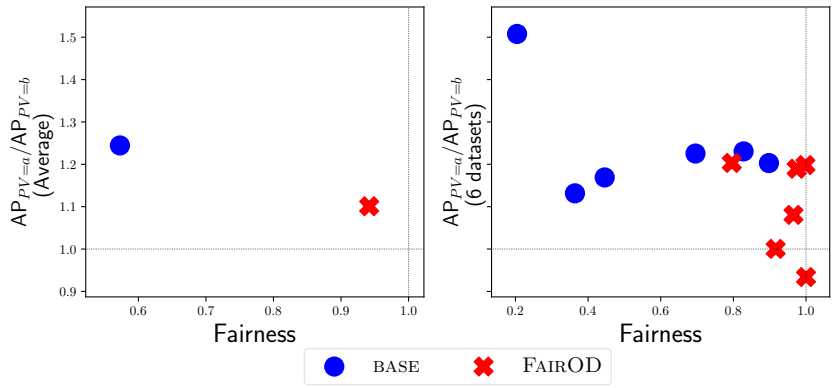
(a) Fairness vs. AP-ratio



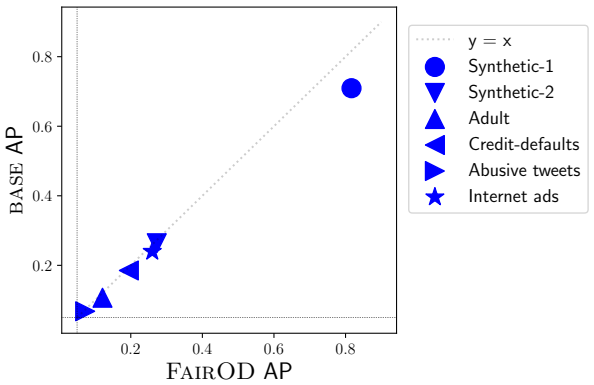
(b) Fairness vs. Precision@top-5% ratio

Fig. 5. FairOD outperforms all the competitors on the averaged metrics over datasets (left of each sub-figure) and across individual datasets (right of each sub-figure). (a.) Group AP-ratio vs. Fairness is reported for each method for the datasets. (b.) Group P@k-ratio vs. Fairness is reported for each method for the datasets.

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)). Fig. 5b reports evaluation of methods against Fairness and another label-aware parity measure – specifically, group P@k-ratio (ideal P@k-ratio = 1). As



(a) Fairness vs. AP-ratio



(b) AP of FAIROD against BASE

Fig. 6. FAIROD approaches ideal Fairness values and matches or improves detection performance, reported in terms of Average Precision (AP), as compared to BASE detector. (a) Group AP-ratio vs. Fairness for each method on individual datasets (right) and on average (left). (b) AP of FAIROD vs. BASE for all datasets.

shown in Fig. 5b (left), FAIROD outperforms all the baselines in expectation as averaged over all datasets. On the other hand, in Fig. 5b (right), FAIROD consistently approaches ideal P@k-ratio across datasets. In contrast, the preprocessing baselines are up to 8× worse than FAIROD over P@k-ratio measure across datasets.

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 average precision (AP). First we present the results comparing only FAIROD and BASE on a label-aware fairness measure – AP-ratio, and then we report their detection performance in terms of AP. In Fig. 6a, we compare FAIROD to BASE against Fairness

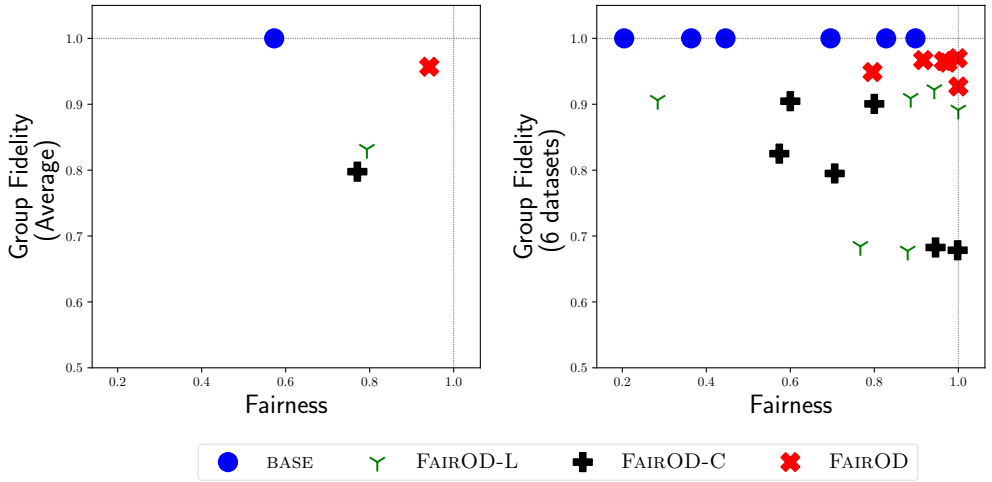


Fig. 7. FAIROD is compared to its variants FAIROD-L and FAIROD-C across datasets to evaluate the effects of different regularization components. FAIROD-L achieves `Fairness` comparable to FAIROD while suffers on `GroupFidelity` due to "laziness". FAIROD-C improves `Fairness` as compared to BASE, but under-performs FAIROD on most datasets, indicating that preserving entire group rankings may be a harder task.

and group `AP-ratio`. FAIROD outperforms BASE and attains SP as `Fairness` nears 1 across datasets and approaches a group `AP-ratio` close to 1 on majority of datasets.

Next in Fig. 6b, we compare the AP of FAIROD to that of BASE detector obtained for all the datasets. Notice that each of the datasets is slightly below the diagonal line indicating that FAIROD achieves equal or sometimes even better (!) detection performance as compared to BASE. Since FAIROD enforces SP and does not allow "laziness", it addresses the issue of falsely flagged minority samples (i.e. false positives) from BASE ranked list, thereby, improving detection performance.

From Fig. 6a- 6b, 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 detector.

[Q3] Ablation study

Finally, we evaluate the effect of various components in the design of FAIROD's fairness-aware 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 [8], which minimizes the correlation between model prediction and group membership to the protected variable. In FAIROD-L, the reconstruction error term substitutes the classification loss used in the optimization criteria in [8]. 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.

$$\mathcal{L}_{GF} = -\sum_{v\in\{a,b\}} \left| \frac{\left(\sum_{X_i\in\mathcal{X}_{PV=v}} s(X_i) - \mu_s\right)\left(\sum_{X_i\in\mathcal{X}_{PV=v}} s^{\text{BASE}}(X_i) - \mu_{s^{\text{BASE}}}\right)}{\sigma_s\ \sigma_{s^{\text{BASE}}}} \right| \tag{19}$$

where,

$$\mu_{s^{\text{BASE}}} = \frac{1}{|\mathcal{X}_{PV=v}|}\sum_{X_i\in\mathcal{X}_{PV=v}} s^{\text{BASE}}(X_i), \qquad \sigma_{s^{\text{BASE}}} = \frac{1}{|\mathcal{X}_{PV=v}|}\sum_{X_i\in\mathcal{X}_{PV=v}} (s^{\text{BASE}}(X_i) - \mu_{s^{\text{BASE}}})^2,$$

$v \in \{a, b\}$, and μ_s, σ_s are defined similarly for FairOD-C.

The comparison of FairOD and its variants are presented in Fig. 7. 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 all datasets. FairOD-L approaches SP and achieves comparable `Fairness` to FairOD except on one dataset as shown in Fig. 7 (right), which results in lower `Fairness` as 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 FairOD-L 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 compared to FairOD-L. The results show that preserving entire group ranking may be a harder task than to preserving top of the ranking. As a result, we also observe that FairOD outperforms FairOD-C across datasets.

5 RELATED WORK

Fairness in machine learning has received a considerable attention in the literature in recent years. A number of different notions of fairness have been considered [5, 16, 67], general impossibility results have been shown [6, Chapter 2], and numerous algorithmic techniques have been developed [7, 24, 27, 30, 34, 69, 70]. Most of these works focus on supervised learning problems, and do not readily apply to our setting, which centers on fairness for unsupervised outlier detection. As such, we do not elaborate further on these topics. We refer to [6, 49] for an excellent overview of these.

We organize related work in three subareas as they relate to our Fair Outlier Detection problem: 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, 14, 26], and finds numerous applications in high-stakes domains such as health-care [46], security [25], finance [55], among others. Therefore, a wide variety of methods addressing various challenges have been proposed that can be organized into broad categories of statistical based methods [59], density based [12, 36, 53], distance based [4], angle based [41], reconstruction based [13, 63], model based [61, 65], and ensemble based detection methods [15, 44, 54, 58].

Despite the vast body of work on designing new detection algorithms, there exists a minimal amount of work on the fairness aspects of OD. Specifically, we are aware of only two existing pieces of work [19, 51] on the subject. P and Sam Abraham [51] propose a detector called FairLOF, that applies an ad-hoc procedure to introduce fairness specifically to the LOF algorithm [12]. However, this approach suffers from several drawbacks. First, it invites disparate treatment at decision time that necessitates access to protected variable values, which may not be available in certain application domains or is otherwise unlawful to use, e.g. in domains like housing, employment, etc.

Second, it only prioritizes statistical parity, which as we discussed in Sec. 2, may permit "laziness." Lastly, since the approach is based on LOF, it is not end-to-end, and therefore cannot optimize a concrete objective function but is rather a heuristic procedure. The other fairness related work on OD is by Davidson and Ravi [19], which focuses on quantifying the fairness of an OD model's outcomes *post hoc* (i.e., proceeding detection) rather than tackling the fair outlier detection problem, which thus has a different scope.

To our knowledge, we are the first to systematically formalize and address the fair OD problem via an end-to-end, optimization-based solution. We expect the desiderata that we established for the fair OD problem in this work to yield further studies by the community on the subject. To facilitate this process, we share the source code of our proposed method as well as the datasets used in this paper at https://tinyurl.com/fairOD.

Fairness-aware Representation Learning

There is an increasing focus on designing methods for learning fair representations [1, 9, 21, 45, 48, 72, 73] due to their flexibility. Roughly put, those approaches learn new embeddings of the input samples so as to obfuscate/mask the membership of samples to protected groups. In other words, they map the input samples to an embedding space in which the new representations are independent of the protected variable and thus indistinguishable amongst groups.

Most recently, adversarial training processes have been applied to learn fair representations that obfuscate protected group membership while still enabling accurate classification [1, 9, 21, 48, 73]. While most of these methods fall under supervised learning as they utilize ground-truth labels, they can be plausibly extended to the OD setting by substituting classification loss with reconstruction loss. One can then strive to achieve fair detection outcomes by training an OD model on such masked data representations. However, a common shortcoming is that in all these fair methods, statistical parity (SP) has been employed as the primary criterion of fairness. In absence of ground-truth labels, other methods that utilize SP along with some label-aware parity measure, such as [1, 9, 73], fall back to SP as the default fairness criterion. On the unsupervised side, fair principal component analysis [50] and fair variational autoencoder [45] are unsupervised representation learning methods, which however also solely consider SP as their fairness criterion.

In short, fair representation learning techniques exhibit two key drawbacks for the task of unsupervised OD. First, they employ SP as the sole notion of fairness, which may be insufficient and prone to "laziness." Secondly, data embedding as an isolated step prior to detection is oblivious to the detection task itself, and therefore can yield poor detection performance (as shown in our experiments in Sec. 4).

Strategies for Data De-Biasing

There exist data manipulation strategies that are designed to modify the input data distribution in the original space, such that the outcome of a subsequent learning method would be fair with respect to the protected variables. Some of the popular de-biasing methods [33, 39] draw from topics in learning with imbalanced data [29] that employ under- or over-sampling or point-wise weighting of the instances based on the class label proportions to obtain balanced data. Similar ideas are extended to fairness-aware learning, as introduced in [33], where the sampling/weighting is instead done based on the protected variable so as to counterbalance the under-representation of minority samples. Other strategies [23] involve editing feature values such that protected variables can not be predicted based on other variables (i.e. features), thus targeting group fairness via SP. 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 that is prone to "laziness." In short, data de-biasing strategies share the same drawbacks as fair representation learning techniques.

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.

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] Fabrizio Angiulli and Clara Pizzuti. 2002. Fast outlier detection in high dimensional spaces. In *European conference on principles of data mining and knowledge discovery*. Springer, 15–27.

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

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

[7] Richard Berk, Hoda Heidari, Shahin Jabbari, Matthew Joseph, Michael Kearns, Jamie Morgenstern, Seth Neel, and Aaron Roth. 2017. A convex framework for fair regression. *arXiv preprint arXiv:1706.02409* (2017).

[8] 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.

[9] 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).

[10] Su Lin Blodgett, Lisa Green, and Brendan O'Connor. 2016. Demographic dialectal variation in social media: A case study of African-American English. *arXiv preprint arXiv:1608.08868* (2016).

[11] 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.

[12] 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.

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

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

[15] 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.

[16] 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

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

[18] Dipankar Dasgupta and Fernando Nino. 2000. A comparison of negative and positive selection algorithms in novel pattern detection. In *Smc 2000 conference proceedings. 2000 ieee international conference on systems, man and cybernetics.'cybernetics evolving to systems, humans, organizations, and their complex interactions'(cat. no. 0*, Vol. 1. IEEE, 125–130.

[19] Ian Davidson and S. S. Ravi. 2020. A Framework for Determining the Fairness of Outlier Detection.. In *ECAI*, Vol. 325. 2465–2472. http://dblp.uni-trier.de/db/conf/ecai/ecai2020.html#DavidsonR20

[20] Richard A Derrig. 2002. Insurance fraud. *Journal of Risk and Insurance* 69, 3 (2002), 271–287.

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

[22] Eleazar Eskin, Andrew Arnold, Michael Prerau, Leonid Portnoy, and Sal Stolfo. 2002. A geometric framework for unsupervised anomaly detection. In *Applications of data mining in computer security*. Springer, 77–101.

[23] 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 international conference on knowledge discovery and data mining*. 259–268.

[24] 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.

[25] 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.

[26] 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.

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

[28] Milos Hauskrecht, Iyad Batal, Michal Valko, Shyam Visweswaran, Gregory F Cooper, and Gilles Clermont. 2013. Outlier detection for patient monitoring and alerting. *Journal of biomedical informatics* 46, 1 (2013), 47–55.

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

[30] Lingxiao Huang and Nisheeth K Vishnoi. 2019. Stable and fair classification. *arXiv preprint arXiv:1902.07823* (2019).

[31] K. Järvelin and J. Kekäläinen. 2002. Cumulated gain-based evaluation of IR techniques. *ACM Transactions on Information Systems (TOIS)* 20, 4 (2002), 422–446. http://scholar.google.de/scholar.bib?q=info:6Bdw8cs-UYMJ:scholar.google.com/&output=citation&hl=de&as_sdt=0,5&ct=citation&cd=0

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

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

[34] Toshihiro Kamishima, Shotaro Akaho, Hideki Asoh, and Jun Sakuma. 2012. Fairness-aware classifier with prejudice remover regularizer. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Springer, 35–50.

[35] 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.

[36] JooSeuk Kim and Clayton D Scott. 2012. Robust kernel density estimation. *The Journal of Machine Learning Research* 13, 1 (2012), 2529–2565.

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

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

[39] 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.

[40] 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).

[41] Xiaojie Li, Jian Cheng Lv, and Dongdong Cheng. 2015. Angle-based outlier detection algorithm with more stable relationships. In *Proceedings of the 18th Asia Pacific Symposium on Intelligent and Evolutionary Systems, Volume 1*. Springer, 433–446.

[42] Moshe Lichman et al. 2013. UCI machine learning repository.

[43] 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.

[44] Fei Tony Liu, Kai Ming Ting, and Zhi-Hua Zhou. 2008. Isolation forest. In *ICDM*. IEEE, 413–422.

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

[46] Wei Luo and Marcus Gallagher. 2010. Unsupervised DRG upcoding detection in healthcare databases. In *2010 IEEE International Conference on Data Mining Workshops*. IEEE, 600–605.

[47] 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.

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

[49] 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).

[50] 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.

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

[52] 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).

[53] Emanuel Parzen. 1962. On estimation of a probability density function and mode. *The annals of mathematical statistics* 33, 3 (1962), 1065–1076.

[54] Tomáš Pevnỳ. 2016. Loda: Lightweight on-line detector of anomalies. *Machine Learning* 102, 2 (2016), 275–304.

[55] 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).

[56] Marcel Prastawa, Elizabeth Bullitt, Sean Ho, and Guido Gerig. 2004. A brain tumor segmentation framework based on outlier detection. *Medical image analysis* 8, 3 (2004), 275–283.

[57] 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.

[58] Shebuti Rayana, Wen Zhong, and Leman Akoglu. 2016. Sequential Ensemble Learning for Outlier Detection: A Bias-Variance Perspective.. In *ICDM*, Francesco Bonchi, Josep Domingo-Ferrer, Ricardo Baeza-Yates, Zhi-Hua Zhou, and Xindong Wu (Eds.). IEEE Computer Society, 1167–1172. http://dblp.uni-trier.de/db/conf/icdm/icdm2016.html#RayanaZA16

[59] Peter J Rousseeuw and Annick M Leroy. 2005. *Robust regression and outlier detection*. Vol. 589. John wiley & sons.

[60] 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.

[61] 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.

[62] 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.

[63] M-L Shyu. 2003. A novel anomaly detection scheme based on principal component classifier. In *Proc. ICDM Foundation and New Direction of Data Mining workshop, 2003*. 172–179.

[64] Clay Spence, Lucas Parra, and Paul Sajda. 2001. Detection, synthesis and compression in mammographic image analysis with a hierarchical image probability model. In *Proceedings IEEE workshop on mathematical methods in biomedical image analysis (MMBIA 2001)*. IEEE, 3–10.

[65] David MJ Tax and Robert PW Duin. 2004. Support vector data description. *Machine learning* 54, 1 (2004), 45–66.

[66] 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.

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

[68] Stijn Viaene, Richard A Derrig, Bart Baesens, and Guido Dedene. 2002. A comparison of state-of-the-art classification techniques for expert automobile insurance claim fraud detection. *Journal of Risk and Insurance* 69, 3 (2002), 373–421.

[69] 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.

[70] Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rogriguez, and Krishna P Gummadi. 2017. Fairness constraints: Mechanisms for fair classification. In *Artificial Intelligence and Statistics*. PMLR, 962–970.

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

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

[73] 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.

[74] 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.

[75] Chong Zhou and Randy C Paffenroth. 2017. Anomaly detection with robust deep autoencoders. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 665–674.

A PROOFS

A.1 Proof of Claim 1

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

$$\begin{aligned}\text{Now,}\quad P(Y=1|O=1) =& P(PV=a|O=1)\cdot P(Y=1|PV=a,O=1)+\\ & P(PV=b|O=1)\cdot P(Y=1|PV=b,O=1)\end{aligned}$$

Given SP, we have

$$\begin{aligned}& P(O=1|PV=a) = P(O=1|PV=b)\\ \implies & P(PV=a|O=1) = P(PV=a),\ \text{and } P(PV=b|O=1) = P(PV=b)\end{aligned}$$

Therefore, we have

$$\begin{aligned}\text{Now,}\quad P(Y=1|O=1) =& P(PV=a)\cdot P(Y=1|PV=a,O=1)+\\ & P(PV=b)\cdot P(Y=1|PV=b,O=1)\end{aligned} \tag{20}$$

Now,

$$P(Y=1) = P(PV=a)\cdot P(Y=1|PV=a) + P(PV=b)\cdot P(Y=1|PV=b)$$

Therefore, if want $P(Y=1|O=1) > P(Y=1)$, then

$$\begin{aligned}& P(PV=a)\cdot P(Y=1|PV=a,O=1)+\\ & P(PV=b)\cdot P(Y=1|PV=b,O=1)\end{aligned} \tag{21}$$

$$>$$

$$P(PV=a)\cdot P(Y=1|PV=a) + P(PV=b)\cdot P(Y=1|PV=b)$$

$$\implies \exists v\in\{a,b\}\quad s.t.\ P(Y=1|PV=v,O=1) > P(Y=1|PV=v)$$

□

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 $\frac{P(Y=1|PV=a)}{P(Y=1|PV=b)} = \frac{P(Y=1|PV=a,O=1)}{P(Y=1|PV=b,O=1)} = \frac{1}{r}$ then
Case 1: When $P(Y=1|PV=b,O=1) < P(Y=1|PV=b)$

$$\begin{aligned}& P(Y=1|PV=b,O=1) < P(Y=1|PV=b)\\ \implies & P(Y=1|PV=b,O=1) < r\cdot P(Y=1|PV=a)\\ \implies & P(Y=1|PV=b,O=1) < r\cdot P(Y=1|PV=a,O=1),\\ & \qquad [\because P(Y=1|PV=a,O=1) > P(Y=1|PV=a)]\end{aligned}$$

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 $P(Y = 1|PV = b, O = 1) = P(Y = 1|PV = b)$

$$
\begin{aligned}
& P(Y = 1|PV = b, O = 1) = P(Y = 1|PV = b) \\
\implies & P(Y = 1|PV = b, O = 1) = r \cdot P(Y = 1|PV = a) \\
\implies & P(Y = 1|PV = b, O = 1) < r \cdot P(Y = 1|PV = a, O = 1), \\
& \qquad [\because P(Y = 1|PV = a, O = 1) > P(Y = 1|PV = a)]
\end{aligned}
$$

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 $P(Y = 1|PV = b, O = 1) > P(Y = 1|PV = b)$.

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,

$$
\begin{aligned}
& P(Y = 1|PV = a) \cdot P(Y = 1|PV = b, O = 1) = P(Y = 1|PV = b) \cdot P(Y = 1|PV = a, O = 1) \\
\implies & P(Y = 1|PV = a) \cdot P(Y = 1|PV = b, O = 1) = P(Y = 1|PV = b) \cdot K \cdot P(Y = 1|PV = a) \\
\implies & P(Y = 1|PV = b, O = 1) = K \cdot P(Y = 1|PV = b) \\
\implies & P(Y = 1|PV = b, O = 1) > P(Y = 1|PV = b)
\end{aligned}
$$

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. □