Proceedings of the Seventh AAAI/ACM Conference on AI, Ethics, and Society (AIES 2024)

Outlier Detection Bias Busted: Understanding Sources of Algorithmic Bias through Data-centric Factors

Xueying Ding, Rui Xi, Leman Akoglu

Carnegie Mellon University
xding2@cs.cmu.edu, rxi2@andrew.cmu.edu, lakoglu@cs.cmu.edu

Abstract

The astonishing successes of ML have raised growing concern for the fairness of modern methods when deployed in real world settings. However, studies on fairness have mostly focused on supervised ML, while unsupervised outlier detection (OD), with numerous applications in finance, security, etc., have attracted little attention. While a few studies proposed fairness-enhanced OD algorithms, they remain agnostic to the underlying driving mechanisms or *sources of unfairness*. Even within the supervised ML literature,there exists debate on whether unfairness stems solely from algorithmic biases (i.e. design choices) or from the biases encoded in the data on which they are trained.

To close this gap, this work aims to shed light on the possible sources of unfairness in OD by auditing detection models under different data-centric factors.By injecting various known biases into the input data—as pertain to sample size disparity, under-representation, feature measurement noise, and group membership obfuscation—we find that the OD algorithms under the study all exhibit fairness pitfalls, although differing in which types of data bias they are more susceptible to. Most notable of our study is to demonstrate that OD algorithm bias is not merely a data bias problem. A key realization is that the data properties that emerge from bias injection could as well be organic—as pertain to natural group differences w.r.t. sparsity, base rate, variance, and multi-modality. Either natural or biased, such data properties can give rise to unfairness as they interact with certain algorithmic design choices.

Our work provides a deeper understanding of the possible sources of OD unfairness, and serves as a framework for assessing the unfairness of future OD algorithms under specific data-centric factors. It also paves the way for future work on mitigation strategies by underscoring the susceptibility of various design choices.

1 Introduction

With ML claiming its unprecedented place in the society, there exists growing concern for its responsible use and potential harm to already under-served societal groups that may exacerbate the pre-existing inequities. While much work has been dedicated to measuring and mitigating unfairness of ML algorithms, they remain mostly agnostic to the underlying sources of unfairness—treating the symptom rather than the cause. In fact, the literature is quite short on understanding the underlying *sources of algorithmic unfairness*, i.e. what drives unfairness to emerge in the first place.

Data is widely acknowledged as the key influencer of the fairness of the algorithms that are trained on it (Barocas and Selbst 2016; Mehrabi et al. 2021). However, what kind of data-centric factors give rise to certain algorithmic behavior is not well understood. Moreover, the community has mainly focused on supervised ML while the fairness of unsupervised algorithms such as for outlier detection (OD) has attracted significantly less attention, despite the numerous applications and punitive decision-making scenarios that OD algorithms are employed in.

In this paper, we aim to contribute to a deeper understanding of the possible sources of algorithmic unfairness for unsupervised OD. Specifically, we empirically investigate the role of various types of data bias as potential contributors to OD unfairness. Our findings highlight the shortcomings of several established OD algorithms in the presence of four different carefully injected data biases in a controlled simulation framework. Data biases that we study are group sample size disparity, target under-representation, feature measurement noise, and group membership obfuscation.

In addition, we identify certain data-centric properties—in particular, those related to group-wise differences in sparsity, base rate, variance, and multi-modality—that emerge as a result of each specific data bias, making it easier to understand the interaction between such data properties and the underlying working assumptions of a given OD algorithm. Perhaps more importantly, we remark that such data properties as pertain to group differences that emerge from data bias could as well be organic, i.e. natural characteristics of the input data. Then our results, when seen through the lens of data-centric factors (and not just from the perspective of data bias), provide evidence toward moving beyond the "algorithmic bias is a data problem" debate (Hooker 2021).

Our work is one of the few that presents a rigorous empirical framework toward understanding the effect of various data-centric factors on algorithmic bias, and the first one specifically focusing on unsupervised OD algorithms. We summarize our main contributions as follows and in Fig. 2.

- **Data bias as possible source of harm (§2):** We curate a list of data biases in the real world that may lead to OD unfairness; as pertain to group sample size

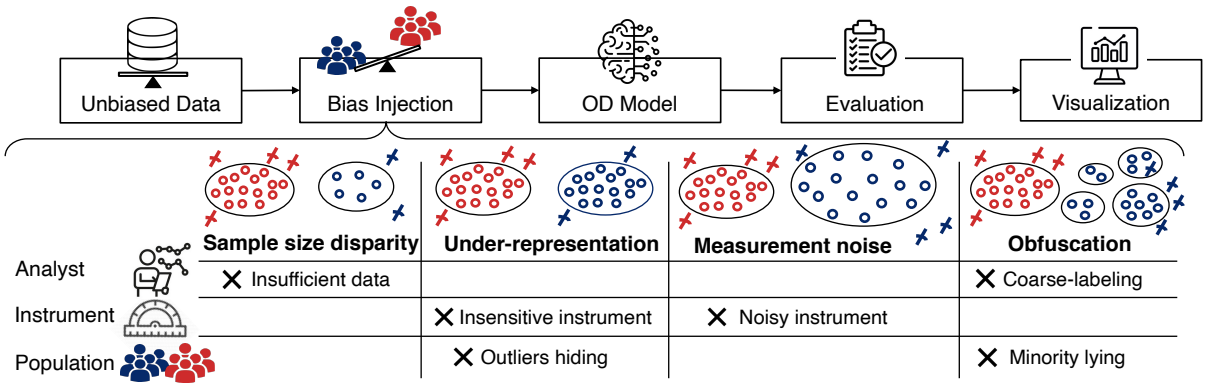

Figure 1: Overview of our study: Starting with simulated unbiased data containing outliers, we inject known types of bias into the data in a controlled setup. We then evaluate the fairness and performance of outlier detection (OD) models under various measures and report any vulnerabilities. We study four different OD models, under four different types of data bias with potential implications on OD: size disparity, target under-representation, measurement error, membership obfuscation (§2.2). These data biases are graphically illustrated when data points (shown with ○) are composed of two protected groups (in color; red & blue). Outliers are shown with cross-marks (✗).

disparity, target under-representation, feature measurement noise, and group membership obfuscation.

- **Fairness (stress-)testing popular OD models (§3):** In extensive controlled simulation settings whereby we inject known data biases, we (stress-)test several established OD techniques w.r.t. both fairness and performance metrics.
- **Extensive empirical analysis (§4):** Our analysis shows that all OD methods under study are susceptible to data bias, although their robustness vary notably depending on the type of bias. This suggests that without an understanding of the type of biases a dataset may exhibit, it would be a challenge to choose an effective OD model to employ. Further, not only data bias may be a source of OD unfairness, it may also severely harm the algorithm performance.
- **Theoretical analysis (§5):** We provide detailed analyses of how data biases interact with the working mechanism of various OD models, leading to disparate impact on different populations in the data.
- **Evidence to move beyond "(OD) algorithm bias is a data problem" (§5):** A key understanding derived from our study is that what drives OD algorithm bias is mainly the misalignment between an algorithm's working assumptions and certain input data characteristics; e.g. whether the detector respects density variability in the feature space or is susceptible to outlier masking (the notion that outliers get hidden when clustered). Crucially, while such data-centric factors may arise as a result of data bias, they could also simply be organic. This implies that OD algorithm bias can arise due to reasons beyond data bias.

Reproducibility: All code and datasets are available at: https://github.com/xyvivian/ODBias.git.

2 Data Bias: Types and Sources

2.1 Preliminaries

In general, algorithmic decisions could be punitive (e.g. imprisoning) or assistive (e.g. loan approval). Our work focuses on the former, where outlier detection (OD) is often applied on population data to flag risky individuals.

For simplicity, we represent a population as composed of two protected groups, associated with a sensitive attribute $G \in \{a, b\}$. We assume group b is the underprivileged. We denote by $\mathbf{X} \in \mathbb{R}^d$ and $Y \in \{0, 1\}$ the sets of input features and the target. In OD scenarios Y is the *true risk*, and the task is to flag the individuals with $Y = 1$. Let $S \in \mathbb{R}$ indicate the output outlierness score of an OD algorithm and $O \in \{0, 1\}$ is the decided label based on S given a threshold. The *base rate* (also, prevalence) of a group is defined as $br_G = P(Y = 1|G)$. On the other hand, the *flag rate* of a group by a detector is given as $fr_G = P(O = 1|G)$.

2.2 Four Types of Data Bias

Based on a survey of the algorithmic fairness literature that covers possible sources of algorithmic harm (Suresh and Guttag 2021; Mehrabi et al. 2021; Wang et al. 2020; Akpinar et al. 2022), we curated a list of four data biases most relevant to OD applications in the real world. We also discuss possible sources—the data collector, the measurement instrument or the population—that may drive the presence of each type of bias in the data. Fig. 2 presents a flowchart and a graphical illustration of the data biases.

Group Sample Size Bias: This type of bias reflects the real-world scenarios with data scarcity for the underprivileged sub-population (group b in our study). In effect, group b with a smaller size constitutes the statistical minority samples in the dataset. It is easy to see that OD would be susceptible to this bias as its goal is identifying outliers, i.e. statistically rare instances in the data.

The presence of sample size bias can be attributed to the data collector whom might have failed to collect enough samples for some group(s). It may also be driven by the population itself, where some groups are in minority by nature.

Target Under-Representation Bias: This type of bias is similar to the sample size bias, but only impacting the *positively-labeled* (i.e. target) individuals in the underprivileged group b. In particular, the rate of positive instances (i.e. the base rate) appear smaller in group b than that in group a. In effect, base rates are purported and the assumption of equal base rates across groups may no longer hold.

This bias could stem in the real world from the measurement instrument doing a poor job in "sensing" such individuals, as well as from the fact that such individuals may be better at "hiding" from being sampled. Alternately, the base rates may simply be different between the groups by nature.

Feature Measurement or Response Bias: This type of bias is reflective of systematically erroneous or noisy measurements in the real world associated with individuals from the underprivileged group. Such errors and noise may inflate the variance of certain features, and thereby increasing the propensity of extreme values. Extreme values appearing as outliers makes OD susceptible to this type of data bias.

The measurements for some group(s) may also display systematic under- or over-estimations. Consider the number of (re)arrests when used as a feature for risk assessment in various domains, which could be over-represented for African-Americans due to racial disparities in arrests (Schleiden et al. 2020). Another example is the systematically under-estimated SAT scores for African-American students due to implicit biases in the questions (Rattani 2016).

The source of this bias is often the measurement instrument (such as questions in the survey or test, the camera, the lab test, etc.) which is not well tuned to measuring underprivileged individuals accurately. The source could also be the population itself due to self-reporting; for example, when surveys are used to collect the measurements, some group(s) may provide more erroneous or noisy responses.

Membership Obfuscation Bias: This bias mimics the real world scenarios in which some underprivileged individuals misreport their group membership for various reasons including fear of disclosure or discrimination (Wang et al. 2020). In addition to obfuscating race/gender/etc., we expect them to also alter several of their demographic/proxy features that correlate with the sensitive attribute, to camouflage and better align with this obfuscation. For example, individuals who falsify race may also misreport their country of origin. While obfuscation may occur in all groups, we assume it to be more prevalent for the underprivileged.

This bias, stemming from the population forging various feature values, induces "fragments" in the data, i.e. subgroups within a group. In the presence of obfuscation, an underprivileged group would break into smaller subgroups, resembling other groups in some ways but not others. This can be considered as heterogeneity or multi-modal distribution within the group. Interestingly, such heterogeneity could be an organic property (e.g. in a large country like India, not all Indians are alike). In such a case, it can be seen as a labeling misstep attributed to the data collector to file multiple heterogeneous subgroups under a single attribute value.

Remarks. The four data biases we focus on in this study are not comprehensive. While there may be other OD-relevant data issues that may trigger unfair OD outcomes, there are also other known data biases that are not applicable in OD settings. For example, the problem of "tainted" (historical) labels does not apply with unsupervised OD algorithms. Moreover, train/test data distribution mismatch problem is not a concern, as we consider transductive OD wherein outliers are to be detected in a given dataset, with test data being the same as train data.

3 OD under Data Bias: Sandbox Setup

Our goal is to study the effectiveness of OD algorithms under counterfactually injected bias into the simulated unbiased data. Thus, we restrict our study to simulated data. In this regard, our work parallels the study by Akpinar et al. (2022) which performed similar fairness and fidelity analyses on simulated controlled environments for classification.

In this section, we present the details of our (unbiased) data simulation (§3.1), bias injection steps (§3.2), the OD models under study (§3.3), and the evaluation metrics (§3.4).

3.1 Data Simulation

To ensure we start with an unbiased dataset (into which we will inject specific, known biases), we simulate a population with the input features and the target described as follows.

The individuals are represented in a feature space consisting of three types of variables; $\mathcal{X} = \mathcal{X}_g \cup \mathcal{X}_c \cup \mathcal{X}_o$, where

- $X_g \in \mathcal{X}_g$ denotes the set of *proxy* variables that correlate with *G*roup membership G, but not with the target Y;
- $X_c \in \mathcal{X}_c$ depicts the "Culprit" or *incriminating* variables that are correlated with or reflective of Y; and
- $X_o \in \mathcal{X}_o$ capturing the *non-incriminating* *O*cclusion (or irrelevant) variables that neither correlate with G nor Y. The inclusion of such attributes make the OD task more realistic and non-trivial, as outliers often hide within feature *subspaces*, i.e. they stand out only w.r.t. a few relevant (but not all) features (in this case X_c).

Note that the culprit variables associate with true risk by design, while the proxy variables do not. This implies that we assume *equal base rates* across groups in the unbiased dataset. As we discuss later, this may not be the case in real world settings, where unequal base rates across groups (e.g. Internet crime propensity by ethnicity) may exist by nature.

Definition 1 (Fair Outlier Detection)**.** *In terms of X_g, X_c and X_o, the output of an OD algorithm is considered fair as*

$$P(O = 1|X_c, X_g, X_o) \;=\; P(O = 1|X_c)\;, \qquad (1)$$

that is, when the assigned outlier labels are independent of group membership/proxy variables as well as irrelevant features, given the incriminating variables.

To stress-test the fairness of OD algorithms in a controlled setting, we simulate an equal number of (1000) samples per group, and also set equal base rates, $br_a = br_b$ at 0.05 or

0.1, since outliers are rare. We use Gaussians to simulate the inliers, with group-wise means respectively at 5 and 20 for $\mathcal{X}_g$, and zero-mean for $\mathcal{X}_c$, both with unit standard deviation. $\mathcal{X}_o$ is drawn from a standard Gaussian uniformly at random.

We create two separate datasets by injecting **clustered** (repetitive or collusive) and **scattered** outliers, respectively. Clustered outliers have the same distribution for $\mathcal{X}_g$ and $\mathcal{X}_o$ as with the inliers, but in $\mathcal{X}_c$ they are drawn from a Gaussian with a higher mean of 3. Scattered outliers are created by randomly sampling a subset of the dimensions in $\mathcal{X}_c$, and inflating the variance (originally, 1) by a factor of $\{3, 6, 9, 12, 15\}$ chosen uniformly at random per outlier.

Finally, we set an equal number of dimensions (5) for $\mathcal{X}_g$, $\mathcal{X}_c$, and $\mathcal{X}_o$, since we focus on unfairness for the case when they all have equal proportions in the feature space.

3.2 Data Bias Injection

The crux of our sandbox is the study of OD algorithms under different types of data biases, as discussed in §2. We describe the steps for injecting our unbiased data as follows. Additional details are given in Apdx. §A.

1. **Group sample size bias:** With probability β_s, we independently exclude samples from the dataset where $G = b$. We inject varying degrees of sample size disparity using $\beta_s \in \{0.01, 0.05, 0.10, 0.2, 0.4, 0.6, 0.8\}$.
2. **Target under-representation bias:** With probability β_u, we independently exclude samples where $G = b$ *and* $Y = 1$. This is sample size bias impacting only the positively-labeled (i.e. target) individuals in the underprivileged group. We study bias $\beta_u \in \{0.01, 0.05, 0.10, 0.2, 0.4, 0.6, 0.8\}$.
3. **Feature measurement or response bias:** To reflect measurement noise, we inflate (i.e. multiply) the variance of the distributions of $\mathcal{X}_g$ and $\mathcal{X}_c$ for $G = b$ by a factor of β_v. To ensure that the supports of the group-wise distributions remain separate after variance inflation, we set the means of $\mathcal{X}_g$ to 5 and 20 for group a and b, respectively. Similarly, to be able to distinguish outlier-vs-inlier distributions upon inflation, we set the means of $\mathcal{X}_c$ to 0 and 10 for inliers and outliers, respectively. We vary $\beta_v \in \{0, 0.05, 0.1, 0.2, 0.5, 1, 2, 4, 6\}$.

   To mimic systematic over-estimation of task-relevant variables, we also experiment with mean-shift bias, where the mean of $\mathcal{X}_c$ is shifted additively for $G = b$ by $\beta_m \in \{0, 2, 4, 6, 8\}$.
4. **Group-membership obfuscation bias:** With probability β_g, we flip/swap the group membership of individuals with $G = b$ to $G = a$. Note that following un-awareness, OD models do not use the group membership indicator/variable G for detection. Thus, for individuals in $G = b$ whose membership has been flipped, we also swap/draw a random subset of their feature values in $\mathcal{X}_g$ from the distributions of $G = a$.

   This bias mimics the real world scenarios in which some underprivileged individuals misreport their group membership for various reasons including fear of disclosure or discrimination (Wang et al. 2020), in addition to altering several of their demographic/proxy features in $\mathcal{X}_g$ to better align with this obfuscation. While obfuscation may occur in both groups, we assume more prevalent obfuscation in the underprivileged group; setting the rate for $G = a$ to be zero for simplicity, and performing our study as β_g varies for $G = b$, in $\{0.05, 0.1, 0.15, 0.2, 0.3, 0.4\}$.

3.3 Outlier Detection Models

There is a long list of algorithms for OD (Aggarwal 2016), including the modern deep neural network models (Pang et al. 2021). Recently, a handful of fairness-enhanced OD algorithms are proposed (see §6). We consider four OD models comprising both "shallow" & deep, and both fairness-unaware & fairness- enhanced detectors: Local Outlier Factor (**LOF**) (Breunig et al. 2000), Isolation Forest (**iForest**) (Liu, Ting, and Zhou 2008), **DeepAE** (Zhou and Paffenroth 2017), and **FairOD** (Shekhar, Shah, and Akoglu 2021). LOF and iForest are shallow techniques that have been shown to be most effective on benchmark evaluations (Emmott et al. 2013). Those two are also *mechanistic*, i.e. directly model/define what an outlier is[1]. On the other hand, DeepAE and FairOD are *learning-based* and leverage deep neural networks with end-to-end learnable parameters. Moreover, FairOD is *fairness-enhanced* while others are standard detectors. We present more details as follows. (Further details on LOF and iForest can be found in Apdx. §D.2, based on which we theoretically analyze these mechanistic OD models later in §5 and Apdx. §D.3.)

Local Outlier Factor. LOF score is based on the reachability distance of a point to its k nearest neighbors (NN) relative to those distances for its NNs. As its name suggests, LOF evaluates a point w.r.t. the *local density*.

Isolation Forest. iForest makes random threshold cuts sequentially on randomly chosen features, thus building an ensemble of trees, and considers the average number of steps required to *isolate* a point from others as its outlier score.

DeepAE. Deep auto-encoder employs *compression* followed by decompression, while reconstruction error is the outlier score. The working assumption is that the inliers exhibit patterns which can be compressed well while minimizing the total reconstruction loss, while outliers that do not obey such patterns receive poor reconstruction.

FairOD. Finally, the fairness-enhanced FairOD also uses a deep autoencoder as a base model but enhances its loss objective with two additional terms for fairness regularization; one enforcing statistical parity and another toward achieving a heuristic approximation of the equality of opportunity (a.k.a. recall or TPR parity).

[1]For LOF, a point with larger reachability distance than its neighbors is an outlier. For iForest, an outlier is a point that can be isolated with few randomized axis-splits.

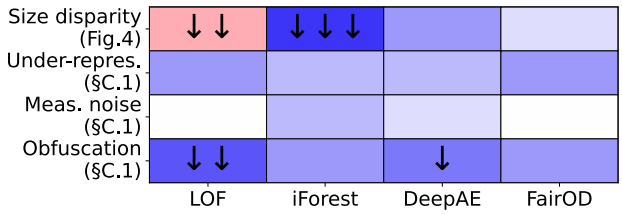

Figure 2: Qualitative summary of results for fairness stress-testing of various OD algorithms (columns) when different type data bias (rows) is applied on group b with clustered outliers. Shades of colors red and blue depict the degree of *unfairness*, when disproportionately inflicted on group a or group b, respectively; blank implying no notable difference. Arrow counts reflect relative change in overall detection *performance*; no arrows implying no notable change.

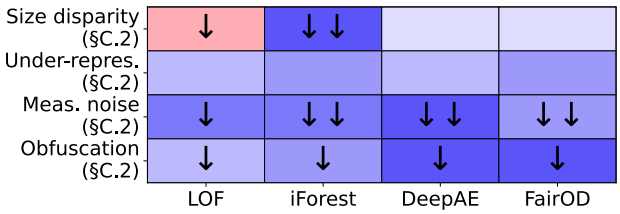

Figure 3: Qualitative summary of results for fairness stress-testing of various OD algorithms (columns) when different type data bias (rows) is applied on group b with scattered outliers. Shades of red and blue depict the degree of *unfairness*, if disproportionately inflicted on group a or group b, respectively; blank implying no notable difference. Arrows reflect the relative change in detection *performance*; no arrows implying no notable change.

Hyperparameter (HP) Tuning. OD algorithms come with HPs; e.g. number of nearest neighbors k for LOF and many others for deep models including architectural (depth and width), regularization (e.g. dropout rate) and optimization (e.g. learning rate) HPs. Critically, OD model performance is quite sensitive to HP choices (Ma et al. 2023; Ding, Zhao, and Akoglu 2022), which is nontrivial to set.

Notably, we search for the *best HP configuration* on each testbed, to prevent the situation where poor HP setting becomes a possible confounding source of unfairness (Apdx. §B). Even under optimal HPs, we observe negative implications of data bias not only on fairness but also on detection performance, as we present results shortly in §4.

3.4 Evaluation Metrics

We evaluate detection performance by AUROC; area under the ROC curve as well as F1; the harmonic mean of Precision and Recall. AUROC quantifies the overall ranking, while F1 requires a threshold on the outlier scores. We set a threshold on S to obtain as many flagged outliers with $O = 1$ as number of true outliers with $Y = 1$.

We evaluate fairness based on *group-wise ratios* w.r.t. (i) positive or flag rates (FR), (ii) true positive rates (TPR), (iii) false positive rates (FPR) and (iv) positive predictive values (PPV). When FR ratio $\frac{fr_a}{fr_b}$ is equal to 1, demographic or statistical parity is satisfied. One caveat is when group base rates differ, i.e. $br_a \neq br_b$. Then, it is suitable to measure bias amplification (Li, Goel, and Ash 2022) as the ratio between FR ratio and the ground truth base rate ratio. For all other ratios, the ideal/unbiased value is 1. For shallow and deep models, we report results averaged over 10 and 5 independent simulation runs.

4 OD on Biased Data: Empirical Findings

Starting with the unbiased datasets with an equal number of samples in each group as well as equal group base rates, we create different biased datasets by varying β_s, β_u, β_v (or β_m), and β_g and report the algorithm performances.

Fig. 2 and Fig. 3 provide a qualitative summary of our results across OD models and bias types for datasets with *clustered* and *scattered* outliers, respectively. We discuss several of our notable findings with details in Apdx. §C.Overall, as depicted by the varying shades of blue, we see that the bias-injected group b is impacted disproportionately across models and bias types, although the severity of unfairness against b varies. Moreover, data bias impacts not only fairness but also detection performance of OD models.

It is interesting to note the stark difference between the detectors in regard to susceptibility to different biases. For example, LOF is most susceptible to Obfuscation in the clustered-outliers setting and to Measurement bias for scattered outliers. iForest is susceptible to Sample size disparity, while DeepAE is most sensitive to Obfuscation in both settings. Fairness-enhanced FairOD is no exception; as it remains comparably brittle under Obfuscation. These suggest the lack of a "winner" detector.

In addition, we find that the models behave quite differently against a certain bias depending on the dataset characteristics. Notable is the Measurement bias, where models are fairly robust when outliers are clustered, which however lead to considerable unfairness as well as performance drop on datasets with scattered outliers.

4.1 Group Sample Size Bias

When samples from group b are dropped at random to inject size disparity, the inliers and outliers both sparsify relative to group a. We find that OD models react to density differences in the feature space differently.

As shown in Fig. 4, LOF's flag rate for group b *drops*, which in turn disadvantages group a with increased FPR and decreased Precision (top row). In contrast, iForest behaves in the *opposite* fashion, disadvantaging group b with significantly higher flag rate, higher FPR and lower Precision (2nd row). This contrast is due to these models treating density locally or globally. LOF evaluates outlierness *locally*, by comparing points to their neighbors. When group b inliers sparsify, the clustered b-outliers "hide" better due to masking (Jiang, Cordeiro, and Akoglu 2022). iForest, on the other hand, can more quickly isolate now-globally-sparser points in group b, overly flagging them, hence maintaining

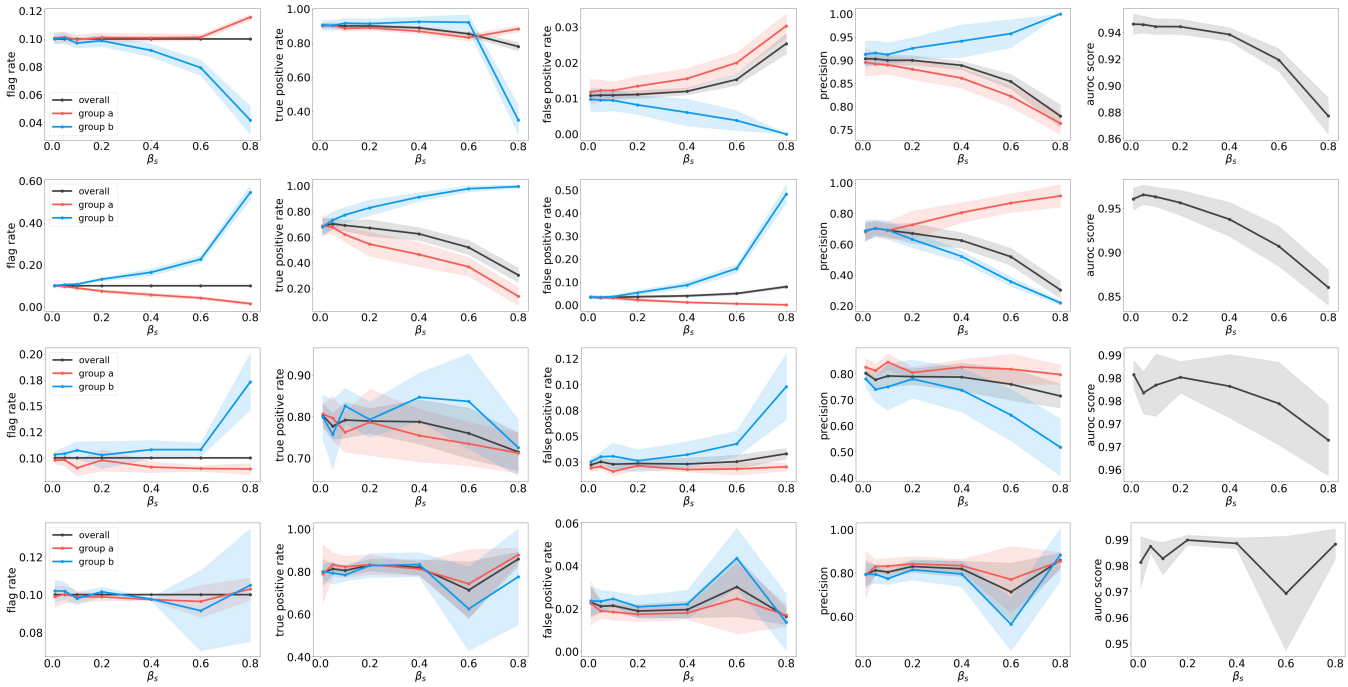

Figure 4: (best in color) Group-wise fairness metrics and AUROC for (top to bottom) LOF, iForest, DeepAE and FairOD under sample size bias on *clustered* outliers.

high TPR, but larger FPR and lower Precision. iForest's brittleness is also evident from its overall performance falling more drastically with increasing bias.

Among the deep models, DeepAE (3rd row) behaves similar to iForest in terms of group fairness, although through a different mechanism. Since subsampling group b renders them rare *minority* samples, their impact on the total loss of the compression-based DeepAE diminishes. As a result, their poor reconstruction reflects as higher flag rate, larger FPR and lower Precision. As these group-wise differences are less extreme than for iForest, DeepAE overall detection performance remains relatively stable. Finally, we find FairOD (bottom row) to be robust against size disparity, where it achieves statistical as well as TPR parity through explicit optimization. These findings continue to hold on scattered outliers as given in Apdx. §C.2 Fig. 13.

4.2 Target Under-Representation Bias

The algorithm behaviors change considerably when exposed to the under-representation bias, where we drop only the target (positive) samples from group b. Results are given for all models in Apdx. §C.1 Fig. 8 and Apdx. §C.2 Fig. 14 in the clustered and scattered outliers settings, respectively.

Note that dropping outlier samples from b renders group a's observed base rate higher. With a higher base rate, group a's clustered outliers are masked for LOF, which reduces group a's flag rate and TPR, while increasing those for b. FPR increases for both groups, proportionate to the overall increase. The masking effect goes away when outliers are scattered, with no notable TPR and FPR difference between groups. iForest is robust to masking as it finds cuts in the feature space that isolate the outlier clusters at once, with higher TPR but lower Precision for group b.

In fact, lower Precision for group b is a common trend across all models. Group b outliers continue to stand out for both models as dropping few outlier samples does not change the bulk of the data and hence the compression quality of group b's inliers. However, because group a's base rate and hence frequency is higher, its flag rate is relatively lower. This translates to a higher flag rate for b at the cost of lower Precision. These results show the vulnerability of all OD models in our study in the face of unequal base rates.

4.3 Feature Measurement Bias

Variance Shift for Measurement Noise. In the clustered setting (see Apdx. §C.1 Fig. 9), feature variance loosens the clusters among both inliers and outliers in group b. As a result, LOF ranks group a outliers strictly above those from group b as they showcase a starker contrast to their relative inliers, while it is still able to flag almost all true outliers. iForest is similarly robust with only slightly higher FPR and slightly lower TPR and Precision for group b than for a, while retaining high overall performance. Deep models maintain high performance as the smaller outlier clusters do not compress as well as the inliers despite higher variance.

While we find all detectors to be quite robust against feature variance disparity on clustered outliers, the results differ considerably when outliers are scattered (see Apdx. §C.2 Fig. 15). We show DeepAE as a representative case in Fig. 5 on clustered versus scattered outliers for comparison.

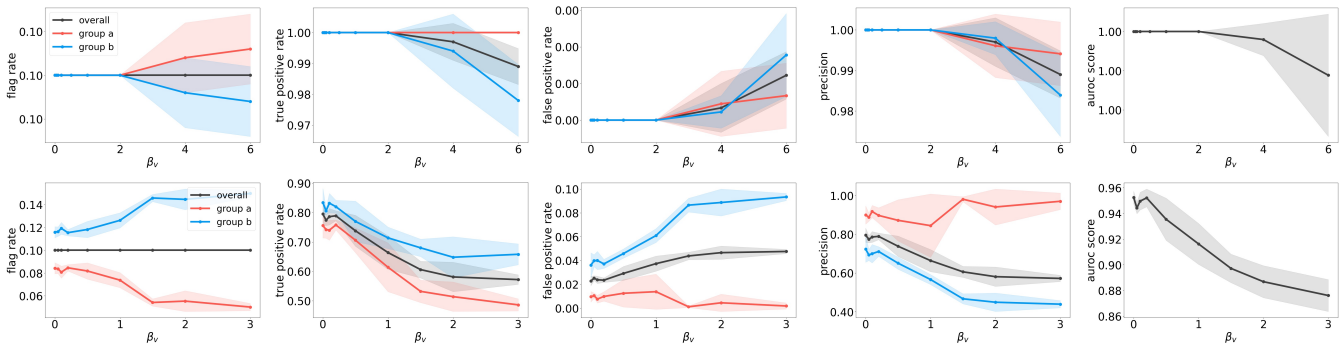

Figure 5: (best in color) Group-wise fairness metrics and AUROC for DeepAE under feature measurement bias (variance shift) on (top) *clustered* and (bottom) *scattered* outliers.

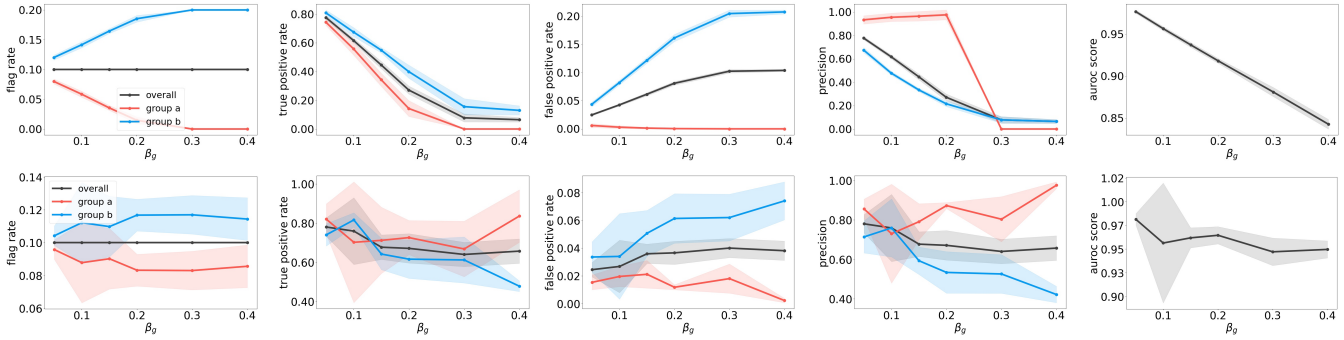

Figure 6: (best in color) Group-wise fairness metrics and AUROC for (top) LOF and (bottom) FairOD under membership obfuscation bias on *clustered* outliers.

In the scattered outliers setting, feature variance loosens the distribution of group b leading to extreme-valued inliers and outliers that are both the at the "skirts" of the distribution. As a result, as group b data sparsifies, LOF tends to blur the inliers and outliers at the periphery, with higher FPR, and lower TPR and Precision for group b. Performance also drops accordingly with increasing bias levels. Results are similar for iForest, which overly flags the now-sparser group b samples. Both deep models are also sensitive in this case, yielding unfair outcomes for group b along with reduced performance. While clustered outliers, despite larger variance, form a separate modality in the data that is hard to compress (hence easier to detect), scattered outliers are harder to detect as they are more similar to the inliers at the outskirts of the inlier distribution.

Mean Shift for Over-Estimation. We find that the OD models are *not* impacted by mean-shifting group b, provided that the groups are sufficiently apart along the proxy variables. Specifically, the shift does not impact the locality of points for LOF, the axis-cuts of iForest that isolate the outliers simply shift accordingly once the groups are cut separate along the proxy variable axes, and the compression-based deep models remain intact when one of the modalities shifts in the feature space. Results are shown for LOF and iForest for brevity in Apdx. §C.1 Fig. 10 and Apdx. §C.2 Fig. 16 in the clustered and scattered outliers settings.

4.4 Membership Obfuscation Bias

Different from earlier bias scenarios in which the data collector or their instrument may be inflicting the bias, here the individuals in the population themselves bias the data distribution. They do so by inducing mixed-up subpopulations, which resemble in some parts to group b and in other parts disguising as group a. As the obfuscation variables are randomized, this leads to *multi-modality* within group b, comprising many small subpopulations. Results can be found in Apdx. §C.1 Fig. 11 and Apdx. §C.2 Fig. 17 in the clustered and scattered outliers settings, respectively.

As shown in Fig. 6 (top), LOF is extremely sensitive to this type of bias as it mistakenly flags more and more of those small micro-clusters in group b. In fact, it seizes to flag any group a instances beyond a certain bias level as group b breaks down into more subpopulations. This leads to reduced TPR for both groups, and high FPR and low Precision for group b. LOF's overall performance also drops with increasing levels of bias. iForest, on the other hand, is relatively more robust because it is *subspace*-based; that is, it is able to isolate outliers as long as it makes axis-cuts on the incriminating variables. As long as it does not solely pick proxy variables, it does not wrongly flag group b micro-clusters. Hence it acquires slight unfairness against group b, with high overall performance retained.

Similar to LOF, DeepAE is sensitive to the presence of

small micro-clusters as more variability in data patterns imply harder compression. Thus its overall performance drops, with higher flag rate, higher FPR and lower Precision for group b. Higher FPR and lower Precision for group b are common across models including FairOD, as shown in Fig. 6 (bottom), which better balances flag rate and TPR rates between groups thanks to explicit regularization.

5 Algorithmic Bias is NOT Merely a Data (Bias) Problem

It is important to note that differences in data distributions between groups may as well be *natural*. In fact, it is perhaps easier to imagine that different groups would follow different data distributions organically, w.r.t the proxy variables.

In the following subsections, we aim to make connections between specific data properties that emerge as a result of bias and how such properties may also appear in the real world organically. These connections support the argument toward moving beyond the "algorithmic bias is a data (bias) problem" debate (Hooker 2021). That is, algorithmic bias can arise solely from the interaction between modeling assumptions and certain properties the input data exhibits, without the data being necessarily inflicted with any bias.

In a nutshell, we argue that group-wise differences in certain data properties (as pertain to sparsity, prevalence or base rate, variance, and multi-modality) — *either induced by data bias or exhibited naturally* — could result in unfair outcomes. Whichever the source may be, it is useful to understand which OD algorithms are more prone to unfairness in the face of such group-wise differences.

Theoretical Analysis. We mathematically show the (in)sensitivity of the mechanistic (i.e. non-learning based) models, LOF and iForest, in the presence of sparsity difference and multi-modality in the clustered outliers setting (Propositions 1–4). Analysis can be repeated for scattered outliers, e.g. Proposition 5. We refer to Apdx. D for preliminaries, notation, and the detailed proofs.

5.1 Group Sample Size Bias Mimics Sparsity Difference Between Groups

In injecting group sample size bias, the process of down-sampling data from one group creates a population that is sparsely sampled from its underlying manifold. As presented in §4.1, we observe that this property makes isolation-based iForest quite brittle, as it tends to more easily yet mistakenly isolate the samples in sparser regions.

However, it may as well be natural that different groups exhibit different distributions with unequal sparsity. This would be especially realistic for proxy variables. For example hair length, a proxy for gender, could follow distributions with varying density between groups. A similar argument can be made for income, a potential proxy for race, with different group-wise density distributions.

Note that in group size bias injection, down- or sparsely-sampling group b's distribution, induced not only size disparity but also sparsity disparity. Here we argue that OD models could remain prone to producing unfair outcomes when group a and b exhibit *different* density distributions even in the *absence* of any group size disparity.

Proposition 1*: In the clustered outliers setting, let groups a and b have equal size $n_a = n_b$ and base rate $P(Y = 1|a) = P(Y = 1|b)$. Let d and D denote intra-group distance between inlier pairs (same for outlier pairs) for group a and b, respectively, where $D > d$ as group b is sparser. Denote by Δ_a and Δ_b the average distance between outliers and inliers in each group (See Fig. 19). Then, assuming LOF hyperparameter k is set s.t. $k > n_a \cdot P(Y = 1|a)$ and $\Delta_a = \Delta_b$, LOF tends to assign higher scores to group a-outliers, increasing flag rate $P(O = 1|a)$. Further, when $\Delta_b \approx D$, LOF score of b-outliers is ≈ 1, i.e. close to inlier scores, leading to low TPR for group b due to masking.*

In plain words, in the presence of group-wise sparsity variation, the proposition states that the local reachability distance (i.e. LOF score) of a-outliers tends to be larger since their inlier neighbors' reachability distance is smaller because group a is denser. Further, as group b continues to sparsify such that the gap between inlier and outlier clusters of group b shrinks, b-outliers get harder for LOF to distinguish from inliers, decreasing TPR. Proposition 5 in Apdx. D.3 shows a similar result for scatter outliers.

Proposition 2: *In the clustered outliers setting, let sp denote an iTree's split, and $P(x_a < sp < x_b)$ denote the probability for a split to occur between points x_a and x_b. Let groups a and b have equal base rate $P(Y = 1|a) = P(Y = 1|b)$ and equal average distance between outliers and inliers $\Delta_a = \Delta_b$. As group b is sparser, $d < D$ denotes the intra-group distances between inlier and outlier pairs for groups a and b, respectively (See Fig. 22). Let $\{\mathcal{O}_a, \mathcal{I}_a\}$, $\{\mathcal{O}_b, \mathcal{I}_b\}$ represent the outliers, inliers sets for groups a and b. Then, for $o_b, p_b \in \mathcal{O}_b$, and $o_a, p_a \in \mathcal{O}_a$, we have $P(p_b < sp < o_b) > P(p_a < sp < o_a)$, indicating an iTree is more likely to split among group-b outliers. The difference between $P(p_b < sp < o_b)$ and $P(p_a < sp < o_a)$ becomes larger when iTree is built in higher dimensions. In addition, for $q_b \in \mathcal{I}_b$, $q_a \in \mathcal{I}_a$, $P(o_b < sp < q_b) = P(o_a < sp < q_a)$, i.e. an iTree is equally likely to split between the inlier and outlier clusters.*

iForest assigns higher outlier scores to points that can be easily isolated with few axis splits. The proposition shows that an iTree is more likely to split among the sparser b-outliers than a-outliers, yielding higher flag rate and FPR for group b. The discrepancy is more notable in higher dimensional splits. iTree is equally likely to split between the inliers and outliers for both groups. Due to this special split that isolates the clustered outliers at once, iForest can flag some fraction of a-outliers, retaining TPR for group a.

Empirically, variance shift experiments could serve as a similar scenario where samples of group b are more sparsely scattered than those of a due to inflated variance, which we observed is subject to higher FPR and lower TPR and Precision, consistent across all models (see Apdx. §C.2 Fig. 15). This shows one possible scenario, wherein group b organically exhibits greater sparsity, and group-wise sparsity difference leads to disparate OD outcomes.

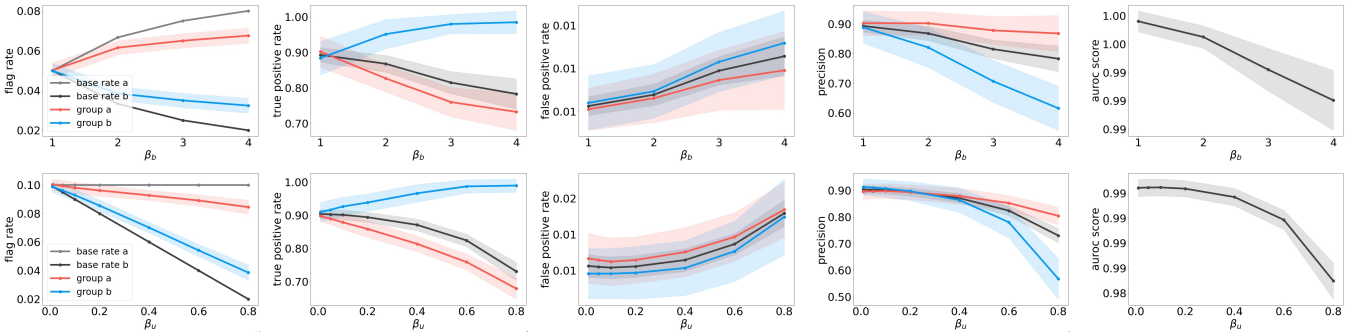

Figure 7: (best in color) Group-wise fairness metrics and AUROC for LOF under (top) unequal base rates and (bottom) target under-representation on *clustered* outliers. The results are qualitatively similar.

5.2 Target Under-Representation Bias Mimics Base-Rate Difference Between Groups

In injecting target under-representation bias, downsampling group b outliers induces its population to have a lower observed base rate than group a. Here we argue that the protected groups may as well exhibit unequal base rates naturally, where $P(Y = 1|G = a) \neq P(Y = 1|G = b)$.

For example, some groups may naturally be more inclined to criminal activities (like Internet crime, human trafficking, etc.). This highlights a key difference between punitive OD and assistive ML settings. While assistive ML algorithms should provide equal assistance across protected groups without bias, it may be appropriate to penalize different groups at varying rates when there is evidence suggesting differing base rates of criminal activity.

Then, in the presence of (natural) *unequal* base rates, unfair OD outcomes observed in §4.2 continue to hold. To demonstrate this phenomenon, we repeat our experiments by varying the base rates between groups while keeping the total number of outliers (and group-wise inliers) the same. Results are shown in Fig. 7 (top) for LOF, which remain qualitatively similar to those under target under-representation bias as shown in Fig. 7 (bottom). Results for all models are in Apdx. §C.1 Fig. 12 and Apdx. §C.2 Fig. 18.

5.3 Feature Measurement Bias Mimics Variance Difference Between Groups

Variance-shift due to measurement noise can alter the likelihood of extreme values and rare observations.This can lead to unfair OD outcomes for the high-variance group, especially with scattered outliers.

In the real world, certain features may naturally vary between groups. For instance, biomarker features used for screening can show genetic variation by age, gender, or race. A notable example is the 2011 Nymwars controversy, where tech companies like Google and Facebook aimed to block users with uncommon names, potentially discriminating against those with African first names.

We argue that variance difference between groups mimics the scenario when such difference is induced by additive measurement noise inflicted on some group(s) but not others. This is true for Gaussians; Gaussian noise added to a Gaussian distribution yields another Gaussian with larger variance. In fact, for any distribution from the stable distribution family, a linear combination of two distributions remains in the same distribution, only with different location and scale parameters (Mainardi 2007). This implies that the unfair OD outcomes observed in §4.3 would continue to hold when one group exhibits naturally higher-variance distribution for certain features.

5.4 Membership Obfuscation Bias Mimics Multi-modality Within Groups

Membership obfuscation, where group members disguise themselves as those from another group, creates subgroups that differ in the proxy feature space. This fragmentation induces multi-modality within the group. Members of a group may naturally exhibit within-group heterogeneity, forming many clusters. It is neglectful to assign such diverse groups to a single attribute. For instance, "Asian" or "Hispanic" are coarse classifications with distinct subpopulations.

When a protected group comprises multiple smaller subgroups, the risk increases that each may stand out as a minority micro-cluster. As we presented in §4.4, OD models are typically tuned to flagging micro-clusters as (clustered) outliers, leading to many false positives at the expense of low recall for the multi-modal group. In the following, we mathematically show that LOF is quite vulnerable in the presence of this data characteristic, supporting the empirical observations in Fig. 6 (top). (See all proofs in Apdx. D.3).

Proposition 3*: In the clustered outliers setting, let groups a and b have equal size $n_a = n_b$ and equal base rate $P(Y = 1|a) = P(Y = 1|b)$. Assume group b-inliers form f smaller populations, $\mathcal{I}_b^i$, for $i \in [\![1, f]\!]$, and group b-outliers form g smaller populations, $\mathcal{O}_b^j$ for $j \in [\![1, g]\!]$. Let d and D denote intra-group distance between inlier pairs (same for outlier pairs) for group a and each sub-population in b, respectively, where $D = d$. Denote Δ_a as the average distance between outliers and inliers in group a, and Δ_b as the average distance between each subpopulation in group b for both inliers and outliers (See Fig. 21). Then, assuming LOF hyperparameter k is set s.t. $k > n_a \cdot P(Y = 1|a)$, and $\Delta_a = \Delta_b$ while $\Delta_a > d$ and $\Delta_b > D$, LOF tends to assign higher scores to group b-outliers, increasing flag rate*

$P(O = 1|b)$. *Further, when* $|\mathcal{I}_b^j| < |\mathcal{O}_b^i| < k$*, LOF score of* b*-inliers is larger than of* b*-outliers, leading to both high FPR and low TPR for group* b*.*

We also show via Proposition 4 in Appx. D.3 that iForest is relatively more robust than LOF against within-group multi-modality since, unlike LOF that relies on nearest neighbor distances in the *full* feature space, it seeks outliers in *sub*spaces. As protected groups form clusters only w.r.t. proxy features, iForest is prone to false positives when axis-splits are limited to proxy features, yet, making axis-splits along the incriminating features allows iForest to isolate true outliers and hence maintain recall.

6 Related Work

Measuring and Mitigating Unfairness: Various work studied possible definitions and measures of ML fairness (Kusner et al. 2017; Verma and Rubin 2018; Beutel et al. 2019; Garg, Villasenor, and Foggo 2020; Tang, Zhang, and Zhang 2023), fairness auditing (Saleiro et al. 2018; Galdon Clavell et al. 2020; Le Merrer, Pons, and Trédan 2023), with most emphasis on mitigation in ML (Mehrabi et al. 2021; Reddy et al. 2021), through optimization (Xinying Chen and Hooker 2023), regularization (Kamishima, Akaho, and Sakuma 2011), constraints (Zafar et al. 2017), adversarial learning (Delobelle et al. 2021), and representation learning (Zemel et al. 2013), to name a few. In contrast, only a handful of work studied auditing (Davidson and Ravi 2020) and mitigation (Abraham et al. 2021; Song, Li, and Liu 2021; Shekhar, Shah, and Akoglu 2021; Liu et al. 2021; Zhang and Davidson 2021) for unsupervised OD.

Sources of Unfairness: On developing a deeper understanding of what drives discrimination by ML, various work studied the impact of data collection (Chen, Johansson, and Sontag 2018), algorithmic factors (Blanzeisky and Cunningham 2021), as well as model and data interactions (Pombal et al. 2022), while others aimed to identify and list possible sources of harm (Mehrabi et al. 2021; Suresh and Guttag 2021). Our work is inspired by the testbed by Akpinar *et al.* (Akpinar et al. 2022) that studied the explicit impact of counterfactually injected biases on *supervised* ML models. In similar vein, our study is the first to demonstrate that popular *unsupervised* OD models are susceptible to certain forms of data bias.

Fair Unsupervised Learning: Despite OD models, several works have discussed fair ML without labels, such as fair clustering (Backurs et al. 2019; Bera et al. 2019; Chen et al. 2020; Chierichetti et al. 2018),fair representation learning (Buet-Golfouse and Utyagulov 2022), and fair feature selection (Xing et al. 2021).

7 Conclusion and Discussions

Summary. We presented a descriptive measurement study that stress-tested the fairness and performance of various OD models when exposed to certain data biases. Our analyses have been expository, unearthing the pitfalls of various algorithmic design choices as they interact with certain data characteristics—such as group-wise differences in sparsity, prevalence, variance and multi-modality.

All models, whether shallow or deep, and regardless of fairness awareness, can disproportionately impact underprivileged groups, depending on data bias and model assumptions. Shallow detectors directly define outliers, leading to unfair results if their definitions misalign with the data. Deep models identify what is compressible and flag outliers as deviations from normal patterns. When multiple norms exist and some are more dominant, these models also risk producing unfair outcomes.

It is also worthy to note that while data bias may induce such data characteristics, they could also be natural. This implies that OD models can naturally fall into the same pitfalls we underscored, even when the analyst, measurement instrument and the population are not at fault of inflicting bias.

Limitations. Our study, based on simulated data, explores the impact of known types of data bias. Although our simulations may not closely mirror real-world datasets, they effectively highlight potential issues that could naturally arise in practice. We did not perform experiments on real-world datasets intentionally. Our aim has been to study which data bias gives rise to unfair outcomes in a controlled setup. Explaining unfairness observed in the wild by pinpointing the root-causes would be a reverse-engineering effort, which is not our intended scope.

Future Directions. Our work establishes a foundation for identifying biases in data that affect fair OD outcomes and suggests potential mitigation strategies, leaving deeper investigation for future research.

Fairness interventions are classified into pre-, post-, and in-processing. Pre-processing alters input data, post-processing adjusts output decisions, and in-processing incorporates fairness during training. Pre-processing is ineffective when group differences are natural. Post-processing can use different thresholds to achieve demographic parity (Corbett-Davies et al. 2017; Menon and Williamson 2018), but optimizing fairness metrics in unsupervised OD is challenging without ground truth. In-processing techniques like decoupling (Dwork et al. 2018; Ustun, Liu, and Parkes 2019) train separate detectors with a joint loss.

Both post-processing and decoupling may cause treatment disparity, assuming the use of sensitive attributes at decision time is ethical and legal. Addressing differences between groups may require accepting disparate treatment to avoid disparate impact (Lipton, McAuley, and Chouldechova 2018). However, these methods don't address within-group discrimination. Exploring clustering-based OD or defining more granular sensitive attributes for subpopulations could be potential directions.

Acknowledgments

This work is sponsored by the National Science Foundation IIS-2310481. Any conclusions expressed in this material are those of the authors and do not necessarily reflect the views, expressed or implied, of the funding parties.

References

Abraham, S. S.; et al. 2021. Fairlof: fairness in outlier detection. *Data Science and Engineering*, 6(4): 485–499.

Aggarwal, C. C. 2016. Outlier analysis second edition.

Akpinar, N.-J.; Nagireddy, M.; Stapleton, L.; Cheng, H.-F.; Zhu, H.; Wu, S.; and Heidari, H. 2022. A sandbox tool to bias (Stress)-test fairness algorithms. *arXiv preprint arXiv:2204.10233*.

Backurs, A.; Indyk, P.; Onak, K.; Schieber, B.; Vakilian, A.; and Wagner, T. 2019. Scalable Fair Clustering. In Chaudhuri, K.; and Salakhutdinov, R., eds., *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, 405–413. PMLR.

Barocas, S.; and Selbst, A. D. 2016. Big data's disparate impact. *California law review*, 671–732.

Bera, S. K.; Chakrabarty, D.; Flores, N. J.; and Negahbani, M. 2019. Fair Algorithms for Clustering. arXiv:1901.02393.

Beutel, A.; Chen, J.; Doshi, T.; Qian, H.; Woodruff, A.; Luu, C.; Kreitmann, P.; Bischof, J.; and Chi, E. H. 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.

Blanzeisky, W.; and Cunningham, P. 2021. Algorithmic factors influencing bias in machine learning. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, 559–574. Springer.

Breunig, M. M.; Kriegel, H.-P.; Ng, R. T.; and Sander, J. 2000. LOF: identifying density-based local outliers. In *Proceedings of the 2000 ACM SIGMOD international conference on Management of data*, 93–104.

Buet-Golfouse, F.; and Utyagulov, I. 2022. Towards Fair Unsupervised Learning. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '22, 1399–1409. New York, NY, USA: Association for Computing Machinery. ISBN 9781450393522.

Chen, I.; Johansson, F. D.; and Sontag, D. 2018. Why is my classifier discriminatory? *Advances in neural information processing systems*, 31.

Chen, X.; Fain, B.; Lyu, L.; and Munagala, K. 2020. Proportionally Fair Clustering. arXiv:1905.03674.

Chierichetti, F.; Kumar, R.; Lattanzi, S.; and Vassilvitskii, S. 2018. Fair Clustering Through Fairlets. arXiv:1802.05733.

Corbett-Davies, S.; Pierson, E.; Feller, A.; Goel, S.; and Huq, A. 2017. Algorithmic decision making and the cost of fairness. In *Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining*, 797–806.

Davidson, I.; and Ravi, S. S. 2020. A framework for determining the fairness of outlier detection. In *ECAI 2020*, 2465–2472. IOS Press.

Delobelle, P.; Temple, P.; Perrouin, G.; Frénay, B.; Heymans, P.; and Berendt, B. 2021. Ethical adversaries: Towards mitigating unfairness with adversarial machine learning. *ACM SIGKDD Explorations Newsletter*, 23(1): 32–41.

Ding, X.; Zhao, L.; and Akoglu, L. 2022. Hyperparameter sensitivity in deep outlier detection: Analysis and a scalable hyper-ensemble solution. *NeurIPS*, 35: 9603–9616.

Dwork, C.; Immorlica, N.; Kalai, A. T.; and Leiserson, M. 2018. Decoupled classifiers for group-fair and efficient machine learning. In *Conference on fairness, accountability and transparency*, 119–133. PMLR.

Emmott, A. F.; Das, S.; Dietterich, T.; Fern, A.; and Wong, W.-K. 2013. Systematic construction of anomaly detection benchmarks from real data. In *Proceedings of the ACM SIGKDD workshop on outlier detection and description*, 16–21.

Galdon Clavell, G.; Martín Zamorano, M.; Castillo, C.; Smith, O.; and Matic, A. 2020. Auditing algorithms: On lessons learned and the risks of data minimization. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, 265–271.

Garg, P.; Villasenor, J.; and Foggo, V. 2020. Fairness metrics: A comparative analysis. In *2020 IEEE international conference on big data (Big Data)*, 3662–3666. IEEE.

Hooker, S. 2021. Moving beyond "algorithmic bias is a data problem". *Patterns*, 2(4).

Jiang, S.; Cordeiro, R. L.; and Akoglu, L. 2022. D. MCA: Outlier Detection with Explicit Micro-Cluster Assignments. In *2022 IEEE International Conference on Data Mining (ICDM)*, 987–992. IEEE.

Kamishima, T.; Akaho, S.; and Sakuma, J. 2011. Fairness-aware learning through regularization approach. In *2011 IEEE 11th international conference on data mining workshops*, 643–650. IEEE.

Kusner, M. J.; Loftus, J.; Russell, C.; and Silva, R. 2017. Counterfactual fairness. *Advances in neural information processing systems*, 30.

Le Merrer, E.; Pons, R.; and Trédan, G. 2023. Algorithmic audits of algorithms, and the law. *AI and Ethics*, 1–11.

Li, N.; Goel, N.; and Ash, E. 2022. Data-centric factors in algorithmic fairness. In *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*, 396–410.

Lipton, Z.; McAuley, J.; and Chouldechova, A. 2018. Does mitigating ML's impact disparity require treatment disparity? *Advances in neural information processing systems*, 31.

Liu, F. T.; Ting, K. M.; and Zhou, Z.-H. 2008. Isolation forest. In *2008 eighth ieee international conference on data mining*, 413–422. IEEE.

Liu, H.; Ma, F.; He, S.; Chen, J.; and Gao, J. 2021. Fairness-aware outlier ensemble. *arXiv preprint arXiv:2103.09419*.

Ma, M. Q.; Zhao, Y.; Zhang, X.; and Akoglu, L. 2023. The need for unsupervised outlier model selection: A review and evaluation of internal evaluation strategies. *ACM SIGKDD Expl. Newsl.*, 25(1).

Mainardi, F. 2007. LECTURE NOTES ON MATHEMATICAL PHYSICS: Lévy Stable Distributions in the Theory of Probability. Preliminary Version. Available online at www.fracalmo.org.

Mehrabi, N.; Morstatter, F.; Saxena, N.; Lerman, K.; and Galstyan, A. 2021. A survey on bias and fairness in machine learning. *ACM computing surveys (CSUR)*, 54(6): 1–35.

Menon, A. K.; and Williamson, R. C. 2018. The cost of fairness in binary classification. In *Conference on Fairness, accountability and transparency*, 107–118. PMLR.

Pang, G.; Shen, C.; Cao, L.; and Hengel, A. V. D. 2021. Deep learning for anomaly detection: A review. *ACM computing surveys (CSUR)*, 54(2): 1–38.

Pombal, J.; Cruz, A. F.; Bravo, J.; Saleiro, P.; Figueiredo, M. A.; and Bizarro, P. 2022. Understanding Unfairness in Fraud Detection through Model and Data Bias Interactions. *arXiv preprint arXiv:2207.06273*.

Rattani, S. A. 2016. SAT: does racial bias exist? *Creative Education*, 7(15): 2151–2162.

Reddy, C.; Sharma, D.; Mehri, S.; Romero Soriano, A.; Shabanian, S.; and Honari, S. 2021. Benchmarking Bias Mitigation Algorithms in Representation Learning through Fairness Metrics. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*.

Saleiro, P.; Kuester, B.; Hinkson, L.; London, J.; Stevens, A.; Anisfeld, A.; Rodolfa, K. T.; and Ghani, R. 2018. Aequitas: A bias and fairness audit toolkit. *arXiv preprint arXiv:1811.05577*.

Schleiden, C.; Soloski, K. L.; Milstead, K.; and Rhynehart, A. 2020. Racial disparities in arrests: a race specific model explaining arrest rates across black and white young adults. *Child and adolescent social work journal*, 37: 1–14.

Shekhar, S.; Shah, N.; and Akoglu, L. 2021. Fairod: Fairness-aware outlier detection. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, 210–220.

Song, H.; Li, P.; and Liu, H. 2021. Deep clustering based fair outlier detection. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, 1481–1489.

Suresh, H.; and Guttag, J. 2021. A framework for understanding sources of harm throughout the machine learning life cycle. In *Equity and access in algorithms, mechanisms, and optimization*, 1–9.

Tang, Z.; Zhang, J.; and Zhang, K. 2023. What-is and how-to for fairness in machine learning: A survey, reflection, and perspective. *ACM Computing Surveys*, 55(13s): 1–37.

Ustun, B.; Liu, Y.; and Parkes, D. 2019. Fairness without harm: Decoupled classifiers with preference guarantees. In *International Conference on Machine Learning*, 6373–6382. PMLR.

Verma, S.; and Rubin, J. 2018. Fairness definitions explained. In *Proceedings of the international workshop on software fairness*, 1–7.

Wang, S.; Guo, W.; Narasimhan, H.; Cotter, A.; Gupta, M.; and Jordan, M. 2020. Robust optimization for fairness with noisy protected groups. *Advances in neural information processing systems*, 33: 5190–5203.

Xing, X.; Liu, H.; Chen, C.; and Li, J. 2021. Fairness-Aware Unsupervised Feature Selection. arXiv:2106.02216.

Xinying Chen, V.; and Hooker, J. N. 2023. A guide to formulating fairness in an optimization model. *Annals of Operations Research*, 326(1): 581–619.

Zafar, M. B.; Valera, I.; Rogriguez, M. G.; and Gummadi, K. P. 2017. Fairness constraints: Mechanisms for fair classification. In *Artificial intelligence and statistics*, 962–970. PMLR.

Zemel, R.; Wu, Y.; Swersky, K.; Pitassi, T.; and Dwork, C. 2013. Learning fair representations. In *International conference on machine learning*, 325–333. PMLR.

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

Zhou, C.; and Paffenroth, R. C. 2017. Anomaly detection with robust deep autoencoders. In *Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining*, 665–674.