Foundations for Unfairness in Anomaly Detection - Case Studies in Facial Imaging Data

Michael Livanos, Ian Davidson

University of California, Davis
mjlivanos@ucdavis.edu, indavidson@ucdavis.edu

Abstract

Deep anomaly detection (AD) is perhaps the most controversial of data analytic tasks as it identifies entities that are then specifically targeted for further investigation or exclusion. Also controversial is the application of AI to facial imaging data. This work explores the intersection of these two areas to understand two core questions: *"Who"* these algorithms are being unfair to and equally important *"Why"*. Recent work has shown that deep AD can be unfair to different groups despite being unsupervised with a recent study showing that for portraits of people: men of color are far more likely to be chosen to be outliers. We study the two main categories of AD algorithms: autoencoder-based and single-class-based which effectively try to compress all the instances with those that can not be easily compressed being deemed to be outliers. We experimentally verify sources of unfairness such as the under-representation of a group (e.g. people of color are relatively rare), spurious group features (e.g. men are often photographed with hats), and group labeling noise (e.g. race is subjective). We conjecture that lack of compressibility is the main foundation and the others cause it but experimental results show otherwise and we present a natural hierarchy amongst them.

Introduction

Anomaly detection (AD) is a central part of data analytics and perhaps the most controversial given that it is employed for high-impact applications that identify individuals for intervention, policing, and investigation. Its use is prevalent to identify unusual behavior in finance (transactions)(Huang et al. 2018; Zamini and Hasheminejad 2019), social media (posting and account creation)(Yu et al. 2016; Savage et al. 2014), and government services (medicare claims)(Zhang and He 2017; Bauder and Khoshgoftaar 2017).

Perhaps one of the most controversial applications of AI is to facial imaging. This is due to our faces being uniquely identifying and personal. Further, the AI's ability to identify us and make decisions (without consent) crosses many cultural and legal barriers (Garvie, Bedoya, and Frankle 2016). Existing work on facial data has focused predominantly on facial recognition, that is, given a large collection of people in a known database, identify if any of them occur in an image. Though legislation and progress have been made towards regulating facial recognition technology (Almeida, Shmarko, and Lomas 2022) other technologies in particular AD involving facial images are starting to emerge which gives rise to new ethical considerations and understanding.

Previous work (Zhang and Davidson 2021) has just begun to explore the unfairness at the intersection of AD applied to facial imaging data. For example, our previous work showed that applying AD to a collection of celebrity images overwhelmingly showed the anomalies being people of color and males (see Figure 1). However, our previous work was mainly focused on making AD algorithms fairer. We recreate our earlier results for not only the one-class AD method and the celebrity image dataset the authors used but also for the popular auto-encoder AD method and a more challenging dataset (Labeled Face In The Wild(Huang et al. 2007)).

Our experimental section attempts to address the "Who" and "Why" questions. We create a measure of unfairness (Disparate Impact Ratio (DIR)) which measures how over-represented a protected group (or its complement) is in the anomaly set. We then experimentally investigate who these algorithms are being unfair to and more nuanced questions such as is the same group always being treated unfairly regardless of algorithm. We also explore why an unsupervised algorithm can be biased. We conjecture four main foundations of unfairness, propose metrics to measure them, and outline a series of experiments to test a hypothesis on how they are structured.

The contributions of this work as are as follows:

- We study the "Who" and "Why" questions when anomaly detection is applied to facial imaging data - a topic to our knowledge has not been addressed before.
- Our experiments addressing the "Who" question show that group-level unfairness is due to an interaction between the dataset and the algorithm.
- We conjecture four main reasons for the "Why" question: i) incompressibility, ii) sample size bias (SSB), iii) spurious feature variance (SFV) within a group, and iv) attribute/group labeling noise (ALN).
- We postulate an intuitive structure to our conjectured reasons, showing it is not empirically verified, but our experimental results suggest an alternative structure.

We begin by discussing background and related work. We

Figure 1: Example of AD Being Unfair When Applied to Facial Imaging Data. Reproduced from (Zhang and Davidson 2021).

then introduce how we measure unfairness in AD and our four proposed foundations of unfairness. Next, our experimental results addressing the "Who" and "Why" questions are presented after which we discuss and conclude our work.

Background and Related Work

Applications of AD to Facial Data. AD algorithms have been used on imaging data for a variety of reasons. Perhaps the most ubiquitous is for data cleaning where anomalies are viewed as being "noise" (Ng and Winkler 2014) which are removed and then a downstream supervised algorithm is applied. However, if the AD algorithm is biased this creates an under-representation in the down-stream training tasks.

Another common use of AD is to view the outliers as "signal" and in doing so flag the outliers for extra attention. Examples include using AD to identify facial expressions to recognize emotions (Zhang et al. 2020) such as surprise. However, if the AD is biased towards some groups this will over-predict certain emotions for certain groups. Similarly, AD can be used to identify aggressive behavior (Cao et al. 2021). However, if the AD has a bias towards some groups this will incorrectly identify the group as being overly aggressive.

Source of Bias. It has been well established that supervised learning algorithms can have bias due to a variety of reasons. In particular class labeling bias has been extensively studied in the context of the Compas dataset (Angwin et al. 2016). Even though features (e.g. race) associated with this bias are removed, deep learning offers the ability to learn surrogates (e.g. zip code)(Raghavan et al. 2020).

The work on fair AD starts in 2020 (Davidson and Ravi 2020; Abraham et al. 2021) and has shown that AD algorithms can cause bias. Most work has focused on how to correct unfairness for a certain algorithm. This involves understanding the limitations in the algorithm's computation and then correcting for it. This has been explored for classic density-based methods such as LOF (Abraham et al. 2021) and deep learning methods for autoencoder (Shekhar, Shah, and Akoglu 2021), one class (Zhang and Davidson 2021) and multi-class deep AD methods. However, despite this earlier body of work, there has been surprisingly little work discussing what produces unfairness in unsupervised anomaly detection.

Four Reasons for Unfairness And Their Measurement

Here we outline our four premises for unfairness in AD and explain them at a conceptual level using Figure 1. We then describe how we measure them.

Incompressability of Data

We begin by discussing how AD methods work in particular what causes an instance to be an outlier. Deep AD methods at their core employ compression either directly or indirectly. Instances that cannot be compressed well are deemed outliers and if a group is unusual in some sense it will be unfairly treated as it will be hard to compress and hence overwhelmingly flagged as an outlier.

To understand this further, we present a common taxonomy of anomaly detection algorithms(Pang et al. 2021).

Autoencoder for Anomaly Detection. Let ϕ_e be the encoding network which maps the data X into the compressed latent space and ϕ_d be the decoding network which maps the latent representation $\phi_e(X)$ back to the original feature space(Hinton 1989). Given the network parameters θ_e, θ_d the standard reconstruction objective to train the autoencoder is:

$$argmin_\theta \left(\frac{1}{n} \sum_{i=1}^{n} \|x_i - \phi_{\theta_d}(\phi_{\theta_e}(x_i))\|^2 + R \right) \quad (1)$$

The term R denotes the regularization to the encoder and decoder. The anomaly score $s(x)$ for instance x is calculated from the reconstruction error:

$$s(x) = ||x - \phi_{\theta_d}(\phi_{\theta_e}(x))||^2 \quad (2)$$

Here clearly an outlier is defined as being an instance that the AE cannot easily compress and hence cannot easily reconstruct(Japkowicz, Myers, and Gluck 1995).

One-Class/Cluster Anomaly Detection Next, consider one class anomaly detection which is still unsupervised. Given the training data of instances $X \in R^{n \times d}$, one class AD method such as the the popular deep SVDD (Ruff et al. 2018) network is trained to map all the n instances close to a fixed center c. Denote function ϕ as a neural network with parameters θ the training objective function is:

$$argmin_\theta \frac{1}{n} \sum_{i=1}^{n} ||\phi_\theta(x_i) - c||^2 + R \quad (3)$$

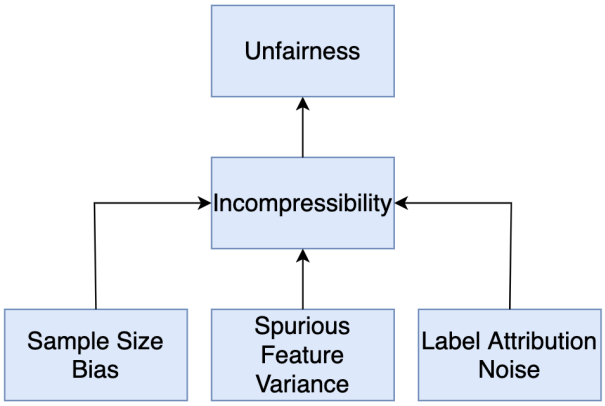


Figure 2: A Diagrammatic view of the expected reasons behind biased outlier detection.

where the term R represents the regularization function. Then the anomaly score is naturally the distance to c.

$$s(x) = ||\phi_\theta(x) - c||^2 \tag{4}$$

Here the aim is to compress all points to map onto a central point c and those that cannot be are deemed outliers.

Deep Clustering for Anomaly Detection Deep Embedded Clustering (DEC) (Xie, Girshick, and Farhadi 2016) is one of the earlier deep clustering methods that combines representation learning with clustering using a clever self-supervision approach. Recently this work was extended to perform outlier detection (Song, Li, and Liu 2021).

The distance a point is from its closest centroid $\{c_1, ...c_K\}$ is naturally an anomaly score $s(x)$:

$$s(x) = \frac{\min_{k\in[1,K]} ||\phi_{\theta_e}(x) - c_k||^2}{\max_{j\in[1,n]\wedge m_j=k} ||\phi_{\theta_e}(x_j) - c_k||^2} \tag{5}$$

where $m_j = k$ denotes instance x_j belongs to cluster c_k, K denotes the total number of clusters, and $\phi_{\theta_e}(x_i)$ is the deep learner embedding function.

The core idea here is an extension to the one-class AD method mentioned earlier but extended to k clusters.

Causes Beyond Incompressibility

The above states that outliers are inherently points that the deep learner cannot compress. Hence it is natural to consider reasons why a deep learner cannot compress a group as being a key issue for unfairness. Here we conjecture three main reasons with the view they are related to biased outliers as shown in Figure 2.

Group Underrepresentation. Here we have a group that is relatively rare in the dataset but has some unique properties so the deep learner cannot compress it well. For example in Figure 1 many outliers are African Americans as they only consist of under 15% of the dataset hence the deep learner uses its limited encoding space to encode more populous properties.

Spurious Features for Groups. In this situation, the group has a property that is not critical for the outlier detection task but is highly variable. For example in Figure 1 many groups who are over-represented in the outliers wear different styles of hats.

Label Attribution Noise. Here the labeling of a group is inaccurate and hence can be a reason a group is labeled as being overly abundant in the outlier group. For example in Figure 1 the second to the bottom line of outliers all have the tag `Male` but this is erroneous.

Measurements of Unfairness and Four Properties

Before discussing our empirical results, we first define each of the properties and how anomaly unfairness is measured. Many of these metrics are the maximum between some expression and their reciprocal. This is because the presence of a tag is equally important as the absence of a tag: for example, disparate treatment of young people and disparate treatment of old (i.e. not young) people are equally important phenomena to study. We first describe how we measure unfairness for anomalies and then how we measure our four properties.

Anomaly DIR: The unfairness of an AD algorithm's output for particular group a is measured by the disparate impact ratio (DIR), which is (Feldman et al. 2015):

$$DIR(X, AD, a) = \max\left(\frac{P(AD(X) = 1|A = a)}{P(AD(X) = 1|A = \neg a)}, \frac{P(AD(X) = 1|A = \neg a)}{P(AD(X) = 1|A = a)}\right) \tag{6}$$

Here X is the dataset the AD algorithm (AD) has made predictions (normal vs anomaly) with $AD(x) = 1$ implying x is an anomaly and $AD(x) = 0$ implying it is a normal instance, and a is the group in question. This is a natural choice for anomaly detection as it compares the rate at which different attributes are flagged as anomalies, normalized by how often the rest of the data is considered anomalous. It is also the most widely used metric in fair unsupervised learning(Verma and Rubin 2018). The range for this metric is $[1, \infty)$ with the larger the number the more unfairly group a is treated.

Incompressibility: To measure this feature, we extend the typical measure of reconstruction error into the novel metric of reconstruction ratio, which is defined:

$$RR(X, f, a) = \max\left(\frac{Loss_{MSE}(X, f(X)|A = a)}{Loss_{MSE}(X, f(X)|A = \neg a)}, \frac{Loss_{MSE}(X, f(X)|A = \neg a)}{Loss_{MSE}(X, f(X)|A = a)}\right) \tag{7}$$

Here X and a are the data used for AD and group again, with f being the autoencoder model (both encoder and decoder). The range of Equation 7 is therefore also $[1, \infty)$, where a higher number indicates that a group is harder to compress than the rest of the data. For example, a RR of 2 indicates that the attribute/group (or absence of the attribute/-group) is twice as difficult to compress than the rest of the data.

Sample Size Bias (SSB): SSB (sometimes referred to as representation bias) is determined by the proportion of that tag or lack in the dataset X and is measured as(Suresh and Guttag 2021):

$$SSB(X, a) = max(P(A = a|X), P(A = \neg a|X)) \quad (8)$$

Where X and a are again the data and the group in question. Because all groups are binary (or encoded as one-hot encoding), the range of this metric is $[0.5, 1]$, with 0.5 indicating perfect balance of the group (i.e. males and females are equally likely) and 1 indicating that the group is always on or always off. Most groups will fall between these two extremes.

Spurious Feature Variance (SFV): SFV refers to the amount of variance in the background objects in the image and is measured as a proportion of the reconstruction error of the image:

$$SFV(X, f, a, b) = 1- \max\left(\frac{Loss_{MSE}(X[b], f(X)[b]|A = a)}{Loss_{MSE}(X, f(X)|A = a)}, \frac{Loss_{MSE}(X[b], f(X)[b]|A = \neg a)}{Loss_{MSE}(X, f(X)|A = \neg a)}\right) \quad (9)$$

Where X is the data, f is the autoencoder, a the tag, and b is a bounding rectangle around the foreground/focus of the image (i.e. the face), either provided by the data or estimated(Kumar et al. 2009). As the denominator is clearly always greater than or equal to the numerator, SFV ranges between $[0, 1]$, where higher values indicate that more error comes from spurious features.

Label Attribute Noise (LAN): This is a metric of how noisy the labeling of a particular group is, as provided by the academic literature((Lingenfelter, Davis, and Hand 2022) for CelebA and (Kumar et al. 2009) for LFW). Some groups such as Gender tend to have very low LAN, whereas other tags have very high LAN such as Blurry(Kumar et al. 2009). We define LAN as:

$$ALN(X, a, a^*) = 1 - (P(a = a^*|X) + P(\neg a = \neg a^*|X)) \quad (10)$$

Where X is the data, a the group in question, and a^* the true label for the group. This property has a range $[0, 1]$ where the higher the value the less reliable the group labeling.

Experimental Results - Who Is AD Unfair To?

Here we answer the question: Who are the groups of individuals most adversely affected? Following this, we explore more nuanced inquiries, such as whether the unfairness is attributable solely to the data, the algorithm, or a combination of both. In the subsequent section, we aim to investigate the underlying reasons for the unfairness inherent in AD.

Our experiments consist of two core AD algorithms: A reconstruction based autoencoder anomaly detection algorithm (hereby referred to as AE) and Deep one-class SVDD (Ruff et al. 2018). As mentioned earlier, clustering-based AD is a generalization of one-class algorithms and the AE methods. Our datasets consist of the CelebFaces Attributes Dataset(Liu et al. 2015) (the 50,000 instance version to reduce compute) which consists primarily of popular individuals in the movies, music, or arts whilst our Labeled Faces in the Wild(Huang et al. 2007) consists of approximately 13,000 instances and includes a wider variety types of popular individuals such as politicians, sports stars, and criminals. Attribution for CelebA is given and attribution for LFW is provided by(Kumar et al. 2009). These two datasets were chosen as they are well-annotated, including analyses of labeling error, and have been extensively studied. Among all of our datasets, we test a total of 63,233 facial images covering 111 attribute tags. We examine each algorithm individually for a total of 222 data points on fairness. Both the CelebA and LFW data sets are publicly available.

For each dataset and algorithm, we determine the unfairness of each group using the Anomaly DIR. Results are collected over five random-initializations of the network and the median results for each property are reported. The list of all raw results is in the appendix, below we outline some key insights.

The Algorithms are Overwhelming Fair to Most Groups. In total amongst both the two algorithms and two datasets there are 222 groups and a frequency distribution shows that overwhelmingly the algorithms are fair with respect to over 70% of the groups as shown in Figure 3. A score of less than 1.2 indicates that the occurrence of the group in the anomalies is not more than 20% greater than the rate of all other groups (together) being labeled anomalies.

However, there are significant examples of unfairness whose properties we now discuss.

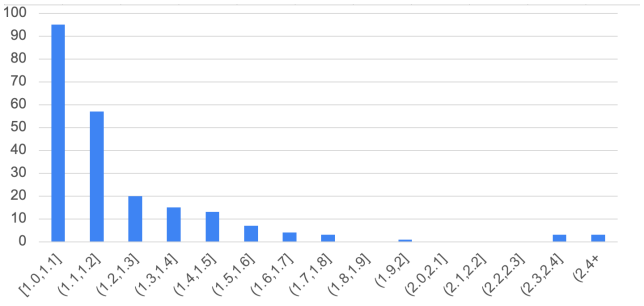


Figure 3: A frequency distribution of the Anomaly DIR score versus how often it occurs across all algorithms and datasets.

Few Groups Are Always Treated Unfairly. We found that there are several groups that are always (regardless of algorithm or dataset) treated unfairly but they are relatively rare. These include the groups centered around weight having the annotations `Chubby`, `Double-Chin` and those centered around very unusual image properties such as `Wearing-Hats`. This is not unexpected given a very rare

	CelebA	LFW
AE	**Beard (3.244)**	Beard (1.061)
	Senior (N/A)	**Senior (1.8)**
	Gray Hair (1.053)	Gray Hair (1.028)
	Unattractive (1.075)	Unattractive(1.158)
SVDD	Beard (1.267)	Beard (1.0876)
	Senior (N/A)	Senior (1.0018)
	Gray Hair (2.449)	Gray Hair (1.197)
	Unattractive (1.094)	**Unattractive (1.566)**

Table 1: Examples of groups treated unfairly only for a particular algorithm and dataset interaction. The Fairness DIR is reported in parentheses and indicates the relative overabundance of the group in the anomalies. The tag being treated unfairly in these cases is in bold. For example, people with a Beard are 3.224 times more likely to be an anomaly than a normal instance for the AE algorithm applied to the CelebA dataset, though people with beards are treated relatively fairly otherwise. Not that "Senior" is not a tag in CelebA and is therefore absent from the in these cells, and "Unattractive" in LFW is labeled "Unattractive Male".

group with unusual properties (not shared by other groups) are unlikely to be well compressed. In total less than 2% of all groups are treated unfairly all the time.

Unfairness Varies Due to Both Algorithm and Dataset. A more likely occurrence is that some groups are treated very unfairly but only for some datasets and some algorithms. Table 1 shows in bold groups treated unfairly (the Anomaly DIR is shown in parentheses) but only for that dataset and algorithm combination. For other algorithm-dataset combinations, they are treated fairly as the Table shows. This result is surprising and shows the strong interaction between the algorithm and the data. Consider that the AE method labeled `¬No Beard` (reported as "`Beard`") in the CelebA dataset at a rate over 3 times greater than the other groups. Yet, the SVDD algorithm on the very same dataset produced just a 1.27 DIR for the `Beard` group, and in the LFW dataset both algorithms the DIR was below 1.2.

The More Focused The Dataset The More Likely Unfairness Can Occur. When we aggregated all fairness DIR scores (see Appendix) for each group and all algorithms we found that the CelebA dataset (Mean DIR = 1.4) causes significantly more unfairness than the LFW dataset (Mean DIR = 1.13).

This is likely due to the CelebA dataset having a much more focused selection bias as it is limited to people who are overwhelmingly in the arts (film, television, music) whereas the LFW dataset consists of a larger representation of popular people. Hence, the definition of normality learned is very specific and there are many ways to deviate from the norm. Examples of groups that are found to be unfairly treated in the CelebA dataset but NOT the LFW dataset are: `Wearing Hat`, `Big Nose`, `Eye-Glasses`, `Goatee`, `Wavy-Hair`.

The More Focused The Algorithm The More Likely Unfairness Can Occur.

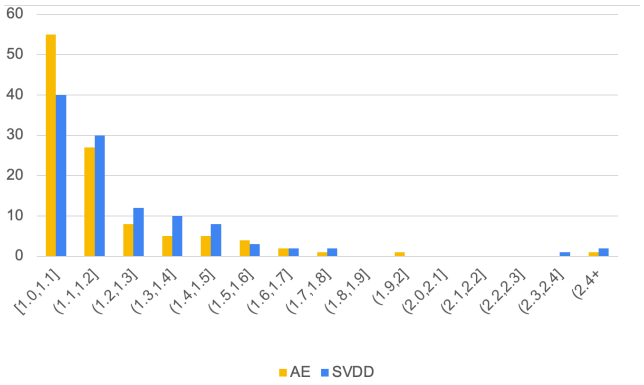


Figure 4: A frequency distribution of the Anomaly DIR score by algorithm. We see that the AE with a more flexible definition of normality is more fair.

Similarly, the way the algorithm defines normality is influential in who it identifies as an anomaly. The SVDD algorithm has the strictest definition of normality as it attempts to find just one group of normal instances (centered around c see equation 3) whereas the AE algorithm with k encoding nodes can in practice (assuming perfect disentanglement) find at least k definitions of normality. Hence not surprisingly the SVDD algorithm is more unfairer than the AE algorithm as shown by the histogram of unfairness for both algorithms in Figure 4.

Experimental Results - Why is AD Unfair

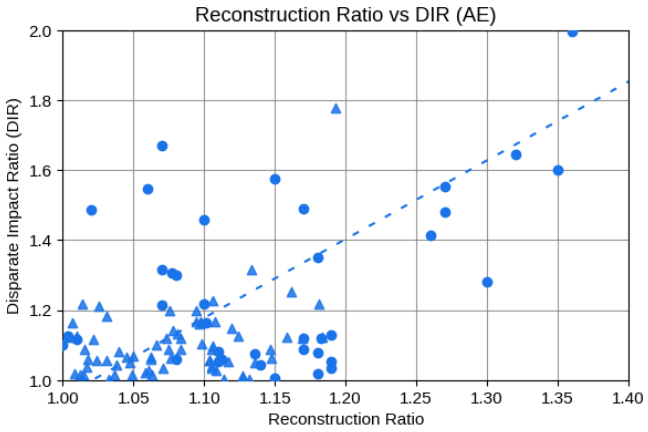
Here we attempt to experimentally answer the following questions:

- How strong are our four properties correlated to unfairness?
- How are our four properties related to each other and in particular is there a hierarchical structure to them?
- How can these properties be combined to create a model to explain unfairness in anomaly detection?

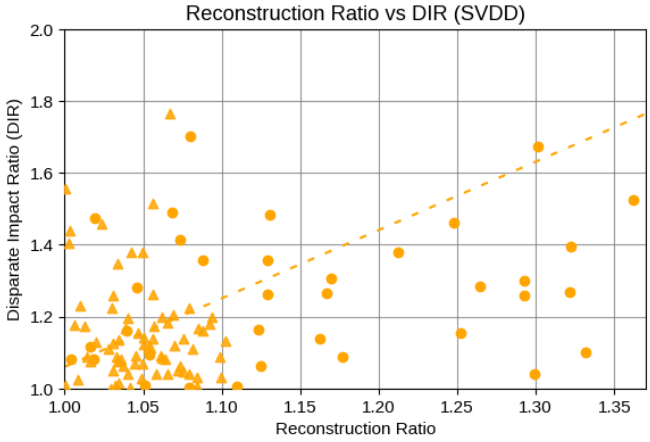
Relationship between Unfairness and Each Property

Our experiments (see Figure 5) demonstrate strong (Pearson) correlations and moderate to strong RSQ (R-squared values of the regression trendline) for each of the properties studied. Each plot shows the results for two datasets (CelebA and LFW) with each data point representing a group of individuals. A positive trend line indicates positive Pearson correlation (see sub-titles of plots for exact values) and we see that incompressability is the most strongest property correlated with unfairness, then Spurious features, then Attribute label noise, and finally Sample Size Bias. This is an interesting result as earlier seminal results showed that AD using facial images (Zhang and Davidson 2021) was unfair due to an under-representation of African Americans and Males in the underlying datasets.

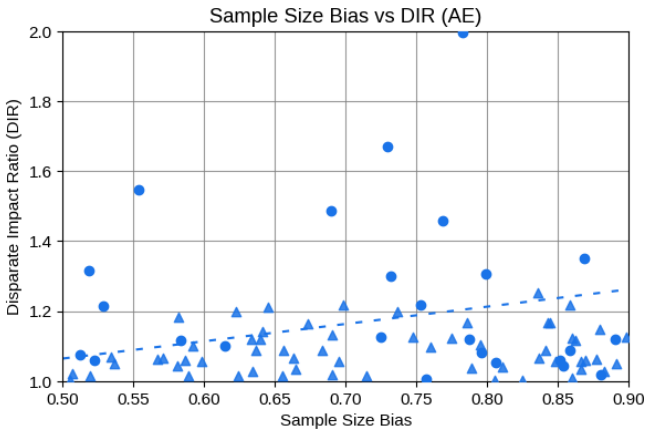
However, it is also clear that no individual property explains unfairness completely by itself. This is shown as each



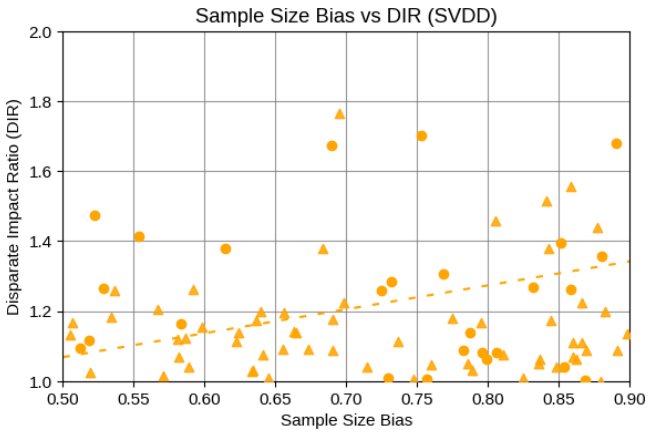
(a) Corr: 0.568, RSQ: 0.334



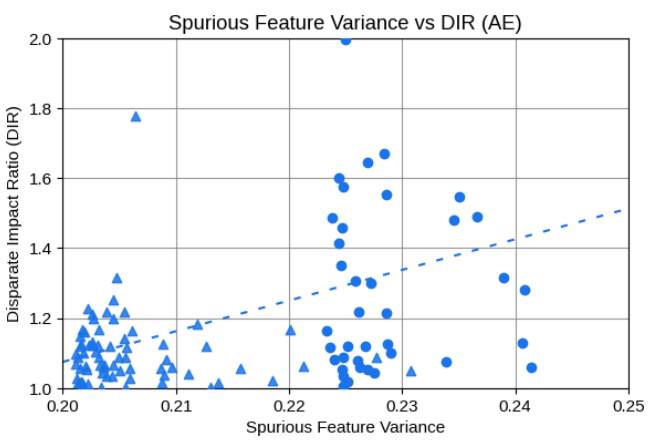
(b) Corr: 0.523, RSQ:0.273



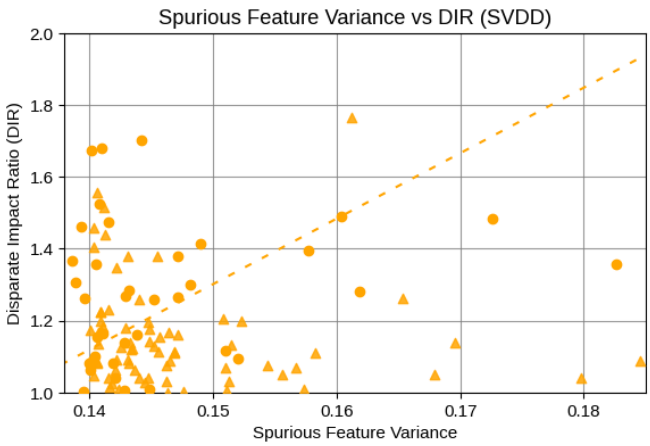
(c) Corr: 0.220, RSQ:0.114



(d) Corr: 0.251, RSQ:0.128



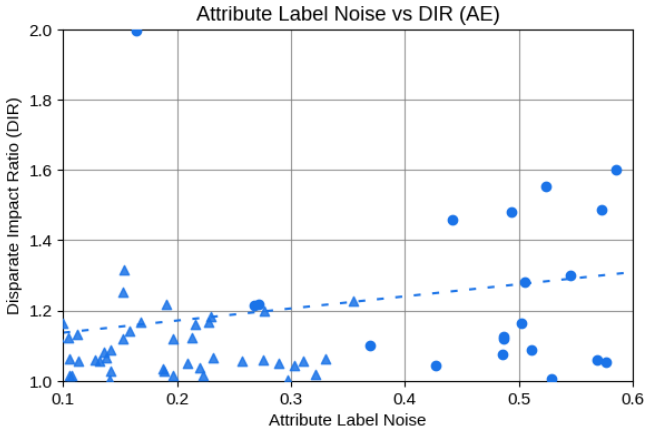
(e) Corr: 0.337, RSQ:0.148



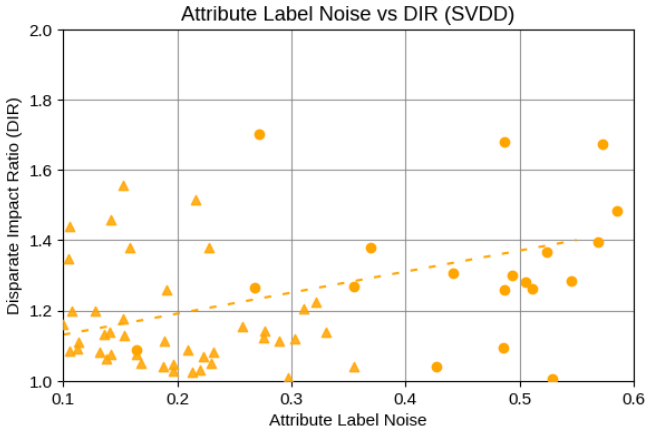
(f) Corr: 0.473, RSQ:0.224

Figure 5: (Figure continues on next page)

.



(g) Corr: 0.261, RSQ:0.167



(h) Corr: 0.328, RSQ:0.108

Figure 5: Plot of different properties against their DIR (unfairness) with the larger the value the more of the property/unfairness. Trendlines are created by minimizing R^2 values. Each mark represents one group. Color denotes algorithm (blue for the AE anomaly detector and orange for the SVDD anomaly detector) and mark denotes dataset (circle for CelebA, triangle for LFW).

graph has points that not only do not fit the trendline, but are contradictory to the relationship implied by the overall data. Further investigation (see next subsection) reveals that when one property fails to explain why that attribute is anomalous, another one typically will.

For example, the group `Bags Under Eyes` (from CelebA) under the AE model has a reconstruction ratio of only 1.077 (it is easy to compress), but a DIR of 1.31 (it is treated unfairly). Following the trend, the expected reconstruction ratio at a group with this DIR would be approximately 1.17. Further, this group has only 20.1% representation, though looking at the DIR one would expect only half that. This group's treatment, however, is explained by the spurious feature variance, as it sits nearly perfectly on the trendline. Similarly, the group `Gray Hair` (from LFW) under Deep SVDD was towards the far end of spurious feature variance at 0.180, but has extremely low anomaly DIR score at 1.04 (i.e. was treated fairly), though it sits just above the trendline for attribute label noise at 1.05.

A full list of these attributes and their squared error for all trendlines is available in the Appendix, and one can see that every tag can be explained by at least one of these properties with high fidelity, with the average sum of square errors being only 0.00351 (std 0.006498), supporting our claim that unfairness in anomaly detection setting can be typically explained by one of these four properties. This claim is rigorously tested in Section .

Relationship between Multiple Properties

We also examine the correlation between the different properties. This analysis is useful in examining potential redundancies and creating our model of unfairness for anomaly detection. Figure 6 examines these relationships. Some features are, indeed, positively correlated with each other, though none have high enough correlation to suggest that they are redundant with each other. In the subsequent subsection, we examine this claim more rigorously via a hypothesis test.

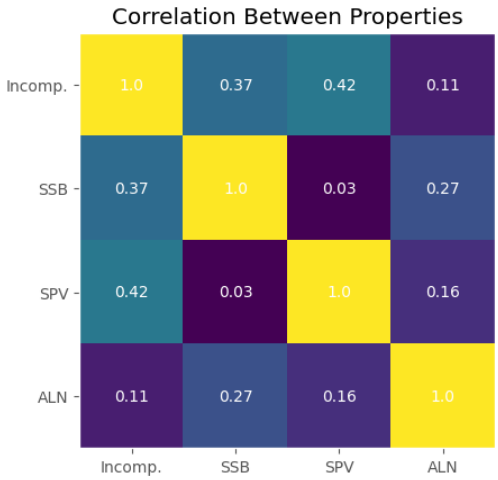


Figure 6: Correlation matrix for all four properties of the model. Pearson correlation is written in each box and is consistent with color (yellow is large, purple is small).

Hypothesis Testing of Relationship Claims In order to test our claims, we create four hypotheses that we verify through hypothesis significance-testing. Those are:

- H1: No individual property is sufficient to always explain unfairness.
- H2: The properties, when combined into a multiple regression, are sufficient to explain unfairness.
- H3: No properties of the multiple regression are redundant and all are needed.
- H4: The results of H2 are significant in that when one property fails to predict unfairness, another does.

Null hypothesised $H1_0 - H4_0$ are constructed straightforwardly. To create the significance test for H1, we perform an F-test on individual regression models crafted from the relationship between each property and DIR. The results of this F-Test (visualized in Figure 7) indicate that individual properties are reasonable though comparably weak predictors of unfairness, with P-values ranging from 0.0137-0.0986 for the AE model and 0.0279-0.0571 for Deep SVDD. Therefore, we reject the null hypothesis $H1_0$ and validate hypothesis H1.

To test hypotheses H2 and H3, we construct a multiple-regression model. Specifically, this is a stacked multiple regression where the meta-function selects the best individual model for the datum. To validate H2, we create such a multiple-regression using all four of the properties (the "full" model). This yields P-Values of 0.00589 for the AE model and 0.0127 for Deep SVDD, significantly lower than those of the respective single-regression models, and indicating that using all four properties is sufficient to explain how unfairness occurs. We reject the null hypothesis $H2_0$ and validate hypothesis H2.

For H3, we conduct a similar experiment except we leave one property out. In every case, the resulting multiple regression models were worse than the full model, with P-Values ranging from 0.00674-0.0109 for the AE model and 0.0138-0.0164 for Deep SVDD, all greater than that of the full model, indicating that every property is necessary and none are redundant. We reject the null hypothesis $H3_0$ and validate hypothesis H3.

One may object to the multiple-regression models used above, given that the model as described will monotonically increase in predictive power given more properties. It is important to note that this model matches the central claim of this paper - that unfairness with respect to a group occurs because of one of the four properties described, though one may still be wary of the statistical significance of the reported results given the technique. To resolve these concerns, we demonstrate that our model is not just combining the predictive power of four different already powerful predictors, but rather when one model fails it is because it is explained by one of the other properties.

To validate this claim, we construct fabricated distributions similar to those of Figure 5. Specifically, unfairness is kept the same, and we create distributions of random fake data which has the same correlation and RSQ as all of those shown. This is accomplished by, for each property, finding random points (sampled across a uniform distribution) along the X-axis, giving them fabricated values perfectly in line with the correlation, and then adding noise such that the correlation is maintained and the RSQ matches that of the actual measured properties. Then, we create the same full model of the multiple regression and measure the P-value. We repeat this process 10,000 times to get 10,000 such distributions.

The distributions therefore should be statistically similar to our real data, but there is no reason to believe that when one of the fabricated models fails, another will explain the unfairness. To validate hypothesis H4, we measure the number of times the fake distributions produce P-values under that of the real data. If the statically similar fabricated data

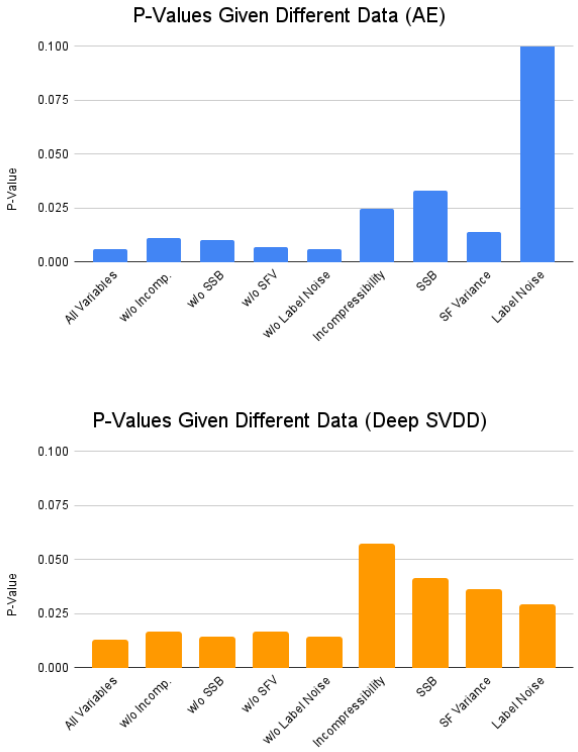


Figure 7: P-Values for the hypotheses H1-H3. The leftmost bar demonstrates that, when all properties are considered, unfairness can be predicted with a very high degree of precision, rejecting the null hypothesis $H2_0$. The next three rows demonstrate that the model is not as powerful if one property was left out, rejecting the null hypothesis $H3_0$. Finally, the higher P-values for the simple regressors indicate that no single feature can be used as a model of unfairness, rejecting null hypothesis $H1_0$.

cannot match the predictive performance of our models, this would validate hypothesis H4.

In the case of the AE model, the fabricated data averaged a P-value of 0.0194 with a standard deviation of 0.00629 and never beat the full model's P-value of 0.00589. Similarly, the model simulating Deep SVDD's data yielded an average P-value of 0.0173 with a standard deviation of 0.00304. Out of the 10,000 trials, only 5 yielded lower P-values. Therefore, we reject the null hypothesis $H4_0$ and validate hypothesis H4. Our model does not simply take four independent good predictors of anomaly and get good statistical results but rather holds the property that when one fails, another property explains it.

A Proposed Model Of Unsupervised Unfairness Relationships

Given the resulting hypothesis tests, we craft our model of unfairness in unsupervised learning. Figure 8 provides a graphical representation of this model. Edges between prop-

erties indicate a relationship (binarized to be correlated at ≥ 0.15). This is supported by the high correlation between each of these properties and unfairness (Figure 5), the result that the properties together form a uniquely powerful multiple-regression to explain unfairness (H2, H4), that no single feature could do this alone (H1), and that no property is redundant (H3).

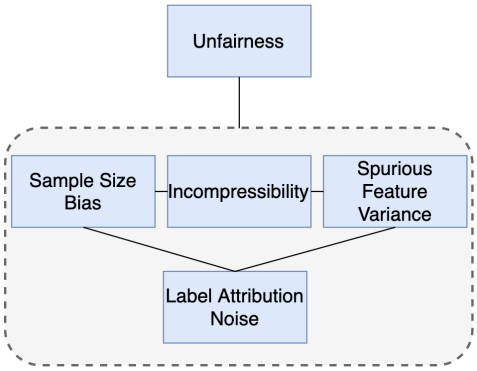


Figure 8: Our model of unfairness determined from our stacked multi-regression model. Compare with the expected model without any analysis in Figure 2.

Conclusion, Limitations, and Future Work

We study the intersection of the controversial deep AD algorithm with facial imaging data to address the "Who" and "Why" questions. We found that overwhelmingly both autoencoder and one-class deep AD algorithms are fair to most groups. However, due to the compression-based focus, they are unfair to some sub-groups.

With regard to the "Who" question we found that it was rare to be consistently unfair to the one group and instead unfairness was due to the interaction of the data and the algorithm. In particular, the more focused the dataset and algorithm the more unfairness was found.

Our study of the "Why" question aimed at developing a deeper understanding on the effect of data related factors on the fairness as well as detection performance of OD algorithms. We postulated four hypotheses and found all to be statistically significant by rejecting the null hypothesis. The first hypothesis is that no single property alone is sufficient to explain unfairness. The second hypothesis is when combined the properties can explain unfairness. The third hypothesis is that all properties are relevant and none are redundant and finally, the fourth hypothesis is that the combination of properties is meaningful beyond the predictive power of each individual property.

Limitations. The use of groups may have varying degrees of applicability to real-world fairness scenarios. For example, some groups such as `Male`, `Black` and `Young` correspond to legally recognized protected classes (88th United States Congress 1964; 90th United States Congress 1967), while others such as `Goatee`, `Wearing Hat` and `attractive` may not. However, we believe that this study still provides meaningful insights into the mechanism of unfairness with respect to different people. Real-world protected attributes may be of varying degrees of visibility, as do our groups, and our analysis reflects this.

Future work. Remediation strategies to improve fairness are left out of scope of our investigation. We briefly discuss them here. Fairness interventions are typically grouped into three: pre-, post-, and in-processing strategies, which respectively, modify the input data, modify the output scores or decisions, and account for fairness during model training.

As we showed, AD unfairness can stem from algorithmic bias alone in the face of natural heterogeneities in the data among or within groups. When this is the case, preprocessing strategies become voided as it is not clear how to modify organic, unbiased data. Post-processing could select different thresholds for each group separately, as in (Corbett-Davies et al. 2017; Menon and Williamson 2018), where the group-specific thresholds could either be "natural" cut-off values, or selected to optimize demographic parity if it is a desired fairness metric. Note that metrics that involve true labels cannot be optimized due to lack of any ground truth during training. In-processing techniques are also limited to only enforcing demographic parity, which as we showed, remains susceptible to unfairness. One such strategy that has not been applied to OD is decoupling, as in (Dwork et al. 2018; Ustun, Liu, and Parkes 2019), where a different detector is trained for each group, while optimizing a joint loss.

We remark that post-processing and decoupling exhibit treatment disparity as they both assume it to be ethical and legal to use the sensitive attribute at test (decision) time - in particular, to select which threshold or detector to employ on a given new sample. When there are differences *among* groups, coming to terms with treatment disparity might be the only get-around to mitigating disparate impact, as argued previously (Lipton, McAuley, and Chouldechova 2018). These solutions, however, do not address unfairness against heterogeneous subpopulations *within* groups, i.e. within-group discrimination. Here, one direction is to explore clustering-based OD algorithms. Alternatively, establishing a more nuanced or granular sensitive attribute, labeling each subpopulation differently.

References

88th United States Congress, T. 1964. Civil Rights Act of 1964. Public Law 88-352, 78 Stat. 241.

90th United States Congress, T. 1967. Age Discrimination in Employment Act of 1967. Public Law 90-202, 81 Stat. 602.

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

Almeida, D.; Shmarko, K.; and Lomas, E. 2022. The ethics of facial recognition technologies, surveillance, and accountability in an age of artificial intelligence: a comparative analysis of US, EU, and UK regulatory frameworks. *AI and Ethics*, 2(3): 377–387.

Angwin, J.; Larson, J.; Mattu, S.; and Kirchner, L. 2016. Machine Bias. *ProPublica*.

Bauder, R. A.; and Khoshgoftaar, T. M. 2017. Multivariate anomaly detection in medicare using model residuals

and probabilistic programming. *The Thirtieth International Flairs Conference*.

Cao, R.; Liu, X.; Zhou, J.; Chen, D.; Peng, D.; and Chen, T. 2021. Outlier detection for spotting micro-expressions. *2021 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, 3006–3011.

Corbett-Davies, S.; Pierson, E.; Feller, A.; Goel, S.; and Huq, A. 2017. Algorithmic decision making and the cost of fairness. *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. *ECAI 2020*, 2465–2472.

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

Feldman, M.; Friedler, S. A.; Moeller, J.; Scheidegger, C.; and Venkatasubramanian, S. 2015. Certifying and Removing Disparate Impact. 259–268.

Garvie, C.; Bedoya, A.; and Frankle, J. 2016. The Perpetual Line-Up: Unregulated Police Face Recognition in America.

Hinton, G. E. 1989. Connectionist Learning Procedures. *Artificial Intelligence*, 40(1-3): 185–234.

Huang, D.; Mu, D.; Yang, L.; and Cai, X. 2018. CoDetect: Financial Fraud Detection With Anomaly Feature Detection. *IEEE Access*, 6: 19161–19174.

Huang, G. B.; Ramesh, M.; Berg, T.; and Learned-Miller, E. 2007. Labeled Faces in the Wild: A Database for Studying Face Recognition in Unconstrained Environments. Technical Report 07-49, University of Massachusetts, Amherst.

Japkowicz, N.; Myers, C.; and Gluck, M. A. 1995. A Novelty Detection Approach to Classification. 518–523.

Kumar, N.; Berg, A. C.; Belhumeur, P. N.; and Nayar, S. K. 2009. Attribute and Simile Classifiers for Face Verification. *2009 IEEE 12th International Conference on Computer Vision*, 365–372.

Lingenfelter, B.; Davis, S.; and Hand, E. 2022. A Quantitative Analysis of Labeling Issues in the CelebA Dataset. *Advances in Visual Computing. ISVC 2022. Lecture Notes in Computer Science*, 13598.

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, Z.; Luo, P.; Wang, X.; and Tang, X. 2015. Deep Learning Face Attributes in the Wild. *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*.

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

Ng, H.-W.; and Winkler, S. 2014. A data-driven approach to cleaning large face datasets. *2014 IEEE international conference on image processing (ICIP)*, 343–347.

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.

Raghavan, M.; Barocas, S.; Kleinberg, J.; and Levy, K. 2020. Mitigating bias in algorithmic hiring: evaluating claims and practices. 469–481.

Ruff, L.; Vandermeulen, R.; Goernitz, N.; Deecke, L.; Siddiqui, S. A.; Binder, A.; Müller, E.; and Kloft, M. 2018. Deep One-Class Classification. PMLR 80: 4393–4402.

Savage, D.; Zhang, X.; Yu, X.; Chou, P.; and Wang, Q. 2014. Anomaly detection in online social networks. *Social networks*, 39: 62–70.

Shekhar, S.; Shah, N.; and Akoglu, L. 2021. Fairod: Fairness-aware outlier detection. *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. *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.

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

Verma, S.; and Rubin, J. 2018. Fairness definitions explained. 1–7.

Xie, J.; Girshick, R.; and Farhadi, A. 2016. Unsupervised Deep Embedding for Clustering Analysis.

Yu, R.; Qiu, H.; Wen, Z.; Lin, C.; and Liu, Y. 2016. A survey on social media anomaly detection. *ACM SIGKDD Explorations Newsletter*, 18(1): 1–14.

Zamini, M.; and Hasheminejad, S. M. H. 2019. A comprehensive survey of anomaly detection in banking, wireless sensor networks, social networks, and healthcare. *Intelligent Decision Technologies*, 13(2): 229–270.

Zhang, G.; Luo, T.; Pedrycz, W.; El-Meligy, M. A.; Sharaf, M. A. F.; and Li, Z. 2020. Outlier processing in multimodal emotion recognition. *IEEE Access*, 8: 55688–55701.

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

Zhang, W.; and He, X. 2017. An anomaly detection method for medicare fraud detection. *2017 IEEE International Conference on Big Knowledge (ICBK)*, 309–314.