SimFair: A Unified Framework for Fairness-Aware Multi-Label Classification

Tianci Liu, Haoyu Wang, Yaqing Wang, Xiaoqian Wang, Lu Su, Jing Gao*

School of Electrical and Computer Engineering, Purdue University, West Lafayette, USA, 47907
{liu3351, wang5346, wang5075, joywang, lusu, jinggao}@purdue.edu

Abstract

Recent years have witnessed increasing concerns towards unfair decisions made by machine learning algorithms. To improve fairness in model decisions, various fairness notions have been proposed and many fairness-aware methods are developed. However, most of existing definitions and methods focus only on single-label classification. Fairness for multi-label classification, where each instance is associated with more than one labels, is still yet to establish. To fill this gap, we study fairness-aware multi-label classification in this paper. We start by extending Demographic Parity (DP) and Equalized Opportunity (EOp), two popular fairness notions, to multi-label classification scenarios. Through a systematic study, we show that on multi-label data, because of unevenly distributed labels, EOp usually fails to construct a reliable estimate on labels with few instances. We then propose a new framework named **Sim**ilarity s-induced **Fair**ness (s_γ-SimFair). This new framework utilizes data that have similar labels when estimating fairness on a particular label group for better stability, and can unify DP and EOp. Theoretical analysis and experimental results on real-world datasets together demonstrate the advantage of s_γ-SimFair over existing methods on multi-label classification tasks.

Introduction

Nowadays, machine learning algorithms play increasingly more important roles in decision-making for a broad spectrum of applications, such as applicant screening in job markets, credit risk analysis, and recommendation systems. However, recent studies (Barocas and Selbst 2016; Buolamwini and Gebru 2018; Dressel and Farid 2018) have discovered that machine learning algorithms tend to make discriminatory decisions. For example, a dataset may contain records of physicians most of whom are male. As a result, a job screening algorithm trained on this dataset may unfairly predict if a person is suitable for a *physician* position based on their gender, instead of education background or professional experience. Obviously, such favorable prediction for male applicants is *unfair* to female applicants.

Formally, the algorithmic fairness issue refers to the phenomenon that machine learning algorithms make discriminatory decisions across different demographic subgroups

and give favorable predictions for some particular subgroups. Intuitively, discriminatory decisions are associated with some demographic features contained in the data, such as age, gender, and race. These features are referred to as *sensitive* features. Ideally, a fair model should be able to make decisions independent of sensitive features. Towards this end, different fairness notions (Pedreshi, Ruggieri, and Turini 2008; Dwork et al. 2011; Hardt, Price, and Srebro 2016; Chouldechova and Roth 2020) have been proposed. Among them, Demographic Parity (DP) (Pedreshi, Ruggieri, and Turini 2008) and Equalized Opportunity (EOp) (Hardt, Price, and Srebro 2016) are two of the most widely-used definitions. DP requires a model's decision to be independent of sensitive features, achieving a population-level fairness (Edwards and Storkey 2015; Madras et al. 2018; Creager et al. 2019). However, Dwork et al. (2011) showed that such population level fairness does not necessarily guarantee fairness in all label groups. To address this limitation, Hardt, Price, and Srebro (2016) proposed to take label information into consideration and defined EOp and its stronger version Equalized Odds (EO). Specifically, EOp requires the decision to be independent of sensitive features conditionally in the label group receiving an *favorable* outcome (Hardt, Price, and Srebro 2016). Examples of favorable outcomes include "being admitted to a position" in job screening, and "approval of credit card application". For brevity, we refer to the label group in which each individual receives the *favorable* outcome as the *advantaged* group. With a more restrictive fairness definition, EO further requires that the decision is independent of sensitive features in each label group, including not only the advantaged group but also the groups receiving other outcomes.

Some methods have been proposed based on the aforementioned fairness definitions. However, they are focused only on scenarios where each instance is associated with a single target label (Hardt, Price, and Srebro 2016; Woodworth et al. 2017; Zafar et al. 2017). In many real-world applications, multiple labels need to be predicted for an instance. For example, in job screening, an applicant may apply for multiple positions, and the admission decision of each position is a target label of the applicant. Similarly, undergraduates usually submit applications to multiple programs when applying to graduate schools, and thus associate themselves with multiple target labels of admission. Scenarios where each instance

*Corresponding author

is associated with more than one target labels are termed as **multi-label classification** (Zhang and Zhou 2014). Obviously, fairness concerns also exist in multi-label classification scenarios. One straightforward approach toward fairness in a multi-label scenario is to decompose multi-label classification into multiple binary classification tasks, each of which judges whether a label is associated with an instance or not, and then apply existing fairness metric separately on each binary classification task (Zhang et al. 2018). However, this naive approach ignores one unique property of multi-label classification, i.e., the correlations among labels. Again, take job screening as an example. Applicants usually apply for positions with similar requirements of skill sets and experiences at the same time, and thus the application outcomes (labels) are correlated. Ignoring such correlations among labels would lead to unsatisfactory classification and let alone fairness results. On the other hand, existing multi-label classification methods consider the correlations among labels but cannot enforce fairness in the predictions.

Therefore, it is critical to define fairness directly in the context of multi-label classification. Unfortunately, we did not find existing work along this direction. This motivates us to study this problem. In a multi-label scenario, since different target labels usually occur together, it is more natural to treat their combinations as an *advantaged* outcome (label). For example, the *advantaged* group in the job screening example with two possible positions can be the applicants who "received offers of position A and position B". Note that this definition allows us to define more general and complex advantaged groups by specifying more than one favorable labels and requires fairness on all of them.

In practice, the discussed fairness objective is usually achieved by incorporating some fairness notions into optimization (Mohler et al. 2018; Scutari, Panero, and Proissl 2021). Such an optimization is non-trivial when tackling fairness issue based on this extended concept of *advantaged* group in multi-label classification, where collected data is usually not evenly distributed among different labels (Dekel and Shamir 2010). When few instances are in the *advantaged* group (i.e., the group that has the favorable label), it may introduce unreliable fairness constraints into optimization and degrade the fairness performance. In this work, we show that the aforementioned optimization challenges can be alleviated by utilizing information sharing among labels. Intuitively, we *group* data with different but similar labels to alleviate data shortage issue, and then enable an EOp-like framework to incorporate fairness constraints on advantaged groups. This will be formalized in Section . We refer to our framework as Similarity s-induced Fairness (s_γ-SimFair), highlighting the crucial requirement of a similarity measure between different labels in the data grouping step.

The proposed framework s_γ-SimFair is principled in the sense that it unifies DP and EOp, bringing the flexibility of leveraging a population level fairness or fairness on some particular label groups (e.g., the *advantaged* group) per desires. The DP and EOp are two extreme cases of s_γ-SimFair. When treating all labels as equally similar and ignoring their differences, we end up with one *label* group of data, in which s_γ-SimFair becomes DP. On the contrary, if two labels are

similar only if they are the same, then each *label* group involves only one label, in which s_γ-SimFair becomes EOp. Moreover, s_γ-SimFair is able to enforce a restrictive fairness notion on the advantaged group even when data is inadequate by utilizing information from other similar label groups.

Our main contributions are summarized below.

- To the best of our knowledge, we are the first to investigate fairness in a multi-label classification setting. We extend DP and EOp to the multi-label classification setting, and recognize the challenge of achieving EOp based on both theoretical and empirical studies.

- To handle the recognized challenge, we propose a novel framework, namely s_γ-SimFair, to achieve the fairness objective for multi-label classification even when imbalanced label distributions exist. We further support the proposed framework with rigorous theoretical analysis.

- The comprehensive experiments show that the proposed framework s_γ-SimFair is able to achieve competitive and even better performance in term of DP and EOp compared to that of directly incorporating DP and EOp into optimization respectively.

Related Work

Algorithmic Fairness

Most existing fairness definitions fall into two categories: group fairness (Pedreshi, Ruggieri, and Turini 2008; Dwork et al. 2011; Hardt, Price, and Srebro 2016; Chouldechova and Roth 2020) and individual fairness (Dwork et al. 2011). Group fairness requires that the probability of being assigned to a group by a model is independent of sensitive features such as gender, age and race. For example, Demographic Parity (DP) requires that the prediction is independent of sensitive features, while Equalized Odds (EO) and Equalized Opportunity (EOp) require that the prediction is conditionally independent of sensitive features in each or some label group. When labels are binary, this is equivalent to requiring an equality of true and false positive rates across different demographic subgroups. Modifications of DP and EO (EOp) have also been studied. For example, in Pleiss et al. (2017), a relaxed condition is required by replacing EO with some calibration. Individual fairness, on the other hand, requires that a model treats similar individuals similarly (Dwork et al. 2011). In this work we focus on group fairness.

In order to correct the unfairness of models, many methods have also been proposed, which can be classified into one of the following three categories: pre-processing biased datasets, in-processing models during training, and post-processing the outputs of models. In-processing is usually the most effective way to intervene an unfair model (Petersen et al. 2021), which can be done by penalizing unfair predictions directly (Mohler et al. 2018; Scutari, Panero, and Proissl 2021), or by disentangling some intermediate representations (on which final predictions are made) from sensitive features (Locatello et al. 2019; Creager et al. 2019). Nevertheless, penalty-based methods are still good and effective starting points to mitigate unfairness (Mary, Calauzenes, and El Karoui 2019; Kamishima et al. 2012).

Multi-Label Classification

Multi-label classification is a general family of classification tasks where each instance is associated with multiple target labels. This task has very broad applications (El Kafrawy, Mausad, and Esmail 2015), such as recommendation systems (Zheng, Mobasher, and Burke 2014; Zhang et al. 2020), multi-object detection (Gong et al. 2019; Zhao et al. 2020), and text classification (Yang et al. 2009; Nam et al. 2014).

Methods for multi-label classification can be grouped into two categories (Zhang and Zhou 2014; Tsoumakas, Katakis, and Vlahavas 2006): problem transformation and algorithm adaptation. Problem transformation tackles multi-label classification by transforming the task into other well-defined tasks. One possible transformation is binary relevance (Boutell et al. 2004), which ignores all dependencies among different targets and predicts each target separately. Classifier chain, the other extreme case, learns the joint distribution of different labels by applying the chain rule of probability (Read et al. 2011). In summary, these problem transformation multi-label classification tasks into other well-established learning problems and adopt existing methods to solve them (Tsoumakas and Vlahavas 2007; Fürnkranz et al. 2008). Algorithm adaptation, on the other hand, modify existing algorithms such as kNN (Zhang and Zhou 2007) and decision tree (Clare and King 2001) to model multi-label data directly. We refer readers to Zhang and Zhou (2014); Tsoumakas, Katakis, and Vlahavas (2006) for more details.

Deep learning has advanced multi-label classification as well (Liu et al. 2021). Recently, Chen, Xue, and Gomes (2018); Bai, Kong, and Gomes (2020) revisited the Multivariate Probit (MP) model (Chib and Greenberg 1998) with the equipment of deep learning tools. MP model assumes that the joint distribution of labels is controlled by a multivariate Gaussian random variable, and the probability of a label is determined by the cumulative density function (CDF) at the value of this Gaussian variable. The correlations in the Gaussian variable allows the model to capture pairwise dependencies in a multi-label setting. Chen, Xue, and Gomes (2018) parameterized the MP model with a deep neural network resulting in the deep Multivariate Probit model (DMVP), and Bai, Kong, and Gomes (2020) proposed to combine DMVP and variational autoencoder (Kingma and Welling 2014) to obtain better performance.

Methodology

In this section, we propose s_γ-SimFair, a flexible framework to unify Demographic Parity (DP) and Equalized Opportunity (EOp). We start with deriving DP and EOp in multi-label scenarios. Then we provide a systematic study on the challenges of estimating EOp in multi-label scenarios. We propose s_γ-SimFair based on the these studies to achieve the fairness objective even when imbalanced label distributions exist.

Preliminaries

Notations Throughout this paper, we use bold capital letters (e.g., $\mathbf{X}$) to denote matrices, bold lowercase letters (e.g., $\boldsymbol{x}$) to denote (column) vectors, and calligraphic letters (e.g., $\mathcal{X}$) to denote spaces. Finally, capital P denotes a probability

and lowercase p denotes a distribution. We summarize notations used in this paper in appendix[1] for better readability.

Consider a dataset that contains N samples $\mathcal{D} = \{(\boldsymbol{x}^{(i)}, a^{(i)}, \boldsymbol{y}^{(i)})\}_{i=1}^N$. Without loss of generality, we assume each sample is associated with M non-sensitive features $\boldsymbol{x}^{(i)} \in \mathcal{X} = \mathbb{R}^M$, a K-way scalar sensitive feature $a^{(i)} \in \mathcal{A} = \{1, \ldots, K\}$ where K is the number of demographic subgroups (e.g., $K = 2$ if *gender* is the sensitive feature that takes *female* and *male*), and L binary labels $\boldsymbol{y}^{(i)} \in \mathcal{Y} = \{0, 1\}^L$. We further assume N samples are drawn from an unknown underlying distribution p over space $(\mathcal{X} \times \mathcal{A} \times \mathcal{Y})$, and use $(\boldsymbol{x}, a, \boldsymbol{y}) \sim p$ to denote a random sample. To avoid ambiguity, for $\boldsymbol{y} = (y_1, \ldots, y_L)$, we call $\boldsymbol{y}$ a *label*, and $y_l \in \{0, 1\}$ the l-th *target*, where $y_l = 1$ indicates the presence of l-th target. We use $h : \mathcal{X} \to \mathcal{Y}$ to denote a multi-label classifier that predicts label based on non-sensitive features. Under these settings, $L = 1$ corresponds to single-label classification, and $L > 1$ corresponds to multi-label classification.

Multi-Label Classification Prediction We consider a wide family of multi-label classifiers that satisfy $h = f \circ g : \mathcal{X} \to [0, 1]^L \to \mathcal{Y}$. In particular, a classifier first predicts $\tilde{\boldsymbol{y}} = g(\boldsymbol{x})$, the probability of the presence of L targets given $\boldsymbol{x}$. Then l-th target prediction is given by $\hat{y}_l = \mathbf{1}(\tilde{y}_l \geq 0.5)$ elementwisely, in which f denotes this elementwise thresholding function. This family of classifiers is capable of capturing dependencies between different targets by predicting $\tilde{\boldsymbol{y}}$ given $\boldsymbol{x}$ jointly as shown in Chen, Xue, and Gomes (2018); Bai, Kong, and Gomes (2020).

DP and EOp on Multi-Label Classification

DP and EOp Condition In this section, we establish DP and EOp condition in multi-label scenarios. For classifier $h = f \circ g : \mathcal{X} \to \mathcal{Y}$ and random sample $(\boldsymbol{x}, a, \boldsymbol{y}) \sim p$, h is *fair* in terms of (1) DP if $\hat{\boldsymbol{y}} \perp a$; and (2) EOp if $\hat{\boldsymbol{y}} \perp a \,|\, \boldsymbol{y}_{adv}$, where $\boldsymbol{y}_{adv} \in \mathcal{Y}$ denotes some advantaged label where only favorable outcomes (e.g., "received offer" in the job screening example) present. In essence, DP requires predictions to be independent with sensitive variables, and EOp requires conditional independence to hold on label $\boldsymbol{y}_{adv}$. As assumed, prediction $\hat{\boldsymbol{y}}$ depends on predicted probability $\tilde{\boldsymbol{y}}$ elmentwisely, therefore distribution of $\hat{\boldsymbol{y}}$ is fully parameterized by $\tilde{\boldsymbol{y}}$. Proposition 0.1 gives a condition for DP and EOp to hold in multi-label classification.

Proposition 0.1 (DP and EOp condition for multi-label classifier)**.** For a multi-label classifier that takes the form $h = f \circ g$, where $\tilde{\boldsymbol{y}} = g(\boldsymbol{x})$ is the predicted probability and $\hat{\boldsymbol{y}} = f(\tilde{\boldsymbol{y}})$ is computed elementwisely, DP and EOp hold if for any $k \in \mathcal{A}$

$$\text{DP:} \quad \mathbb{E}[\tilde{\boldsymbol{y}} \,|\, a = k] = \mathbb{E}[\tilde{\boldsymbol{y}}]$$
$$\text{EOp:} \quad \mathbb{E}[\tilde{\boldsymbol{y}} \,|\, a = k, \boldsymbol{y} = \boldsymbol{y}_{adv}] = \mathbb{E}[\tilde{\boldsymbol{y}} \,|\, \boldsymbol{y} = \boldsymbol{y}_{adv}]. \quad (1)$$

Proof. See appendix. $\square$

Remark 1. Proposition 0.1 indicates that on multi-label data where labels are correlated, for classifier h, we can still evaluate its fairness performances in the same way as evaluating

---

[1]https://arxiv.org/abs/2302.09683

traditional single label classifiers by comparing the averaged predicted probability on different subgroups. Moreover, we can construct estimations with finite samples

$$\mathbb{E}[\tilde{\boldsymbol{y}}\,|\,a\!=\!k]\approx\frac{\sum_{i=1}^{N}\tilde{\boldsymbol{y}}^{(i)}\mathbf{1}(a^{(i)}\!=\!k)}{\sum_{i=1}^{N}\mathbf{1}(a^{(i)}\!=\!k)}\quad\mathbb{E}[\tilde{\boldsymbol{y}}]\approx\frac{1}{N}\sum_{i=1}^{N}\tilde{\boldsymbol{y}}^{(i)}$$

$$\mathbb{E}[\tilde{\boldsymbol{y}}\,|\,a\!=\!k,\boldsymbol{y}\!=\!\boldsymbol{y}_{adv}]\approx\frac{\sum_{i=1}^{N}\tilde{\boldsymbol{y}}^{(i)}\mathbf{1}(a^{(i)}\!=\!k)\mathbf{1}(\boldsymbol{y}\!=\!\boldsymbol{y}_{adv})}{\sum_{i=1}^{N}\mathbf{1}(a^{(i)}\!=\!k)\mathbf{1}(\boldsymbol{y}\!=\!\boldsymbol{y}_{adv})}\quad(2)$$

$$\mathbb{E}[\tilde{\boldsymbol{y}}\,|\,\boldsymbol{y}\!=\!\boldsymbol{y}_{adv}]\approx\frac{\sum_{i=1}^{N}\tilde{\boldsymbol{y}}^{(i)}\mathbf{1}(\boldsymbol{y}\!=\!\boldsymbol{y}_{adv})}{\sum_{i=1}^{N}\mathbf{1}(\boldsymbol{y}\!=\!\boldsymbol{y}_{adv})}\quad(3)$$

Estimation Challenge of EOp In multi-label scenarios, a long-tailed phenomenon, i.e., most labels only associate with few samples (Dekel and Shamir 2010), brings additional challenges for EOp estimation. Without sufficient samples, EOp is barely able to construct reliable estimates for fairness and correspondingly may not achieve fairness objective in an in-processing framework on such datasets.

Mathematically, the challenge of estimating EOp stems from terms $\sum_{i=1}^{N}\mathbf{1}(a^{(i)}\!=\!k)\mathbf{1}(\boldsymbol{y}^{(i)}\!=\!\boldsymbol{y}_{adv})$ and $\sum_{i=1}^{N}\mathbf{1}(\boldsymbol{y}^{(i)}\!=\!\boldsymbol{y}_{adv})$ in eqn (2) and (3). When these summations are close to 0, the two estimates are unstable or even undefined. More formally, the conditional expectation in eqn (1) is

$$\mathbb{E}[\tilde{\boldsymbol{y}}\,|\,a\!=\!k,\boldsymbol{y}\!=\!\boldsymbol{y}_{adv}]=\int\tilde{\boldsymbol{y}}p(\tilde{\boldsymbol{y}}\,|\,a\!=\!k,\boldsymbol{y}\!=\!\boldsymbol{y}_{adv})\,\mathrm{d}\tilde{\boldsymbol{y}}$$

$$=\frac{\int\tilde{\boldsymbol{y}}p(\tilde{\boldsymbol{y}},a\!=\!k,\boldsymbol{y}\!=\!\boldsymbol{y}_{adv})\,\mathrm{d}\tilde{\boldsymbol{y}}}{P(a\!=\!k,\boldsymbol{y}\!=\!\boldsymbol{y}_{adv})}.$$

Here $P(a\!=\!k,\boldsymbol{y}\!=\!\boldsymbol{y}_{adv})=\mathbb{E}[\mathbf{1}(a\!=\!k)\mathbf{1}(\boldsymbol{y}\!=\!\boldsymbol{y}_{adv})]$, and

$$\int\tilde{\boldsymbol{y}}p(\tilde{\boldsymbol{y}},a\!=\!k,\boldsymbol{y}\!=\!\boldsymbol{y}_{adv})\,\mathrm{d}\tilde{\boldsymbol{y}}$$

$$=\iiint\tilde{\boldsymbol{y}}\mathbf{1}(a\!=\!k)\mathbf{1}(\boldsymbol{y}\!=\!\boldsymbol{y}_{adv})p(\tilde{\boldsymbol{y}},a,\boldsymbol{y})\,\mathrm{d}a\,\mathrm{d}\boldsymbol{y}\,\mathrm{d}\tilde{\boldsymbol{y}}$$

$$=\mathbb{E}[\tilde{\boldsymbol{y}}\mathbf{1}(a\!=\!k)\mathbf{1}(\boldsymbol{y}\!=\!\boldsymbol{y}_{adv})].$$

This implies

$$\mathbb{E}[\tilde{\boldsymbol{y}}\,|\,a\!=\!k,\boldsymbol{y}\!=\!\boldsymbol{y}_{adv}]=\frac{\mathbb{E}[\tilde{\boldsymbol{y}}\mathbf{1}(a\!=\!k)\mathbf{1}(\boldsymbol{y}\!=\!\boldsymbol{y}_{adv})]}{\mathbb{E}[\mathbf{1}(a\!=\!k)\mathbf{1}(\boldsymbol{y}\!=\!\boldsymbol{y}_{adv})]}\quad(4)$$

$$\mathbb{E}[\tilde{\boldsymbol{y}}\,|\,\boldsymbol{y}\!=\!\boldsymbol{y}_{adv}]=\frac{\mathbb{E}[\tilde{\boldsymbol{y}}\mathbf{1}(\boldsymbol{y}\!=\!\boldsymbol{y}_{adv})]}{\mathbb{E}[\mathbf{1}(\boldsymbol{y}\!=\!\boldsymbol{y}_{adv})]}.\quad(5)$$

Henceforth, eqn (1) is equivalent to

$$\frac{\mathbb{E}[\tilde{\boldsymbol{y}}\mathbf{1}(\boldsymbol{y}\!=\!\boldsymbol{y}_{adv})]}{\mathbb{E}[\mathbf{1}(\boldsymbol{y}\!=\!\boldsymbol{y}_{adv})]}=\frac{\mathbb{E}[\tilde{\boldsymbol{y}}\mathbf{1}(a\!=\!k)\mathbf{1}(\boldsymbol{y}\!=\!\boldsymbol{y}_{adv})]}{\mathbb{E}[\mathbf{1}(a\!=\!k)\mathbf{1}(\boldsymbol{y}\!=\!\boldsymbol{y}_{adv})]},\quad(6)$$

for $\forall\,k\!\in\!\mathcal{A}$. If $\mathbf{1}(\boldsymbol{y}\!=\!\boldsymbol{y}_{adv})\!=\!1$ happens with low probability, i.e., few samples are from label group $\boldsymbol{y}_{adv}$, EOp is difficult and even impossible to estimate from eqn (2) and (3) directly.

Similarity s-Induced Fairness (s$_\gamma$-SimFair)

Motivated by the above analysis, we propose a new framework to help achieve DP or EOp, where hard $\mathbf{1}(\boldsymbol{y}\!=\!\boldsymbol{y}_{adv})\!\in\!\{0,1\}$ is relaxed to some similarity function $s(\boldsymbol{y},\boldsymbol{y}_{adv})\!\in\![0,1]$. Informally, we loosen the membership of the advantaged group requirement in EOp and use a soft conditioning.

For any random sample $(\boldsymbol{x},a,\boldsymbol{y})$, fairness of its prediction is always taken in consideration, but as the affinity of $\boldsymbol{y}$ to $\boldsymbol{y}_{adv}$ decreases, it will be down-weighted when estimating fairness violations with respect to $\boldsymbol{y}_{adv}$.

Definition 1 (s$_\gamma$-SimFair). Given a similarity function s: $\mathcal{Y}\times\mathcal{Y}\!\to\![0,1]$, a multi-label classifier h satisfies **Sim**ilarity s-induced **Fair**ness (s$_\gamma$-SimFair) if for $\forall\,k\!\in\!\mathcal{A}$,

$$\frac{\mathbb{E}[\tilde{\boldsymbol{y}}s(\boldsymbol{y},\boldsymbol{y}_{adv})]}{\mathbb{E}[s(\boldsymbol{y},\boldsymbol{y}_{adv})]}=\frac{\mathbb{E}[\tilde{\boldsymbol{y}}\mathbf{1}(a\!=\!k)s(\boldsymbol{y},\boldsymbol{y}_{adv})]}{\mathbb{E}[\mathbf{1}(a\!=\!k)s(\boldsymbol{y},\boldsymbol{y}_{adv})]}.\quad(7)$$

Same as DP and EOp, terms involved in eqn (7) can be estimated with

$$\frac{\mathbb{E}[\tilde{\boldsymbol{y}}s(\boldsymbol{y},\boldsymbol{y}_{adv})]}{\mathbb{E}[s(\boldsymbol{y},\boldsymbol{y}_{adv})]}\approx\frac{\sum_{i}\tilde{\boldsymbol{y}}^{(i)}s(\boldsymbol{y}^{(i)},\boldsymbol{y}_{adv})}{\sum_{i}s(\boldsymbol{y}^{(i)},\boldsymbol{y}_{adv})}\quad(8)$$

$$\frac{\mathbb{E}[\tilde{\boldsymbol{y}}\mathbf{1}(a\!=\!k)s(\boldsymbol{y},\boldsymbol{y}_{adv})]}{\mathbb{E}[\mathbf{1}(a\!=\!k)s(\boldsymbol{y},\boldsymbol{y}_{adv})]}\approx\frac{\sum_{i}\tilde{\boldsymbol{y}}^{(i)}\mathbf{1}(a^{(i)}\!=\!k)s(\boldsymbol{y}^{(i)},\boldsymbol{y}_{adv})}{\sum_{i}\mathbf{1}(a^{(i)}\!=\!k)s(\boldsymbol{y}^{(i)},\boldsymbol{y}_{adv})}.\quad(9)$$

In this paper, we adopt the Jaccard score to define similarity s. In essence, we use the cardinality ratio between the intersection and union of pair $(\boldsymbol{y},\boldsymbol{y}')$ to measure their similarity, then apply some monotonic transformation for scaling. Formally, for $\boldsymbol{y}\!\in\!\mathcal{Y}$ with $y_l\!=\!1$ represents the presence of the l-th target, we denote $\mathrm{cate}(\boldsymbol{y})\!=\!\{l:y_l\!=\!1,l\!=\!1,\ldots,L\}$, i.e., the collection of indices of present targets, and define

$$\mathrm{Jac}(\boldsymbol{y},\boldsymbol{y}_{adv})\!=\!\frac{|\mathrm{cate}(\boldsymbol{y})\cap\mathrm{cate}(\boldsymbol{y}_{adv})|}{|\mathrm{cate}(\boldsymbol{y})\cup\mathrm{cate}(\boldsymbol{y}_{adv})|}$$

$$s_\gamma(\boldsymbol{y},\boldsymbol{y}_{adv})\!=\!\exp\left(\gamma\left(\mathrm{Jac}(\boldsymbol{y},\boldsymbol{y}_{adv})\!-\!1\right)\right)$$

where γ is a scaling parameter. It is worth mentioning that the choice of s is not unique and can be task- or data-specific.

s$_\gamma$-SimFair Unifies DP and EOp

One key characteristic of s$_\gamma$-SimFair is that it can be seen as an unification of DP and EOp, as formalized by Proposition 0.2 and 0.3.

Proposition 0.2 (DP and EOp are special cases of s$_\gamma$-SimFair). Consider s$_\gamma$-SimFair defined in eqn (7), if similarity s is a constant function $s(\boldsymbol{y},\boldsymbol{y}')\!=\!c$ for some c, then s$_\gamma$-SimFair implies DP; if s is an indicator function $s(\boldsymbol{y},\boldsymbol{y}')\!=\!\mathbf{1}(\boldsymbol{y}\!=\!\boldsymbol{y}')$, then s$_\gamma$-SimFair implies EOp.

Proof. See appendix. □

Proposition 0.3 (s$_\gamma$-SimFair helps achieve DP and EOp). For any multi-label classifier h satisfying s$_\gamma$-SimFair, its violation of DP will be arbitrarily small if γ is sufficiently small; and its violation of EOp will be arbitrarily small if γ is sufficiently large. More generally, its violation of DP is arbitrarily close to its violation of s$_\gamma$-SimFair for sufficiently small γ, and its violation of EOp is arbitrarily close to its violation of s$_\gamma$-SimFair for sufficiently large γ.

Proof. See appendix. □

Remark 2. Proposition 0.2 reveals the connection between s$_\gamma$-SimFair and DP (EOp). Proposition 0.3 further shows that s$_\gamma$-SimFair condition indeed helps achieve DP and EOp, establishing a theoretical foundation of *borrowing information from similar labels*.

14341

s$_\gamma$-SimFair Regularized Model Training

Fairness Violation Violation of s$_\gamma$-SimFair denoted by $\ell_{s_\gamma(\boldsymbol{y},\boldsymbol{y}_{adv})}(h)$, is defined as

$$\sum_{k=1}^{K}\left\|\frac{\mathbb{E}[\tilde{\boldsymbol{y}}s_\gamma(\boldsymbol{y},\boldsymbol{y}_{adv})]}{\mathbb{E}[s_\gamma(\boldsymbol{y},\boldsymbol{y}_{adv})]}-\frac{\mathbb{E}[\tilde{\boldsymbol{y}}\mathbf{1}(a=k)s_\gamma(\boldsymbol{y},\boldsymbol{y}_{adv})]}{\mathbb{E}[\mathbf{1}(a=k)s_\gamma(\boldsymbol{y},\boldsymbol{y}_{adv})]}\right\| \quad (10)$$

where $\|\cdot\|$ is the L_2 norm. In words, we count how the fairness conditions in eqn (7) are violated in all demographic subgroup $a=k$. When $K=2$ (i.e., the sensitive feature is binary), it can also be written as

$$\left\|\frac{\mathbb{E}[\tilde{\boldsymbol{y}}\mathbf{1}(a=1)s_\gamma(\boldsymbol{y},\boldsymbol{y}_{adv})]}{\mathbb{E}[\mathbf{1}(a=1)s_\gamma(\boldsymbol{y},\boldsymbol{y}_{adv})]}-\frac{\mathbb{E}[\tilde{\boldsymbol{y}}\mathbf{1}(a=2)s_\gamma(\boldsymbol{y},\boldsymbol{y}_{adv})]}{\mathbb{E}[\mathbf{1}(a=2)s_\gamma(\boldsymbol{y},\boldsymbol{y}_{adv})]}\right\|. \quad (11)$$

DP and EOp, as discussed, are special cases of s$_\gamma$-SimFair so we omit their forms.

In-processing with s$_\gamma$-SimFair We use s$_\gamma$-SimFair to improve fairness of classifier h in an in-processing framework. Specifically, on each mini-batch during training, we estimate the fairness violation (defined in eqn (10) or (11)) by eqn (8) and (9). The estimate defines the regularization term as parts of training loss. In particular, we train h with stochastic gradient descent-based methods by minimizing

$$\min_{h}\ell_{\text{mlc}}(h)+\lambda\ell_{s_\gamma(\boldsymbol{y},\boldsymbol{y}_{adv})}(h). \quad (12)$$

Here $\ell_{\text{mlc}}(h)$ is the loss for multi-label classification, and $\ell_{s_\gamma(\boldsymbol{y},\boldsymbol{y}_{adv})}(h)$ estimates the violation of s$_\gamma$-SimFair. Hyperparameter $\lambda\geq 0$ balances the two losses.

Multivariate Probit Variational AutoEncoder (MPVAE)
We use MPVAE as a backbone model to illustrate and verify the performance of our work. MPVAE is a multi-label classification method without fairness constraint enforcement, and we adapt it with a fairness penalty to ensure s$_\gamma$-SimFair.

MPVAE is a variational autoencoder structured model that is capable to capture pairwise dependency in label $\boldsymbol{y}$. It learns two encoders to map $\boldsymbol{x}$ and $\boldsymbol{y}$ into a shared representation space and decode with the same decoder. A Multivariate Probit (MP) model is used to predict $\hat{\boldsymbol{y}}$, and model correlations between different labels y_d and y'_d. Fig. 1 illustrates the structure of MPVAE, where green color marks the additional fairness penalty. Algorithm 1 in appendix provides a concise summary of updating MPVAE with one step on a minibatch. Due to the page limitation, we refer readers to Bai, Kong, and Gomes (2020) for details about MPVAE.

Experiments
In this section, we evaluate s$_\gamma$-SimFair with the goal of providing insights from three aspects:

- How does s$_\gamma$-SimFair approximate DP and EOp?
- How does s$_\gamma$-SimFair help achieve DP and EOp?
- How does s$_\gamma$-SimFair affect fairness-accuracy tradeoff?

In the following, we will discuss experiment settings first and then present the details about the evaluation from these three aspects.

Datasets and Experiment Setup
Datasets Due to the lack of existing work in fairness-aware multi-label classification, we transform two tabular datasets
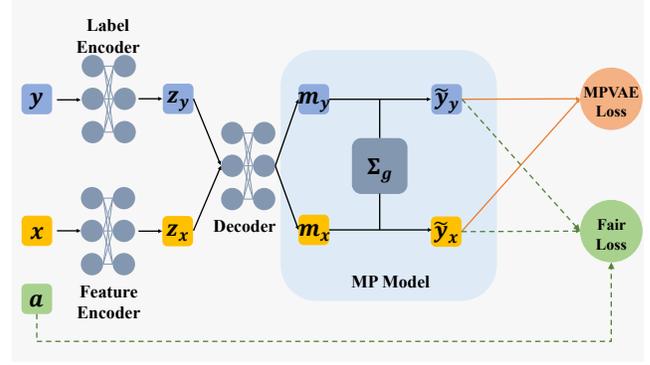

Figure 1: Framework of training MPVAE with fairness regularization (in green). Blue blocks mark the *label* branch and yellow blocks mark the (non-sensitive) *feature* branch. During training, MPVAE predicts two probability vectors $\tilde{\boldsymbol{y}}$ on two branches separately. Both of them are used to construct the s$_\gamma$-SimFair regularizer. During testing, only yellow blocks (prediction from the *feature* branch) are accessible.

that are ubiquitous in fairness literature into multi-label settings. Towards this goal, we select some features and treat them as additional targets. To help focus on the challenge brought by multi-label, we use binary sensitive features, but as defined in eqn (3), our methods can easily generalize to where more complicated sensitive features are used[2].

- *Adult* (Kohavi 1996) is a widely-used fairness dataset from UCI repository that contains 48,842 samples. Original Adult dataset contains 112 features and a binary label *income level*, which denotes whether an individual's yearly income is greater than $50K dollars or not. We further use *workclass* and *occupation* as two other targets. In terms of sensitive features, we follow Reddy et al. (2021) and binarize *age* into *25-44 years old* and *else*. This allows us to construct two balanced demographic subgroups.

- *Credit* (Yeh and Lien 2009) is another popular fairness dataset from UCI repository. It contains 30,000 samples, each sample is associated with 24 features and a binary label indicates the existence of *default payments*. We treat *education level* as an additional target, and use *gender* as the sensitive feature.

Baselines We compare MPVAE h trained with proposed s$_\gamma$-SimFair regularizer with three baseline methods: (1) No regularizer: use ℓ_{mlc} loss only by setting $\lambda=0$ in eqn (12); (2) DP regularizer: use DP violation as a regularizer, can be seen as an extension of Calders, Kamiran, and Pechenizkiy (2009); and (3) EOp regularizer: use EOp violation as a regularizer, which can be seen as an extension of Zafar et al. (2017). Regularizers are constructed according to eqn (11).

Evaluation Metrics To evaluate the fairness mitigation, we report the values of eqn (11) on test sets. As these are violations of fairness, smaller values indicate better performance. To evaluate the multi-label classification, we report three popular metrics in multi-label classifications (Wu and Zhou 2017; Bai, Kong, and Gomes 2020): micro-averaged F1 (micro-F1),

---

[2]See appendix for experiments where a multi-class sensitive feature *race* is considered.
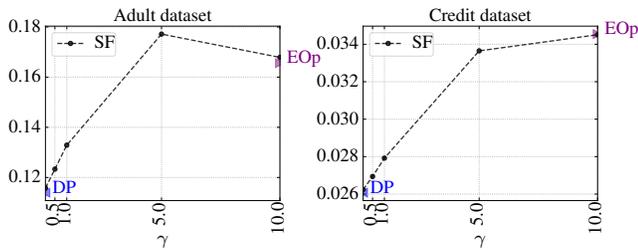
Figure 2: s_γ-SimFair (SF) can estimate DP and EOp (marked on the left and right y-axes) with different hyperparameter γ.

macro-averaged F1 (macro-F1), and example-averaged F1 (example-F1) as defined below

$$\text{micro-F1} = \frac{2\sum_{l=1}^{L}\sum_{i=1}^{N} y_l^{(i)} \hat{y}_l^{(i)}}{\sum_{l=1}^{L}\sum_{i=1}^{N} (y_l^{(i)} + \hat{y}_l^{(i)})}$$

$$\text{macro-F1} = \frac{1}{L}\sum_{l=1}^{L} \frac{2\sum_{i=1}^{N} y_l^{(i)} \hat{y}_l^{(i)}}{\sum_{i=1}^{N} (y_l^{(i)} + \hat{y}_l^{(i)})}$$

$$\text{example-F1} = \frac{1}{N}\sum_{i=1}^{N} \frac{2\sum_{l=1}^{L} y_l^{(i)} \hat{y}_l^{(i)}}{\sum_{l=1}^{L} (y_l^{(i)} + \hat{y}_l^{(i)})}.$$

These metrics compute either F1-score over the label matrix or averaged F1-score over targets or samples.

Implementation Details For s_γ-SimFair, we use $\gamma = 1, 5, 10$ to illustrate when it approximates DP or EOp. We also vary λ, the coefficient of fairness loss $\ell_{s_\gamma(\boldsymbol{y}, \boldsymbol{y}_{adv})}(h)$, from 1 to 5000 to study the trade-off between fairness and accuracy (in terms of micro-, macro-, and example-F1). We randomly choose 70% data for training and 30% for testing. Other hyperparameters for MPVAE training such as batch size, epochs, and learning rates are fixed throughout all experiments. A full list of hyperparameters used in this paper is provided in appendix.

Estimate DP and EOp with s_γ-SimFair

We first evaluate how well s_γ-SimFair can approximate DP and EOp to answer RQ1. To do so, we train a MPVAE without any regularizers for 20 epochs on Adult and Credit datasets and evaluate how it violates DP and EOp. We choose the largest label group (i.e., the label that appears most frequent) as the advantaged group. This allows us to construct a reliable estimate of EOp, which could be used as the ground truth.

Figure 2 shows how fairness violations estimated by s_γ-SimFair change under different γ, with DP and EOp marked on the left and right y-axis. From the figure, the starting points of s_γ-SimFair curves at $\gamma = 0.1$ locate close to DP, and the ending points at $\gamma = 10$ are close to EOp; these observations justify the effectiveness of s_γ-SimFair in approximating DP and EOp, consistent with theoretical analysis.

Next, we study the robustness of three estimators by varying the numbers of samples in the advantaged group to different levels and evaluating how estimates of DP and EOp change. Results summarized in Table 1 are averaged over 10 independent replications. Empirically, EOp estimator degrades drastically as the size of *observed* advantaged group decreases. s_γ-SimFair with large $\gamma (= 5)$, in contrary, produces more stable EOp estimates when EOp estimator fails.

	$\boldsymbol{y}_{adv}$	DP	$s_{0.1}$-SF	$s_{0.5}$-SF	s_1-SF	s_5-SF	s_{10}-SF	EOp
Adult	100%	0.11*	0.12	0.12	0.13	0.18	0.17	0.17*
	70%	0.11	0.11	0.12	0.13	0.18	0.18	0.17
	30%	0.10	0.10	0.11	0.12	0.17	0.18	0.17
	10%	0.10	0.10	0.10	0.11	0.14	0.16	0.17
	5%	0.10	0.10	0.10	0.11	0.15	0.23	0.27
Credit	100%	0.03*	0.03	0.03	0.03	0.03	0.03	0.03*
	70%	0.03	0.03	0.03	0.03	0.03	0.04	0.04
	30%	0.02	0.02	0.02	0.02	0.03	0.03	0.03
	10%	0.02	0.02	0.02	0.02	0.03	0.04	0.04
	5%	0.02	0.02	0.02	0.02	0.03	0.04	0.05

Table 1: DP, EOp, and s_γ-SimFair estimates (s_γ-SF) on Adult and Credit datasets. Certain portions of samples in the advantaged group are kept (column $\boldsymbol{y}_{adv}$) to check the robustness of different estimators. Results are averaged over 10 replications. Estimates of DP and EOp on 100% portion of samples are considered as the *ground truth* (marked with asterisk). s_γ-SimFair estimator is more robust than EOp estimator.

In terms of DP, both its own and s_γ-SimFair estimator produce similarly stable results, which is reasonable as we only decrease the size of the advantaged group.

Performance of Regularization

After showing that s_γ-SimFair can approximate DP and EOp well, we evaluate how well it can help achieve DP and EOp.

We start with reporting fairness violations of MPVAE trained with DP, EOp, and s_γ-SimFair regularizers. On each dataset, two potential advantaged groups are considered. The first group is the largest label group as in the last subsection, and the second group is chosen to be a *small* label group but we can still estimate EOp on the test set. For Adult dataset, since it has more labels, we choose the 18-th largest label group, which is the smallest one that has more than 100 test samples from the advantaged group out of 152 possible labels. For Credit dataset, we choose the 9-th largest, this group has at least 10 test samples from the advantaged label out of 13 possible labels. Throughout experiments, we fix $\lambda = 10$ and run 10 replications to smooth out randomness.

Table 2 shows resultant DP and EOp achieved by different methods. In all experiments, s_γ-SimFair performs competitive to DP regularizer and better than EOp regularizer in terms of minimizing these metrics as objectives. Notably, when the advantaged group is small, the vanilla EOp regularizer mitigates EOp violation poorly, but s_γ-SimFair still reduces it significantly. Moreover, s_γ-SimFair maintains a better DP-EOp balance, even they are known to be incompatible (Barocas, Hardt, and Narayanan 2017). For example, s_1-SF regularzier helps achieve better DP and EOp simultaneously than a DP regularizer on the largest label group on Adult dataset. We interpret this observation as a byproduct of the *biased* estimation given by s_γ-SimFair. As s_γ-SimFair is *biased* towards DP (EOp) when estimating EOp (DP), such bias implicitly considers the other metric and hence strikes a batter balance. These results clearly establish the power of s_γ-SimFair in minimizing DP and EOp.

To better reveal the limitation of EOp regularizer, we further evaluate how fair a model can be achieved by the use of different methods. To do so, we choose a large $\lambda = 5000$. Note that such large λ, will be shown shortly, significantly impedes accuracy. Here we sacrifice all accuracy to check

	$\|\boldsymbol{y}_{adv}\|$	Metric	Regularzier					
			DP	s_1-SF	s_5-SF	s_{10}-SF	EOp	None
Adult	No.1	DP	0.038	**0.031**	0.038	0.043	0.045	0.111
		EOp	0.051	0.042	**0.030**	0.034	0.035	0.161
	No.18	DP	**0.038**	**0.038**	0.043	0.045	0.094	0.111
		EOp	0.076	0.072	0.037	**0.027**	0.066	0.095
Credit	No.1	DP	0.018	0.018	**0.017**	0.018	0.018	0.029
		EOp	0.026	0.026	**0.025**	**0.025**	0.026	0.038
	No.9	DP	**0.018**	**0.018**	0.019	0.019	0.030	0.030
		EOp	0.202	**0.192**	0.193	0.197	0.241	0.241

Table 2: DP and EOp violations of MPVAE trained with DP, EOp, and s_γ-SimFair regularziers. On each dataset, a large and a small advantaged groups (measured by their ranking in col. $\|\boldsymbol{y}_{adv}\|$) are tested. Results are averaged over 10 replications, best results are in bold.
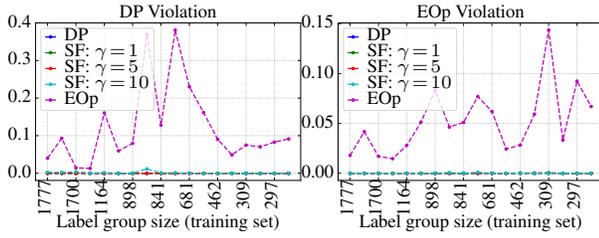
(a) Adult dataset
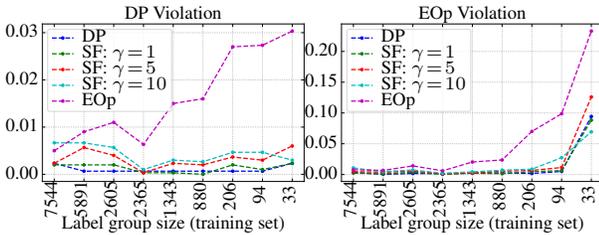
(b) Credit dataset

Figure 3: Achieved DP and EOp as the advantaged group becomes smaller. An extremely large $\lambda = 5000$ is used to enforce fairness mitigation. Compared to EOp regularizer, s_γ-SimFair is more robust to the sample size.

the potential of different methods.

We run 3 replications on top 18 largest label groups in Adult dataset and top 9 largest label groups in Credit dataset as advantaged group separately[3]. Figure 3 shows resultant DP and EOp achieved by different methods. Compared to EOp regularizer, which performs the worst on all labels, s_γ-SimFair is much more stable. In extreme cases, s_γ-SimFair, as a good approximation of EOp, also encounter failure ultimately, but it is much more robust.
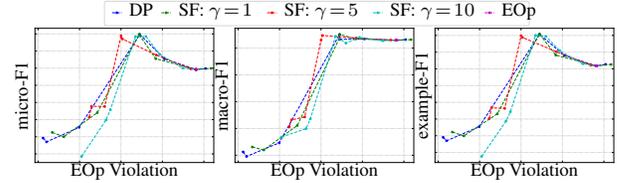
Fairness-Accuracy Tradeoff

We end up this section with a study on the tradeoff between fairness and accuracy on the two label groups from the previous section. Hyperparameter λ varies from 1 to 5000 and results are averaged over 10 replications. Due to the page limit,

---

[3]As described above, these groups have sufficient test samples to check violations.
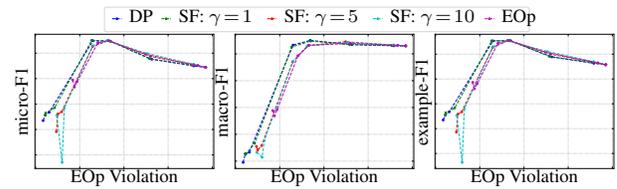
(a) Credit dataset: No.1 label group

(b) Credit dataset: No.9 label group

Figure 4: EOp-accuracy tradeoffs on Credit dataset. EOp regularizer is unstable and ineffective when the advantaged group is small, s_γ-SimFair, in contrast, preserves similar tradeoff trend as DP on both large and small label groups.

we only report EOp-accuracy tradeoffs on Credit dataset in Figure 4 here and defer other figures to appendix. Nevertheless, conclusions drawn here apply to all experiments.

Overall, micro- and example-F1 are much more robust to fairness requirement than macro-F1. On Credit dataset, they are even improved slightly when a small fairness regularization is added. We hypothesize that fairness regularization indirectly adds smooth conditions and penalizes unstable predictions. s_γ-SimFair has similar tradeoff patterns compared with the DP regularizer and does not encounter instability as EOp regularizer does. In addition, on small label groups where EOp regularizer fails, its s_γ-SimFair approximation succeeds in achieving low EOp violation, and performs one of the best in handling tradeoffs.

Conclusions

In this paper, We study the important problem of enforcing fairness on multi-label classification. Given the ubiquitous imbalanced issue with respect to label groups, we propose s_γ-SimFair, an effective framework that helps achieve existing group fairness metric: DP and EOp. We first establish a formal extension of DP and EOp condition to multi-label scenarios, then prove that (extended) DP and EOp can be exactly expressed by s_γ-SimFair, and can be approximated arbitrarily well. Experiments on two real-world datasets echos with theoretical analysis and reveals limitations of EOp regularizer. s_γ-SimFair, in contrary, shows strong robustness against the challenges EOp regularizer cannot overcome.

s_γ-SimFair is a general tool. The concept and technique derived in this paper can apply to multi-class classification as well, so long as a proper similarity function can be defined in the label space $\mathcal{Y}$. In the future, we plan to further conduct theoretical analysis on s_γ-SimFair regularizer and convergence, and apply s_γ-SimFair in a post-processing framework.

Acknowledgements

References

Bai, J.; Kong, S.; and Gomes, C. 2020. Disentangled Variational Autoencoder based Multi-Label Classification with Covariance-Aware Multivariate Probit Model. *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence*.

Barocas, S.; Hardt, M.; and Narayanan, A. 2017. Fairness in machine learning. *Nips tutorial*, 1: 2.

Barocas, S.; and Selbst, A. D. 2016. Big data's disparate impact. *Calif. L. Rev.*, 104: 671.

Boutell, M. R.; Luo, J.; Shen, X.; and Brown, C. M. 2004. Learning multi-label scene classification. *Pattern recognition*, 37(9): 1757–1771.

Buolamwini, J.; and Gebru, T. 2018. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on fairness, accountability and transparency*, 77–91. PMLR.

Calders, T.; Kamiran, F.; and Pechenizkiy, M. 2009. Building classifiers with independency constraints. In *2009 IEEE International Conference on Data Mining Workshops*, 13–18. IEEE.

Chen, D.; Xue, Y.; and Gomes, C. 2018. End-to-end learning for the deep multivariate probit model. In *International Conference on Machine Learning*, 932–941. PMLR.

Chib, S.; and Greenberg, E. 1998. Analysis of multivariate probit models. *Biometrika*, 85(2): 347–361.

Chouldechova, A.; and Roth, A. 2020. A snapshot of the frontiers of fairness in machine learning. *Communications of the ACM*, 63(5): 82–89.

Clare, A.; and King, R. D. 2001. Knowledge Discovery in Multi-label Phenotype Data. In *European conference on principles of data mining and knowledge discovery*, 42–53. Springer.

Creager, E.; Madras, D.; Jacobsen, J.-H.; Weis, M. A.; Swersky, K.; Pitassi, T.; and Zemel, R. 2019. Flexibly Fair Representation Learning by Disentanglement. arXiv:1906.02589.

Dekel, O.; and Shamir, O. 2010. Multiclass-Multilabel Classification with More Classes than Examples. In Teh, Y. W.; and Titterington, M., eds., *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, volume 9 of *Proceedings of Machine Learning Research*, 137–144. Chia Laguna Resort, Sardinia, Italy: PMLR.

Dressel, J.; and Farid, H. 2018. The accuracy, fairness, and limits of predicting recidivism. *Science Advances*, 4(1): eaao5580.

Dwork, C.; Hardt, M.; Pitassi, T.; Reingold, O.; and Zemel, R. 2011. Fairness Through Awareness. arXiv:1104.3913.

Edwards, H.; and Storkey, A. 2015. Censoring Representations with an Adversary. *arXiv preprint arXiv:1511.05897*.

El Kafrawy, P.; Mausad, A.; and Esmail, H. 2015. Experimental comparison of methods for multi-label classification in different application domains. *International Journal of Computer Applications*, 114(19): 1–9.

Fürnkranz, J.; Hüllermeier, E.; Loza Mencía, E.; and Brinker, K. 2008. Multilabel classification via calibrated label ranking. *Machine learning*, 73(2): 133–153.

Gong, T.; Liu, B.; Chu, Q.; and Yu, N. 2019. Using multi-label classification to improve object detection. *neurocomputing*, 370: 174–185.

Hardt, M.; Price, E.; and Srebro, N. 2016. Equality of Opportunity in Supervised Learning. arXiv:1610.02413.

Kamishima, T.; Akaho, S.; Asoh, H.; and Sakuma, J. 2012. Fairness-aware classifier with prejudice remover regularizer. In *Joint European conference on machine learning and knowledge discovery in databases*, 35–50. Springer.

Kingma, D. P.; and Welling, M. 2014. Auto-Encoding Variational Bayes. arXiv:1312.6114.

Kohavi, R. 1996. Scaling up the Accuracy of Naive-Bayes Classifiers: A Decision-Tree Hybrid. In *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*, KDD'96, 202–207. AAAI Press.

Liu, W.; Wang, H.; Shen, X.; and Tsang, I. 2021. The Emerging Trends of Multi-Label Learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1–1.

Locatello, F.; Abbati, G.; Rainforth, T.; Bauer, S.; Schölkopf, B.; and Bachem, O. 2019. On the Fairness of Disentangled Representations. arXiv:1905.13662.

Madras, D.; Creager, E.; Pitassi, T.; and Zemel, R. 2018. Learning Adversarially Fair and Transferable Representations. In *International Conference on Machine Learning*, 3384–3393. PMLR.

Mary, J.; Calauzenes, C.; and El Karoui, N. 2019. Fairness-aware learning for continuous attributes and treatments. In *International Conference on Machine Learning*, 4382–4391. PMLR.

Mohler, G.; Raje, R.; Carter, J.; Valasik, M.; and Brantingham, J. 2018. A penalized likelihood method for balancing accuracy and fairness in predictive policing. In *2018 IEEE international conference on systems, man, and cybernetics (SMC)*, 2454–2459. IEEE.

Nam, J.; Kim, J.; Loza Mencía, E.; Gurevych, I.; and Fürnkranz, J. 2014. Large-scale multi-label text classification—revisiting neural networks. In *Joint european conference on machine learning and knowledge discovery in databases*, 437–452. Springer.

Pedreshi, D.; Ruggieri, S.; and Turini, F. 2008. Discrimination-aware data mining. In *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, 560–568.

Petersen, F.; Mukherjee, D.; Sun, Y.; and Yurochkin, M. 2021. Post-processing for Individual Fairness. *Advances in Neural Information Processing Systems*, 34.

Pleiss, G.; Raghavan, M.; Wu, F.; Kleinberg, J.; and Weinberger, K. Q. 2017. On Fairness and Calibration. arXiv:1709.02012.

Read, J.; Pfahringer, B.; Holmes, G.; and Frank, E. 2011. Classifier chains for multi-label classification. *Machine learning*, 85(3): 333–359.

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 *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 1)*.

Scutari, M.; Panero, F.; and Proissl, M. 2021. Achieving Fairness with a Simple Ridge Penalty. arXiv:2105.13817.

Tsoumakas, G.; Katakis, I.; and Vlahavas, I. 2006. A review of multi-label classification methods. In *Proceedings of the 2nd ADBIS workshop on data mining and knowledge discovery (ADMKD 2006)*, 99–109. Citeseer.

Tsoumakas, G.; and Vlahavas, I. 2007. Random k-labelsets: An ensemble method for multilabel classification. In *European conference on machine learning*, 406–417. Springer.

Woodworth, B.; Gunasekar, S.; Ohannessian, M. I.; and Srebro, N. 2017. Learning non-discriminatory predictors. In *Conference on Learning Theory*, 1920–1953. PMLR.

Wu, X.-Z.; and Zhou, Z.-H. 2017. A unified view of multi-label performance measures. In *international conference on machine learning*, 3780–3788. PMLR.

Yang, B.; Sun, J.-T.; Wang, T.; and Chen, Z. 2009. Effective multi-label active learning for text classification. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, 917–926.

Yeh, I.-C.; and Lien, C.-h. 2009. The comparisons of data mining techniques for the predictive accuracy of probability of default of credit card clients. *Expert systems with applications*, 36(2): 2473–2480.

Zafar, M. B.; Valera, I.; Gomez Rodriguez, M.; and Gummadi, K. P. 2017. Fairness beyond disparate treatment & disparate impact: Learning classification without disparate mistreatment. In *Proceedings of the 26th international conference on world wide web*, 1171–1180.

Zhang, D.; Zhao, S.; Duan, Z.; Chen, J.; Zhang, Y.; and Tang, J. 2020. A multi-label classification method using a hierarchical and transparent representation for paper-reviewer recommendation. *ACM Transactions on Information Systems (TOIS)*, 38(1): 1–20.

Zhang, M.-L.; Li, Y.-K.; Liu, X.-Y.; and Geng, X. 2018. Binary relevance for multi-label learning: an overview. *Frontiers of Computer Science*, 12(2): 191–202.

Zhang, M.-L.; and Zhou, Z.-H. 2007. Ml-knn: A Lazy Learning Approach to Multi-Label Learning. *Pattern recognition*, 40(7): 2038–2048.

Zhang, M.-L.; and Zhou, Z.-H. 2014. A Review on Multi-Label Learning Algorithms. *IEEE Transactions on Knowledge and Data Engineering*, 26(8): 1819–1837.

Zhao, Z.; Guo, Y.; Shen, H.; and Ye, J. 2020. Adaptive object detection with dual multi-label prediction. In *European Conference on Computer Vision*, 54–69. Springer.

Zheng, Y.; Mobasher, B.; and Burke, R. 2014. Context recommendation using multi-label classification. In *2014 IEEE/WIC/ACM International Joint Conferences on Web Intelligence (WI) and Intelligent Agent Technologies (IAT)*, volume 2, 288–295. IEEE.