Fairness without Demographics through Shared Latent Space-Based Debiasing

Rashidul Islam, Huiyuan Chen, Yiwei Cai

Visa Research, USA
{raislam, hchen, yicai}@visa.com

Abstract

Ensuring fairness in machine learning (ML) is crucial, particularly in applications that impact diverse populations. The majority of existing works heavily rely on the availability of protected features like race and gender. However, practical challenges such as privacy concerns and regulatory restrictions often prohibit the use of this data, limiting the scope of traditional fairness research. To address this, we introduce a Shared Latent Space-based Debiasing (SLSD) method that transforms data from both the target domain, which lacks protected features, and a separate source domain, which contains these features, into correlated latent representations. This allows for joint training of a cross-domain protected group estimator on the representations. We then debias the downstream ML model with an adversarial learning technique that leverages the group estimator. We also present a relaxed variant of SLSD, the R-SLSD, that occasionally accesses a small subset of protected features from the target domain during its training phase. Our extensive experiments on benchmark datasets demonstrate that our methods consistently outperform existing state-of-the-art models in standard group fairness metrics.

Introduction

Recent years have witnessed a surge in evidence suggesting that, when trained on historical data without necessary precautions, ML systems can inadvertently exhibit discrimination across various demographic groups (House 2016; Barocas and Selbst 2016; O'neil 2016; Campolo et al. 2017; Noble 2018). Such bias can have a serious impact on diverse aspects of everyday life, ranging from movie recommendations (Islam et al. 2021) to more serious domains like credit scoring (Nedlund 2019), and criminal recidivism prediction (Angwin et al. 2016). Consequently, significant research has been directed towards developing and enforcing various mathematical constructs of bias and fairness in algorithms (Dwork et al. 2012; Zemel et al. 2013; Hardt, Price, and Srebro 2016; Kusner et al. 2017; Kearns et al. 2018; Foulds et al. 2020b). However, a common constraint in these works is their dependence on the model's access to protected attributes such as race and gender, during their training.

In practical contexts, factors such as privacy concerns, legal constraints, and regulatory measures often limit the acquisition or use of protected attributes. For instance, Title VII of the 1964 Civil Rights Act prevents employers from inquiring about an applicant's gender and race (Blankenship 1993). Similarly, the EU GDPR imposes constraints on collecting such data (Voigt and Von dem Bussche 2017). Yet, the imperative to achieve fairness is undiminished, especially to counteract harmful biases against specific protected groups. For example, the CFPB mandates creditors to implement fair lending practices but concurrently restricts them from collecting demographic details from applicants.[1] This paradox is well-acknowledged within the AI community, spanning both the public sector (Veale and Binns 2017) and industry (Holstein et al. 2019), and highlights the urgent need of ensuring fairness without demographics.

Present state-of-the-art solutions to this conundrum (Hashimoto et al. 2018; Lahoti et al. 2020; Chai and Wang 2022) mainly adopt the idea of Rawlsian max-min fairness (Rawls 2001) that maximizes the utility such as accuracy for the most disadvantaged group without demographic information. These methods effectively tackle representation bias adhering to the infra-marginality principle, which posits that a system is biased if and only if its behavior exhibits disparities greater than those in society or the underlying data (Simoiu, Corbett-Davies, and Goel 2017). However, our experiments reveal that they frequently fail to satisfy established parity-based group fairness standards, like demographic parity (Dwork et al. 2012; Zemel et al. 2013; Creager et al. 2019) or the legally recognized 80%-Rule, as specified in the Code of Federal Regulations (Equal Employment Opportunity Commission 1978). Besides addressing representation bias, parity-based fairness notions also capture other critical biases such as societal and intentional prejudices, and societal disadvantages, all of which can often skew the behavior of ML systems (Barocas and Selbst 2016). Furthermore, groups or regions found by these existing fair algorithms without demographics may not necessarily align with the intended protected attributes. For instance, a model may be optimized to maximize utility in terms of race when the desired protected attribute is gender.

Our Contributions. In this paper, we introduce a novel method that leverages a shared latent space to approximate the inherent protected groups for fair learning. Despite the

[1] CFBP Law and Regulations, 12 CFR § 1002.5 Rules concerning requests for information: https://tinyurl.com/mr2hw8km.

system does not directly access these protected groups, the unobserved groups are correlated with observed features x (e.g., zip codes often correlate with race (Hunt 2005)) and outcomes y (e.g. disparities in outcomes often align with specific groups (Lahoti, Gummadi, and Weikum 2019)). While correlates of protected groups often trigger concerns in fairness literature, we demonstrate how they can be beneficial for enhancing fairness metrics. Inspired by domain adaptation (Ben-David et al. 2006; Ganin and Lempitsky 2015), we hypothesize that correlation of protected groups learned in a "source domain with demographics" (e.g., publicly available HMDA loan approval data includes demographics[2]) can be effectively transferred to a "target domain without demographics" (e.g., a bank's internal data for credit card fraud detection where demographics have been excluded to preserve privacy). We further hypothesize that while both domains should have loose connection (i.e., both are financial domains in the running example), they might contain different individuals with varying observed features x and outcomes y (i.e., loan decisions in the source versus fraud detection in the target). Our aim, based on these hypotheses, is to develop fair learning algorithm for the target domain (i.e., a fair fraud detection system) via group estimates learned and adapted from the source domain.

Our method, named *Shared Latent Space-based Debiasing (SLSD)*, transforms data from both source and target domains into correlated latent representations to facilitate the training of a "cross-domain" protected group estimator using just the observed groups in the source data. An adversarial debiasing technique then improves fairness in the downstream task on the target data using the group estimator. However, due to significant domain shifts and out-of-distribution examples, the group estimator's performance might degrade considerably on the target data. We address this by adopting a consistency training approach that refines group estimates on the target data by regularizing the estimator to be invariant to small noise injected to input examples. Specifically, *we use "source data with demographics" to ensure fairness in "target data without demographics"* by first **pre-training** for estimating groups on shared representations between them, then **fine-tuning** for improving group estimates on target data, and finally **debiasing** the downstream model for target data with these estimates. While the SLSD model operates without accessing protected attributes in the target data, we also present a relaxed variant, *R-SLSD*, which considers a very small subset (e.g., 1% of the training set) of the target data that provides protected attributes. Our extensive experiments validate SLSD's and R-SLSD's efficacy, demonstrating their fairness improvements over existing models across multiple fairness benchmark datasets.

Background and Related Work

Fairness in ML: A number of studies have subsequently demonstrated the harmful and pervasive nature of societal biases in ML (Angwin et al. 2016; Bolukbasi et al. 2016; Buolamwini and Gebru 2018). Addressing these concerns, there's been a surge in research to define fairness, typically divided into three categories: 1) individual fairness (Dwork et al. 2012; Kusner et al. 2017) which aims to ensure similar outcomes for similar individuals, 2) group fairness (Zemel et al. 2013; Hardt, Price, and Srebro 2016; Zafar et al. 2017; Islam et al. 2023) which advocates outcome parity across protected groups, and 3) max-min fairness (Hashimoto et al. 2018; Lahoti et al. 2020) which attempts to improve minimum utility across groups. We primarily focus on group fairness due to practical challenges in individual similarity determination (Dwork et al. 2012) and max-min notion's gaps in addressing societal stereotypes (Zhao et al. 2022; Grari, Lamprier, and Detyniecki 2022). There are various techniques to improve fairness, from penalizing violations (Hardt, Price, and Srebro 2016; Islam et al. 2023) and imposing fairness constraints (Zafar et al. 2017; Agarwal et al. 2018) to fair data transformations (Zemel et al. 2013; Louizos et al. 2016) and adversarial debiasing (Zhang, Lemoine, and Mitchell 2018). However, these approaches require the availability of protected attributes, which are often missing in practical applications.

Fairness without demographics: Achieving fairness in the absence of demographic data is an emerging and complex challenge. A common strategy is to use proxy features (Gupta et al. 2018; Zhao et al. 2022; Grari, Lamprier, and Detyniecki 2022) or to operate under the assumption of slightly perturbed (Awasthi, Kleindessner, and Morgenstern 2020; Wang et al. 2020) protected features. However, such proxies, vulnerable to estimation bias (Chen et al. 2019), are not consistently found in data and can be difficult to identify without domain expertise (Grari, Lamprier, and Detyniecki 2022). In fact, Kallus, Mao, and Zhou 2022 demonstrated that it's generally impossible to spot disparities when relying solely on proxies. Alternative approaches utilize pseudo-group formations through clustering (Yan, Kao, and Ferrara 2020; Dai and Wang 2021), but the alignment of these artificially constructed groups with real protected groups highly varies with data distributions (Zhao et al. 2022).

As mentioned earlier, the works closest to ours are DRO (Hashimoto et al. 2018) and ARL (Lahoti, Gummadi, and Weikum 2019), which aim to achieve fair models without demographics and without proxy-based assumptions. DRO uses distributionally robust optimization to tackle worst-case groups, while ARL concentrates on identifiable training errors through adversarially re-weighted learning strategy. Similar re-weighting strategies are seen in fair learning for supervised (Nam et al. 2020; Liu et al. 2021) and self-supervised (Chai and Wang 2022) contexts. Although these max-min fairness-driven approaches have enhanced the underrepresented group's accuracy, they often fail in addressing societal biases from conventional group fairness perspectives (Zhao et al. 2022) and inadvertently amplify inherent biases, a phenomenon we observed in our experiments.

Notably, Coston et al. 2019 explored fairness in domain adaptation in contexts where demographic information was present only in either the source or target domain. However, unlike our approach, their method mandates identical downstream tasks and a consistent feature set across both domains, severely limiting its broader applicability.

[2]https://www.consumerfinance.gov/data-research/hmda/

Modeling Inspirations: Our proposed approach draws inspiration from a wide range of prior techniques. Canonical Correlation Analysis (CCA) is a classical multivariate analysis method that finds maximally correlated linear projections of two random variables (Hotelling 1936). Later works have enhanced CCA for multi-view learning using deep models (Andrew et al. 2013; Zhang et al. 2021). Consistency training methods significantly improve semi-supervised learning by utilizing large unlabeled data to ensure the model's stability against input noise (Tarvainen and Valpola 2017). Later Xie et al. 2020 have demonstrated how to effectively produce noise for consistency training by data augmentation technique to nearly match the performance of the fully supervised models with semi-supervised learning. Domain adaptation work requires a model to be robust and generalizable across different domains (Ganin and Lempitsky 2015).

The Proposed SLSD

In this section, we first formalize the problem. We then present our SLSD approach, structured into three integral stages: 1) the pre-training phase, focusing on group estimations in the shared latent space, 2) the fine-tuning phase, which refines these estimations using consistency training, and 3) the debiasing phase, where we rectify biases in a downstream model with the refined group estimates.

Problem Formulation

In this paper, we consider a binary classification framework with tabular data, although our method can be generalized to other settings. Suppose, we are given a target dataset $\mathcal{T} = \{(x_{\mathcal{T}}^{(i)}, y_{\mathcal{T}}^{(i)})\}_{i=1}^{n_{\mathcal{T}}}$ consisting of $n_{\mathcal{T}}$ individuals, where $x_{\mathcal{T}}^{(i)}$ is an $v_{\mathcal{T}}$ dimensional input vector of non-protected features, and $y_{\mathcal{T}}^{(i)}$ represents its binary class label. We assume that each individual in $\mathcal{T}$ belongs to an unobserved protected group $a_{\mathcal{T}}^{(i)}$, e.g., men or women. To be more precise, $a_{\mathcal{T}}$ remain inaccessible both during training and inference.

We also consider a source dataset consisting of $n_{\mathcal{S}}$ individuals $\mathcal{S} = \{(x_{\mathcal{S}}^{(i)}, y_{\mathcal{S}}^{(i)}, a_{\mathcal{S}}^{(i)})\}_{i=1}^{n_{\mathcal{S}}}$ where again $x_{\mathcal{S}}^{(i)}$ is an $v_{\mathcal{S}}$ dimensional vector of non-protected features and $y_{\mathcal{S}}^{(i)}$ represents its binary class label. In contrast to the target dataset, the source dataset explicitly provides the protected groups $a_{\mathcal{S}}^{(i)}$. It is crucial to note that the source and target datasets might differ significantly in terms of individuals and types of features, with potential disparities in sample sizes ($n\mathcal{S} \neq n_{\mathcal{T}}$), and feature dimensions ($v_{\mathcal{S}} \neq v_{\mathcal{T}}$).

Given this setup, our goal is to leverage explicit groups $a_{\mathcal{S}}$ in $\mathcal{S}$ to estimate group memberships $\hat{a}_{\mathcal{T}}^{(i)}$ for $\mathcal{T}$. This inference serves as a foundation for developing a fair model $M_\Theta(x_{\mathcal{T}})$, parameterized by Θ, for downstream tasks (e.g., fair lending decisions). Despite the absence of explicit $a_{\mathcal{T}}$, we seek to lead $M_\Theta(x_{\mathcal{T}})$ to be fair for a particular group, like gender, by selecting that group from $a_{\mathcal{S}}$.

The Pre-training Phase for SLSD

The purpose of the pre-training stage is to learn complex nonlinear transformations between $x_{\mathcal{S}}$ and $x_{\mathcal{T}}$ such that the resulting representations $z_{\mathcal{S}}$ and $z_{\mathcal{T}}$, respectively, are highly linearly correlated. Following Deep CCA (Andrew et al. 2013), we can model both transformations with a source encoder E_ϑ as $z_{\mathcal{S}} = E_\vartheta(x_{\mathcal{S}})$ and a target encoder E_φ as $z_{\mathcal{T}} = E_\varphi(x_{\mathcal{T}})$, where the corresponding parameters ϑ and φ are jointly learned to maximize the total correlation between $z_{\mathcal{S}}$ and $z_{\mathcal{T}}$. However, Deep CCA was originally designed to find linear relationships between two views of the same dataset, e.g., correlating images with their textual descriptions. Applying this approach directly to our distinct datasets $x_{\mathcal{S}}$ and $x_{\mathcal{T}}$, each with its own unique individuals and features might not be meaningful due to the lack of inherent linkage between them.

To address this, we present a straightforward data sampling technique that establishes an indirect relationship between $x_{\mathcal{S}}$ and $x_{\mathcal{T}}$. Typically, outcomes vary from favorable or positive outcomes, such as loan approvals, to unfavorable or negative outcomes, like loan rejections. Let's denote positive subsets $x_{\mathcal{S}}^+ \subset x_{\mathcal{S}}$ and $x_{\mathcal{T}}^+ \subset x_{\mathcal{T}}$ of $n_{\mathcal{S}}^+$ and $n_{\mathcal{T}}^+$ individuals, respectively, when $y_{\mathcal{S}} = y_{\mathcal{T}} = 1$. The negative counterparts $x_{\mathcal{S}}^-$ and $x_{\mathcal{T}}^-$ consist of the remaining $n_{\mathcal{S}}^-$ and $n_{\mathcal{T}}^-$ individuals. Our sampling ensures that positive instances from both datasets are concurrently transformed by their respective encoders as $z_{\mathcal{S}}^+ = E_\vartheta(x_{\mathcal{S}}^+)$ and $z_{\mathcal{T}}^+ = E_\varphi(x_{\mathcal{T}}^+)$, and similarly for the negative instances. Furthermore, we adjust the sampling rate for $x_{\mathcal{T}}^+$ and $x_{\mathcal{T}}^-$, by either oversampling or downsampling, to ensure $n_{\mathcal{S}}^+ = n_{\mathcal{T}}^+$ and $n_{\mathcal{S}}^- = n_{\mathcal{T}}^-$. This balancing act enables effective optimization of the CCA loss in terms of the covariance and variance as:

$$\mathcal{L}_{\text{CCA}}(z_{\mathcal{S}}, z_{\mathcal{T}}) = -\sum_{i=1}^{n_{\mathcal{S}}^+} \text{cov}(z_{\mathcal{S}}^{+(i)}, z_{\mathcal{T}}^{+(i)}) / \sqrt{\text{var}(z_{\mathcal{S}}^{+(i)})\text{var}(z_{\mathcal{T}}^{+(i)})} - \sum_{i=1}^{n_{\mathcal{S}}^-} \text{cov}(z_{\mathcal{S}}^{-(i)}, z_{\mathcal{T}}^{-(i)}) / \sqrt{\text{var}(z_{\mathcal{S}}^{-(i)})\text{var}(z_{\mathcal{T}}^{-(i)})} \tag{1}$$

Minimizing the $\mathcal{L}_{\text{CCA}}(z_{\mathcal{S}}, z_{\mathcal{T}})$ is equivalent to maximizing

$$\max_{\vartheta,\varphi} \quad \text{Tr}(z_{\mathcal{S}}^\intercal, z_{\mathcal{T}}) \quad \text{s.t.} \quad z_{\mathcal{S}}^\intercal z_{\mathcal{S}} = z_{\mathcal{T}}^\intercal z_{\mathcal{T}} = I\,, \tag{2}$$

where $z_{\mathcal{S}} = [z_{\mathcal{S}}^+, z_{\mathcal{S}}^-]$ and $z_{\mathcal{T}} = [z_{\mathcal{T}}^+, z_{\mathcal{T}}^-]$ are the corresponding concatenations. The representations $z_{\mathcal{S}}$ and $z_{\mathcal{T}}$ need to serve a dual purpose: they should be discriminative enough for group estimations, and simultaneously, invariant to discrepancies between source and target domains. To fulfill this, a "cross-domain" Protected Group Estimator (PGE) model G_Ψ is designed that takes encoded representations as input to estimate group memberships. **Both E_φ and G_Ψ are shared across all three training phases but are not used during the inference of downstream fair model.**

In the pre-training phase, G_Ψ aims to minimize a cross-entropy (CE) loss, using only the observed $a_{\mathcal{S}}$ as:

$$\mathcal{L}_{\text{CE}}(\hat{a}_{\mathcal{S}}, a_{\mathcal{S}}) = -\sum_{i=1}^{n_{\mathcal{S}}} \sum_{k=1}^{K} a_{\mathcal{S},k}^{(i)} \log(\hat{a}_{\mathcal{S},k}^{(i)})\,, \tag{3}$$

where K is the number of groups in the source domain and $\hat{a}_{\mathcal{S}} = \sigma(G_\Psi(z_{\mathcal{S}}))$ is the Softmax output of G_Ψ, with $z_{\mathcal{S}}$ as its input. Therefore, the final pre-training objective becomes:

$$\min_{\vartheta,\varphi,\Psi} \quad \mathcal{L}_{\text{CCA}}(z_{\mathcal{S}}, z_{\mathcal{T}}) + \mathcal{L}_{\text{CE}}(\hat{a}_{\mathcal{S}}, a_{\mathcal{S}})\,. \tag{4}$$

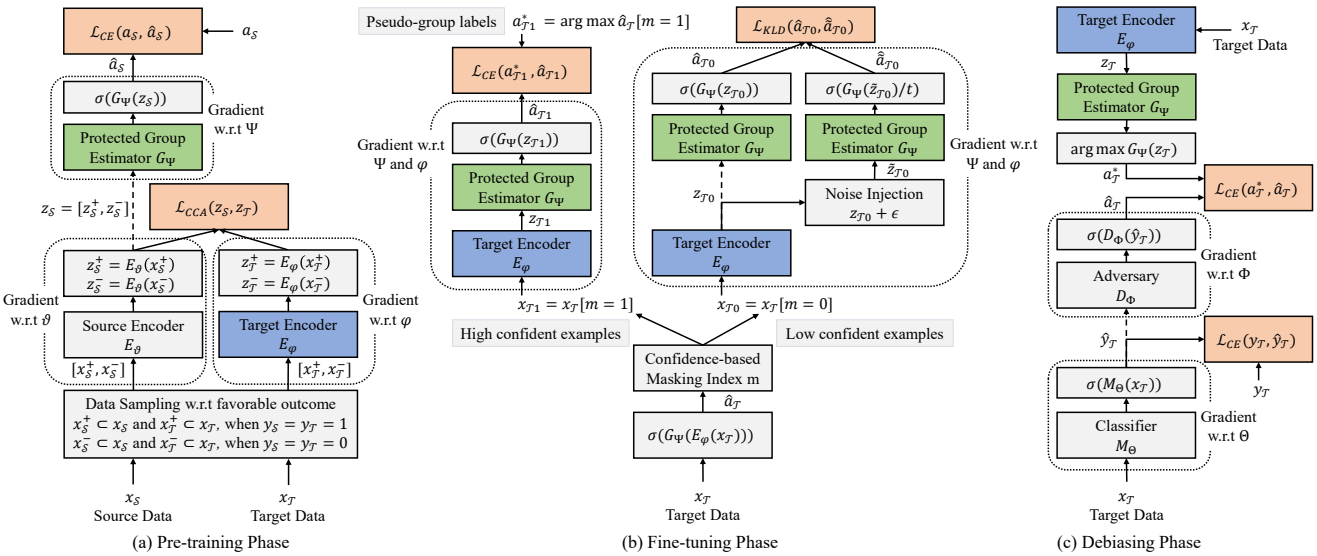

Figure 1: Computational graph for SLSD. Target encoder (*blue*) and group estimator (*green*) are shared across all phases.

The Fine-tuning Phase for SLSD

The fine-tuning is focused on enhancing both the target encoder E_φ and PGE G_Ψ for the target data, given the restrictions of unobserved $a_\mathcal{T}$. Our approach is inspired by the semi-supervised learning method (Xie et al. 2020), which uses data augmentation for consistency training. However, their method relies on a small set of labeled data to optimize the supervised CE loss, while concurrently optimizing the unsupervised consistency loss for the larger unlabeled set. Additionally, their noising operations for data augmentation, specifically designed for image and text data, are not suitable for our tabular data context. To tackle these challenges, we extend their method for our needs by optimizing supervised CE loss with entirely unsupervised data and incorporating an effective noise injection mechanism on E_φ's encoded representations for consistency training.

As G_Ψ is pre-trained solely on $z_\mathcal{S}$, we start by masking out those samples in the target data for which G_Ψ displays low confidence regarding the estimated group probabilities $\hat{a}_\mathcal{T} = \sigma(G_\Psi(z_\mathcal{T}))$, where $z_\mathcal{T} = E_\varphi(x_\mathcal{T})$. To be specific, we define a masking index m. For the top half, $n_\mathcal{T}/2$, of individuals with the highest probabilities across estimated group categories, we set $m = 1$ and extract the corresponding samples as $x_{\mathcal{T}1} = x_\mathcal{T}[m = 1]$. For the remaining samples, we set $m = 0$ and designate them as $x_{\mathcal{T}0} = x_\mathcal{T}[m = 0]$. For the supervised portion of training, group probabilities are then estimated as $\hat{a}_{\mathcal{T}1} = \sigma(G_\Psi(E_\varphi(x_{\mathcal{T}1})))$. Using these high-confidence samples, we generate pseudo-group labels as $a^*_{\mathcal{T}1} = \arg\max \hat{a}_{\mathcal{T}1}$ and plug them in Equation 3.

In the unsupervised consistency training, we explored various noise injections into $z_{\mathcal{T}0} = E_\varphi(x_{\mathcal{T}0})$, ranging from Gaussian to drop-out and Laplace noises. Based on our observations, small random perturbations, particularly jittering drawn from a Cauchy distribution with heavier tails, proved to be the most effective noise mechanism. Our noising operation can be formulated as $\tilde{z}_{\mathcal{T}0} = z_{\mathcal{T}0} + \epsilon$, where $\epsilon \sim \text{Cauchy}(\mu, \gamma)$. We set $\mu = 0$ and $\gamma = 200$ for all our experiments. The consistency loss is then computed as KL divergence between the estimated group probabilities as:

$$\mathcal{L}_{\text{KLD}}(\hat{\tilde{a}}_{\mathcal{T}0}, \hat{a}_{\mathcal{T}0}) = \hat{a}_{\mathcal{T}0} \cdot (\log \hat{a}_{\mathcal{T}0} - \log \hat{\tilde{a}}_{\mathcal{T}0}) \,, \tag{5}$$

where, $\hat{a}_{\mathcal{T}0} = \sigma(G_\Psi(z_{\mathcal{T}0}))$ and $\hat{\tilde{a}}_{\mathcal{T}0} = \sigma(G_\Psi(\tilde{z}_{\mathcal{T}0}/t))$, using a reduced Softmax temperature t. Given that prior studies emphasize the advantages of reducing prediction entropy in noisy scenarios (Grandvalet and Bengio 2004; Xie et al. 2020), we sharpen group predictions on augmented representations by setting t to 0.4. The final fine-tuning objective:

$$\min_{\varphi,\Psi} \quad \mathcal{L}_{\text{CE}}(\hat{a}_{\mathcal{T}1}, a^*_{\mathcal{T}1}) + \mathcal{L}_{\text{KLD}}(\hat{\tilde{a}}_{\mathcal{T}0}, \hat{a}_{\mathcal{T}0}) \,. \tag{6}$$

This fine-tuning procedure by minimizing both CE loss with pseudo group assignments and divergence with noise injection gradually propagates the high confident group assignments from $x_{\mathcal{T}1}$ to low confident $x_{\mathcal{T}0}$.

Relaxed Modeling Variant (R-SLSD)

Our relaxed modeling variant, R-SLSD, assumes that only a small fraction of the target data provides access to protected attributes. For simplicity, let's use our previous notation: $x_{\mathcal{T}1} \subset x_\mathcal{T}$ now represnts a small subset with observed $a_{\mathcal{T}1}$, while $x_{\mathcal{T}0} \subset x_\mathcal{T}$ represents the larger subset where $a_{\mathcal{T}0}$ remains unobserved. To utilize the $a_{\mathcal{T}1}$ while pre-training encoders and PGE models in R-SLSD, the Equation 4 can be extended by incorporating $\hat{a}_{\mathcal{T}1} = \sigma(G_\Psi(E_\varphi(x_{\mathcal{T}1})))$ as:

$$\min_{\vartheta,\varphi,\Psi} \quad \mathcal{L}_{\text{CCA}}(z_\mathcal{S}, z_\mathcal{T}) + \mathcal{L}_{\text{CE}}(\hat{a}_\mathcal{S}, a_\mathcal{S}) + \mathcal{L}_{\text{CE}}(\hat{a}_{\mathcal{T}1}, a_{\mathcal{T}1}) \,. \tag{7}$$

Under the R-SLSD framework, generating pseudo-group labels via confidence-based masking during the fine-tuning phase is no longer necessary. Therefore, the pseudo-groups $a^*_{\mathcal{T}1}$ in Equation 6 can be replaced with observed $a_{\mathcal{T}1}$ as:

$$\min_{\varphi,\Psi} \quad \mathcal{L}_{\text{CE}}(\hat{a}_{\mathcal{T}1}, a_{\mathcal{T}1}) + \mathcal{L}_{\text{KLD}}(\hat{\tilde{a}}_{\mathcal{T}0}, \hat{a}_{\mathcal{T}0}) \,, \tag{8}$$

where, the consistency training for $x_{\mathcal{T}0}$ with unobserved $a_{\mathcal{T}0}$ proceeds in the same manner as the SLSD approach.

The Debiasing Phase for SLSD and R-SLSD

Our debiasing approach for both SLSD and R-SLSD follows the same procedures. In an ideal scenario where our protected group estimations are perfect, e.g., if G_Ψ estimates the groups with absolute accuracy, we could readily apply any existing fairness algorithm to debias the downstream ML model, simply by replacing the true protected groups with our estimates. While achieving a perfect G_Ψ is infeasible, we observe that fair learning methods, which rely on explicit measurements of fairness metric to compute constraints (Agarwal et al. 2018) or penalties (Islam et al. 2023), struggle to effectively debias the downstream model when paired with our approach. This is presumably due to the high sensitivity of the fairness metric to the noisy group estimates, leading the model to converge in a bad solution.

To tackle this issue, we extend the adversarial debiasing method (Louppe, Kagan, and Cranmer 2017; Zhang, Lemoine, and Mitchell 2018) to make the downstream model's predictions independent of our G_Ψ's estimations, eliminating the need for explicit fairness metric measurement during training. Suppose M_Θ is the downstream classifier model which takes $x_\mathcal{T}$ as input and predicts the outcome $\hat{y}_\mathcal{T}$ for each individual, who belongs to the unknown protected group $a_\mathcal{T}$. Given our fine-tuned encoder E_φ and PGE G_Ψ, the group assignments can be estimated as:

$$a^*_\mathcal{T} = \arg\max G_\Psi(E_\varphi(x_\mathcal{T}))\,. \tag{9}$$

An adversarial network D_Φ is then designed that gets classifier's predictions $\hat{y}_\mathcal{T} = \sigma(M_\Theta(x_\mathcal{T}))$ as input and attempts to predict groups as $\hat{a}_\mathcal{T} = \sigma(D_\Phi(\hat{y}_\mathcal{T}))$. The learning objective to debias M_Θ becomes a min-max problem:

$$\min_\Theta \max_\Phi \quad \mathcal{L}_{\text{CE}}(\hat{y}_\mathcal{T}, y_\mathcal{T}) - \lambda \mathcal{L}_{\text{CE}}(\hat{a}_\mathcal{T}, a^*_\mathcal{T})\,, \tag{10}$$

where $\lambda > 0$ is a hyper-parameter that trades between classifier M_Θ's utility and fairness. Larger λ allows to achieve more fairness, but with greater loss in predictive performance, while smaller λ has the opposite impact. In this debiasing procedure, the adversary D_Φ penalizes the classifier M_Θ if the PGE G_Ψ's output is predictable from the M_Θ's output. Specifically, D_Φ aims to assure that predictions from M_Θ are independent of the estimated group assignments $a^*_\mathcal{T}$.

Practical Considerations

Figure 1 summarizes the computational graph of our proposed SLSD approach. In the experiments, we use standard feed-forward networks to implement both SLSD and R-SLSD. The architecture for the source encoder E_ϑ, target encoder E_φ, classifier M_Θ and adversary D_Φ are fully connected three layer feed-forward networks $256 - 128 - 64$, with ReLU activations. Although the PGE G_Ψ can be a deep network, a linear structure without hidden layers proved to be optimal. This is particularly true for the small academic benchmark datasets we experimented with, where the necessary features for group estimations were already extracted by the encoders. Notably, for adversarial debiasing, we observe that a warm start initialization procedure is required before optimizing the min-max problem in Equation 10. The training for the debiasing can be summarized in three steps:

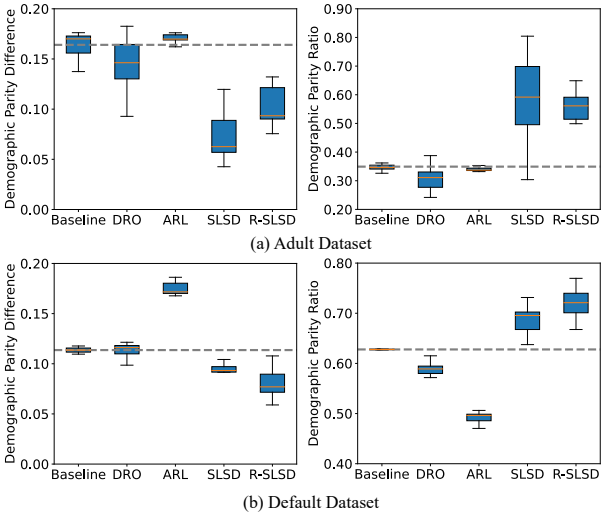

Figure 2: Fairness metrics across 10 runs. Lower is better for parity difference; higher for parity ratio.

1) pre-training M_Θ for the entire data, 2) pre-training D_Φ on the M_Θ's predictions, and 3) finally, alternately train M_Θ and D_Φ for each mini-batch by first training D_Φ while keeping M_Θ fixed and then training M_Θ while keeping D_Φ fixed.

Experimental Results

We conduct a comprehensive evaluation of our SLSD and R-SLSD on three benchmark datasets[3]: 1) Adult (Becker and Kohavi 1996): income prediction, 2) ACSIncome (Ding et al. 2021): another variant of income prediction and 3) Default (Yeh 2016): credit card default prediction [4]. For all datasets, we selected *gender* (men and women) as the protected attribute. Additionally, we conducted a case study on the COMPAS dataset, which has faced criticism for racial bias in criminal recidivism predictions (Angwin et al. 2016), focusing on the protected attribute *race* (white and black).

Evaluation Metrics

To assess predictive accuracy, we measure the area under the ROC curve (AUC) and balanced accuracy (Bal. Acc.), averaged over all (*overall*) instances, given their robustness against class imbalance. For fairness evaluation, we use well-recognized group fairness metrics: *demographic parity difference* (DPD) (Zemel et al. 2013) and *demographic parity ratio* (DPR) (Zafar et al. 2017) that quantify disparities in favorable outcomes between privileged (e.g. men) and unprivileged (e.g., women) groups. In line with DRO and ARL paper, we also report AUC (min) and Bal. Acc. (min) metrics, which denote the minimum AUC and Bal. Acc. values across all protected groups. These metrics serve as representations of Rawlsian max-min fairness. **A lower DPD is de-**

[3]Detailed dataset's descriptions are in the Supplementary.

[4]To replicate Ernst and Young's study (Dudík et al. 2020) on unfairness in credit models, following Microsoft Fairlearn toolkit (https://tinyurl.com/mv3d4npj), we introduced a synthetic feature, which is highly correlated with both gender and the outcome.

Dataset	Method	Bal. Acc. ↑ overall	Bal. Acc. ↑ min	AUC ↑ overall	AUC ↑ min	DPR ↑	DPD ↓
Adult	Baseline	0.773±0.009	0.764±0.007	0.910±0.001	0.888±0.009	0.349±0.016	0.164±0.012
	DRO	0.735±0.022	0.716±0.020	0.903±0.002	0.879±0.002	0.310±0.044	0.146±0.027
	ARL	**0.778**±0.004	**0.769**±0.003	**0.912**±0.000	**0.890**±0.000	0.340±0.007	0.170±0.004
	SLSD	0.700±0.028	0.695±0.023	0.828±0.035	0.822±0.028	**0.575**±0.158	**0.078**±0.033
	R-SLSD	0.763±0.012	0.750±0.012	0.900±0.003	0.885±0.001	0.559±0.047	0.103±0.019
Default	Baseline	**0.781**±0.007	0.737±0.005	**0.885**±0.001	**0.841**±0.001	0.628±0.001	0.114±0.004
	DRO	0.753±0.017	0.726±0.013	0.874±0.002	0.830±0.002	0.588±0.021	0.114±0.007
	ARL	0.778±0.005	**0.744**±0.003	0.880±0.001	0.839±0.002	0.492±0.011	0.175±0.006
	SLSD	0.752±0.092	0.714±0.079	0.835±0.157	0.789±0.146	0.714±0.094	0.084±0.029
	R-SLSD	0.775±0.010	0.737±0.006	0.882±0.003	**0.841**±0.002	**0.721**±0.028	**0.080**±0.013
ACSIncome	Baseline	0.798±0.003	0.794±0.003	0.894±0.000	0.888±0.000	0.698±0.007	0.123±0.004
	DRO	0.771±0.018	0.766±0.020	0.875±0.002	0.870±0.002	0.724±0.032	0.106±0.009
	ARL	**0.801**±0.003	**0.797**±0.003	**0.896**±0.000	**0.890**±0.000	0.700±0.007	0.124±0.003
	SLSD	0.723±0.048	0.715±0.051	0.817±0.056	0.804±0.061	0.727±0.036	0.096±0.020
	R-SLSD	0.797±0.003	0.792±0.003	0.892±0.001	0.889±0.000	**0.786**±0.013	**0.083**±0.006

Table 1: Performance for our SLSD and R-SLSD on the target datasets, compared with standard Baseline, DRO, and ARL.

sirable, while for other metrics, higher values are preferable. The protected features are used for fairness evaluation on the test subset of the target dataset.

Experimental Settings

Our methodology is designed for a transfer learning between source and target datasets. Specifically, **ACSIncome serves as the source when Adult is the target, and conversely, Adult becomes the default source for other target datasets**. While SLSD operates fully unsupervised in terms of protected features in the target, the R-SLSD randomly incorporates these features for 1% of training examples.

We use the same experimental setup, architecture, and hyper-parameter tuning for all the approaches reported in the experimental section. Each dataset is randomly split into 70% training and 30% test sets. Hyper-parameter tuning, including learning rate, mini-batch size, and the fairness tuning parameter λ (from Equation 10), is conducted on the training set. Best hyper-parameter values for all approaches are chosen via grid-search by performing 5-fold cross-validation optimizing for the best *overall* balanced accuracy. Note that we do not use protected features for tuning. Once the hyper-parameters are tuned, we use the independent test set for unbiased performance assessment. Refer to supplementary for further details. All experimental results are averaged across 10 independent runs, with different model parameter initialization.

Main Results - Fairness without Demographics

Our main comparisons are with *DRO* (Hashimoto et al. 2018), a group-agnostic distributionally robust optimization, and *ARL* (Lahoti et al. 2020), a group-agnostic adversarially reweighted learning technique. We also report results for the standard group-agnostic *Baseline* classifier, which emphasizes solely accurate predictions, without any fairness considerations. Table 1 reports average performance metrics with standard deviations across runs, with best results highlighted in bold. We make the following key observations:

- **Both SLSD and R-SLSD improve group fairness:** Our proposed models outperform other models in group fairness metrics across all datasets. Specifically, SLSD is the fairest model for the Adult in terms of both DPR and DPD, while R-SLSD leads in fairness improvement for the Default and ACSIncome datasets. When compared to the Baseline model on these datasets, *SLSD notably improves DPR by 64.8%, 13.7%, and 4.2% and DPD by 52.4%, 26.3%, and 22.0%*, while *R-SLSD improves DPR by 60.2%, 14.8%, and 12.6% and DPD by 37.2%, 29.8%, and 32.5%*, respectively.
- **DRO and ARL often amplify existing biases:** While the intent of any fair learning algorithm is to address biases present in the standard Baseline model, both DRO and ARL often underperform or can even intensify these biases. Figure 2 further shows DPD and DPR for all methods over 10 runs with varied model initializations. For both Adult and Default datasets, *DRO and ARL amplify the Baseline model's biases*. In contrast, our SLSD and R-SLSD models consistently mitigate these biases.
- **Cost of utility in SLSD:** Pursuing improved group fairness often results in a compromise on predictive accuracy, a well-established trade-off (Agarwal et al. 2018; Menon and Williamson 2018; Zhao and Gordon 2022). Given SLSD's dual challenges of improving fairness and bridging domain shifts between source and target, it unsurprisingly sacrifices both AUC and balanced accuracy. As DRO and ARL primarily aim to enhance utility metrics for under-performing groups, easily outperform SLSD in these measures. However, our R-SLSD model offers a promising balance, even overtaking DRO in utility.

Comparison with Fully Supervised Fair Model

To highlight our models' merits, we compare them with the original *adversarial debiasing model (ADM)* (Zhang, Lemoine, and Mitchell 2018), which demands access to protected features for all training instances. Figure 3 shows the

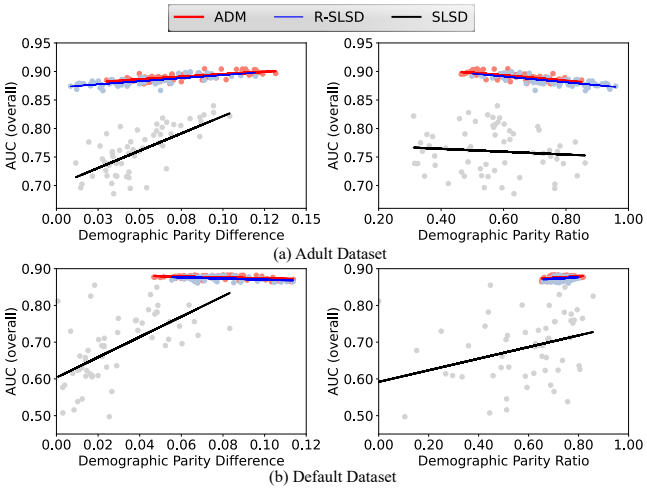

Figure 3: Comparison of SLSD and R-SLSD with ADM, which utilizes protected features for all training instances.

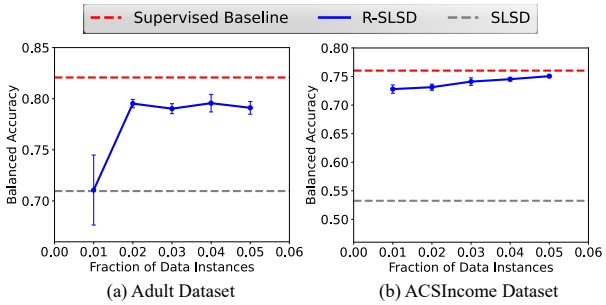

Figure 4: Protected group estimation comparing our SLSD and R-SLSD with a fully supervised model.

cross-validation grid search analysis for SLSD, R-SLSD, and ADM on Adult and Default datasets. SLSD outperforms in both DPD and DPR metrics without any protected target data, at the expense of AUC. Using just 1% of protected data, R-SLSD closely mirrors ADM's performance, which utilizes 100% of protected data, in both utility and fairness.

Performance for Protected Group Estimations

As our debiasing method depends on the group estimations, we further probe into the efficacy of these estimations in Figure 4. We train a fully supervised classifier to predict groups, establishing it as our benchmark. This Supervised Baseline consistently outperforms SLSD in balanced accuracy for group estimations. As expected, when the data fraction of group labels available to R-SLSD increases, its performance approximates the Supervised Baseline.

Case Study of An Extreme Scenario

We analyze an extreme scenario where the COMPAS criminal recidivism is the target dataset for all models, while SLSD and R-SLSD use the Adult income prediction as the source dataset. Given completely distinct domains (financial vs criminal justice), the fairness improvement of our models

Method	Bal. Acc. ↑ overall	AUC ↑ overall	DPR ↑	DPD ↓
Baseline	0.684	**0.745**	0.642	0.235
DRO	0.665	0.740	0.671	0.236
ARL	**0.685**	**0.745**	0.642	0.234
SLSD	0.676	0.736	0.660	**0.221**
R-SLSD	0.653	0.739	**0.718**	0.224

Table 2: Extreme scenario for SLSD and R-SLSD.

sharply decreases, compared to earlier experiments. This anticipated decline is presumably due to the significant domain shift, complicating the alignment between the source and target domains. However, SLSD and R-SLSD still surpass DRO and ARL in DPD. Regarding DPR, R-SLSD outperforms both DRO and ARL, though DRO edges out SLSD.

Discussion and Future Work

In future work, we plan to address potential privacy concerns related to our group estimates by adopting federated learning. This involves training the debiasing network on encrypted group estimates from SLSD in a secure environment. Once the downstream model is adjusted for fairness, the redundant debiasing components can be discarded during inference, eliminating residual privacy risks. We also aim to expand our methodology to a multi-dimensional protected groups setting, which will require more than a one-vs-all approach due to potential computational inefficiency and loss from data-sparsity issue of intersecting groups (Foulds et al. 2020a). To tackle this, we suggest learning multi-dimensional representations where each dimension corresponds to a protected group. Furthermore, SLSD can be extended to multi-class classification and regression tasks by directly utilizing our debiasing approach, where the adversary takes classifier's predicted probabilities or the regression model's continuous outcome. This, however, introduces complexity for data sampling in the pre-training phase, as information on individuals with both favourable and unfavourable outcomes is needed to map the disadvantaged groups in the latent space. We suggest binarizing the output space for these tasks. Our method also offers flexibility in replacing the debiasing phase with other techniques, such as fair representation learning (Louizos et al. 2016) , by providing group estimates instead of the true protected groups.

The journey towards algorithmic fairness is deeply embedded within broader social and historical discourses on equity and justice (Noble 2018; Keyes, Hutson, and Durbin 2019). Existing solutions for fairness without demographics mainly focus on addressing representation bias. However, fairness is not just a technical problem, it also encompasses societal, philosophical, and legal dimensions (Campolo et al. 2017). Our work introduces a promising direction of domain adaptation while acknowledging the complexities of demographic-agnostic fairness. SLSD, with its wide applicability in fairness-aware applications, especially in industries where demographic data collection is legally restricted, mitigates current privacy concerns in the ML fairness.

References

Agarwal, A.; Beygelzimer, A.; Dudík, M.; Langford, J.; and Wallach, H. 2018. A reductions approach to fair classification. In *International Conference on Machine Learning*, 60–69. PMLR.

Andrew, G.; Arora, R.; Bilmes, J.; and Livescu, K. 2013. Deep canonical correlation analysis. In *International Conference on Machine Learning*, 1247–1255. PMLR.

Angwin, J.; Larson, J.; Mattu, S.; and Kirchner, L. 2016. Machine bias: There's software used across the country to predict future criminals. and it's biased against blacks. *ProPublica, May*, 23.

Awasthi, P.; Kleindessner, M.; and Morgenstern, J. 2020. Equalized odds postprocessing under imperfect group information. In *International Conference on Artificial Intelligence and Statistics*, 1770–1780. PMLR.

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

Becker, B.; and Kohavi, R. 1996. Adult. UCI Machine Learning Repository.

Ben-David, S.; Blitzer, J.; Crammer, K.; and Pereira, F. 2006. Analysis of representations for domain adaptation. *Advances in Neural Information Processing Systems*, 19.

Blankenship, K. M. 1993. Bringing gender and race in: US employment discrimination policy. *Gender & Society*, 7(2): 204–226.

Bolukbasi, T.; Chang, K.-W.; Zou, J. Y.; Saligrama, V.; and Kalai, A. T. 2016. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. *Advances in Neural Information Processing Systems*, 29.

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.

Campolo, A.; Sanfilippo, M. R.; Whittaker, M.; and Crawford, K. 2017. AI Now 2017 Report. *AI Now Institute at New York University*.

Chai, J.; and Wang, X. 2022. Self-supervised fair representation learning without demographics. *Advances in Neural Information Processing Systems*, 35: 27100–27113.

Chen, J.; Kallus, N.; Mao, X.; Svacha, G.; and Udell, M. 2019. Fairness under unawareness: Assessing disparity when protected class is unobserved. In *Conference on Fairness, Accountability, and Transparency*, 339–348.

Coston, A.; Ramamurthy, K. N.; Wei, D.; Varshney, K. R.; Speakman, S.; Mustahsan, Z.; and Chakraborty, S. 2019. Fair transfer learning with missing protected attributes. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, 91–98.

Creager, E.; Madras, D.; Jacobsen, J.-H.; Weis, M.; Swersky, K.; Pitassi, T.; and Zemel, R. 2019. Flexibly fair representation learning by disentanglement. In *International Conference on Machine Learning*, 1436–1445. PMLR.

Dai, E.; and Wang, S. 2021. Say no to the discrimination: Learning fair graph neural networks with limited sensitive attribute information. In *Proceedings of the ACM Conference on Web Search and Data Mining*, 680–688.

Ding, F.; Hardt, M.; Miller, J.; and Schmidt, L. 2021. Retiring adult: New datasets for fair machine learning. *Advances in Neural Information Processing Systems*, 34: 6478–6490.

Dudík, M.; Chen, W.; Barocas, S.; Inchiosa, M.; Lewins, N.; Oprescu, M.; Qiao, J.; Sameki, M.; Schlener, M.; Tuo, J.; and Wallach, H. 2020. Assessing and mitigating unfairness in credit models with the fairlearn toolkit. White paper. URL: https://tinyurl.com/2x37jece.

Dwork, C.; Hardt, M.; Pitassi, T.; Reingold, O.; and Zemel, R. 2012. Fairness through awareness. In *Innovations in Theoretical Computer Science Conference*, 214–226. ACM.

Equal Employment Opportunity Commission. 1978. Guidelines on Employee Selection Procedures. *C.F.R.*, 29.1607.

Foulds, J. R.; Islam, R.; Keya, K. N.; and Pan, S. 2020a. Bayesian Modeling of Intersectional Fairness: The Variance of Bias. In *Proceedings of the SIAM International Conference on Data Mining*, 424–432. SIAM.

Foulds, J. R.; Islam, R.; Keya, K. N.; and Pan, S. 2020b. An intersectional definition of fairness. In *IEEE International Conference on Data Engineering*, 1918–1921. IEEE.

Ganin, Y.; and Lempitsky, V. 2015. Unsupervised domain adaptation by backpropagation. In *International Conference on Machine Learning*, 1180–1189. PMLR.

Grandvalet, Y.; and Bengio, Y. 2004. Semi-supervised learning by entropy minimization. *Advances in Neural Information Processing Systems*, 17.

Grari, V.; Lamprier, S.; and Detyniecki, M. 2022. Fairness without the Sensitive Attribute via Causal Variational Autoencoder. In *International Joint Conference on Artificial Intelligence*, 696–702.

Gupta, M. R.; Cotter, A.; Fard, M. M.; and Wang, S. 2018. Proxy Fairness. CoRR abs/1806.11212 (2018). *arXiv preprint arXiv:1806.11212*.

Hardt, M.; Price, E.; and Srebro, N. 2016. Equality of opportunity in supervised learning. *Advances in Neural Information Processing Systems*, 29.

Hashimoto, T.; Srivastava, M.; Namkoong, H.; and Liang, P. 2018. Fairness without demographics in repeated loss minimization. In *International Conference on Machine Learning*, 1929–1938. PMLR.

Holstein, K.; Wortman Vaughan, J.; Daumé III, H.; Dudik, M.; and Wallach, H. 2019. Improving fairness in machine learning systems: What do industry practitioners need? In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, 1–16.

Hotelling, H. 1936. Relations between two sets of variates. *Biometrika*, 28(3-4): 321–321.

House, W. 2016. Big Data: A Report on Algorithmic Systems, Opportunity, and Civil Rights. Executive Office of the President.

Hunt, D. B. 2005. Redlining. *Encyclopedia of Chicago*, 15.

Islam, R.; Keya, K. N.; Pan, S.; Sarwate, A. D.; and Foulds, J. R. 2023. Differential Fairness: An Intersectional Framework for Fair AI. *Entropy*, 25(4): 660.

Islam, R.; Keya, K. N.; Zeng, Z.; Pan, S.; and Foulds, J. 2021. Debiasing career recommendations with neural fair collaborative filtering. In *Proceedings of the Web Conference 2021*, 3779–3790.

Kallus, N.; Mao, X.; and Zhou, A. 2022. Assessing algorithmic fairness with unobserved protected class using data combination. *Management Science*, 68(3): 1959–1981.

Kearns, M.; Neel, S.; Roth, A.; and Wu, Z. S. 2018. Preventing fairness gerrymandering: Auditing and learning for subgroup fairness. In *International Conference on Machine Learning*, 2564–2572. PMLR.

Keyes, O.; Hutson, J.; and Durbin, M. 2019. A Mulching Proposal: Analysing and Improving an Algorithmic System for Turning the Elderly into High-Nutrient Slurry. In *CHI Conference on Human Factors in Computing Systems*. ACM.

Kusner, M. J.; Loftus, J.; Russell, C.; and Silva, R. 2017. Counterfactual fairness. In *Advances in Neural Information Processing Systems*.

Lahoti, P.; Beutel, A.; Chen, J.; Lee, K.; Prost, F.; Thain, N.; Wang, X.; and Chi, E. 2020. Fairness without demographics through adversarially reweighted learning. *Advances in Neural Information Processing Systems*, 33: 728–740.

Lahoti, P.; Gummadi, K. P.; and Weikum, G. 2019. ifair: Learning individually fair data representations for algorithmic decision making. In *IEEE International Conference on Data Engineering*, 1334–1345. IEEE.

Liu, E. Z.; Haghgoo, B.; Chen, A. S.; Raghunathan, A.; Koh, P. W.; Sagawa, S.; Liang, P.; and Finn, C. 2021. Just train twice: Improving group robustness without training group information. In *International Conference on Machine Learning*, 6781–6792. PMLR.

Louizos, C.; Swersky, K.; Li, Y.; Welling, M.; and Zemel, R. S. 2016. The Variational Fair Autoencoder. In *International Conference on Learning Representations*.

Louppe, G.; Kagan, M.; and Cranmer, K. 2017. Learning to pivot with adversarial networks. *Advances in Neural Information Processing Systems*, 30.

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

Nam, J.; Cha, H.; Ahn, S.; Lee, J.; and Shin, J. 2020. Learning from failure: De-biasing classifier from biased classifier. *Advances in Neural Information Processing Systems*, 33: 20673–20684.

Nedlund, E. 2019. Apple Card is accused of gender bias. Here's how that can happen.

Noble, S. U. 2018. Algorithms of oppression. In *Algorithms of Oppression*. New York University Press.

O'neil, C. 2016. *Weapons of math destruction: How big data increases inequality and threatens democracy*. Crown.

Rawls, J. 2001. *Justice as fairness: A restatement*. Harvard University Press.

Simoiu, C.; Corbett-Davies, S.; and Goel, S. 2017. The problem of infra-marginality in outcome tests for discrimination. *The Annals of Applied Statistics*, 11(3): 1193–1216.

Tarvainen, A.; and Valpola, H. 2017. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. *Advances in Neural Information Processing Systems*, 30.

Veale, M.; and Binns, R. 2017. Fairer machine learning in the real world: Mitigating discrimination without collecting sensitive data. *Big Data & Society*, 4(2).

Voigt, P.; and Von dem Bussche, A. 2017. The eu general data protection regulation (gdpr). *A Practical Guide, 1st Ed., Cham: Springer International Publishing*, 10(3152676): 10–5555.

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

Xie, Q.; Dai, Z.; Hovy, E.; Luong, T.; and Le, Q. 2020. Unsupervised data augmentation for consistency training. *Advances in Neural Information Processing Systems*, 33: 6256–6268.

Yan, S.; Kao, H.-t.; and Ferrara, E. 2020. Fair class balancing: Enhancing model fairness without observing sensitive attributes. In *Proceedings of the ACM Conference on Information & Knowledge Management*, 1715–1724.

Yeh, I.-C. 2016. Default of credit card clients. UCI Machine Learning Repository.

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 International Conference on World Wide Web*, 1171–1180.

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

Zhang, B. H.; Lemoine, B.; and Mitchell, M. 2018. Mitigating unwanted biases with adversarial learning. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, 335–340.

Zhang, H.; Wu, Q.; Yan, J.; Wipf, D.; and Yu, P. S. 2021. From canonical correlation analysis to self-supervised graph neural networks. *Advances in Neural Information Processing Systems*, 34: 76–89.

Zhao, H.; and Gordon, G. J. 2022. Inherent tradeoffs in learning fair representations. *The Journal of Machine Learning Research*, 23(1): 2527–2552.

Zhao, T.; Dai, E.; Shu, K.; and Wang, S. 2022. Towards fair classifiers without sensitive attributes: Exploring biases in related features. In *Proceedings of the ACM International Conference on Web Search and Data Mining*, 1433–1442.