DO WE REALLY ACHIEVE FAIRNESS WITH EXPLICIT SENSITIVE ATTRIBUTES?

Anonymous authors
Paper under double-blind review

ABSTRACT

Recently the wide usage of machine learning models for high-stake decision-making raises the concerns about the fairness and discrimination issue. Existing works found that *sensitive information*[1] of a sample could be leaked completely by sensitive attributes or partially by non-sensitive attributes, thus removing the sensitive attributes directly from the original features can not achieve fairness. The current fairness practice is to leverage the explicit sensitive attributes (i.e., as regularization) to debias the prediction, based on a strong assumption that non-sensitive attributes of all samples leak the sensitive information totally. However, we investigate the distribution of leaked *sensitive information* from non-sensitive attributes and make interesting findings that 1) the sensitive information distinctly varies across different samples. 2) the violation of demographic parity for samples prone to leak sensitive information (high-sensitive) are worse than that for low-sensitive samples, indicating the failure of current demographic parity measurements. To this end, we propose a new group fairness (α-Demographic Parity) to measure the demographic parity for samples with different levels of sensitive information leakage. Furthermore, we move one step forward and propose to achieve α-demographic parity by encouraging the independence of the distribution of the sensitive information in non-sensitive attributes and that of downstream task prediction, which is formulated as a cross-task knowledge distillation framework. Specifically, the sensitive teacher models the distribution of the sensitive information and the fair student models the distribution of the downstream task prediction. Then we encourage the independence between them by minimizing the Hilbert-Schmidt Independence Criterion. Our model can naturally tackle the limited sensitive attribution scenario since the teacher models can be trained with partial samples with sensitive attributes. Extensive experiments show the superior performance of our proposed method on the α-demographic parity and performs well on limited sensitive attribute scenarios.

1 INTRODUCTION

Deep neural networks (DNNs) have been increasingly applied to high-stake decision making such as credit scoring (Petrasic et al., 2017; Avery et al., 2012), criminal justice (Berk et al., 2021; Grgic-Hlaca et al., 2018), and healthcare (Rajkomar et al., 2018; Ahmad et al., 2020). Nevertheless, recent literature has exposed the prevalence of undesirable biases in deep neural networks. Despite the rising concerns, research on how to accurately evaluate the bias is still need more exploration.

Existing works found that *sensitive information* of an data sample could be leaked completely by sensitive attributes or partially by non-sensitive attributes [2], thus solely removing the sensitive attributes can not guarantee the achievement of fairness since the non-sensitive attributes ($\mathbf{x}$) can still leak the sensitive information partially Kamishima et al. (2012) The current fair machine learning models aim to remove the sensitive information hidden in non-sensitive attributes ($\mathbf{x}$) by leveraging the explicit sensitive attributes (s) to debias $\mathbf{x}$. Thus it implicitly *assumes* that non-sensitive attributes in all samples can leak the sensitive information totally and equally. However, the amount of sensitive information in non-sensitive attributes may be different, raising the following question:

[1] *sensitive information* is different form *sensitive attribute*, which means the amount of *sensitive attribute*.

[2] e.g., race information could be leaked completely by "race"attribute or partially by "zipcode" attribute

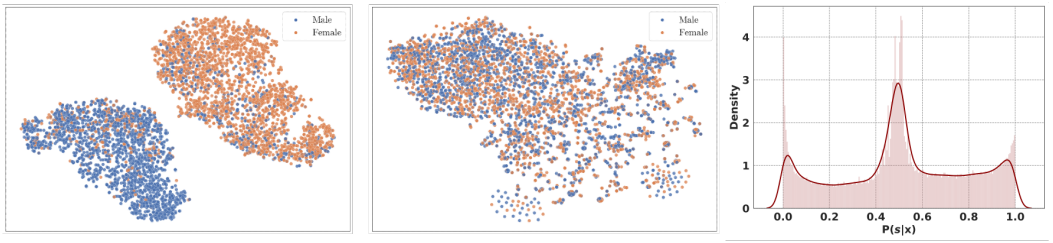

Figure 1: **Left:** t-SNE of high-sensitive samples. **Middle:** t-SNE of low-sensitive samples. **Right:** the distribution of the sensitive information in non-sensitive attributes. The apparent clustering of high-sensitive samples shows high-sensitive samples leak more sensitive information while high-sensitive samples even do not leak any sensitive information.

What happens if non-sensitive attributes leak different levels of sensitive information? And how should we consider it in the training and evaluation phase for fairness?

To answer this question, we conduct a preliminary experiment to investigate the fairness measurement on different samples with different levels of sensitive information. As the results shown in Figure 1, leakage of sensitive information of different samples are distinctly different. The left subfigure are the samples where more sensitive information is leaked by non-sensitive attributes. The middle subfigure shows the two-dimensional representation of the samples where less sensitive information leaked by the non-sensitive attributes. We also leverage a machine learning model to learn the $\mathrm{P}(s|\mathbf{x})$ and plot its distribution in the right subfigure in Figure 1, which show the $\mathrm{P}(s|\mathbf{x})$ are quite different over different samples. In summary, the results in Figure 1 show that the sensitive information in non-sensitive attributes could be various.

To this end, we first propose a new group fairness (α-Demographic Parity) to measure the demographic parity for different levels of leaked sensitive information. The proposed metric ensure each subgroup whose samples have the same level of sensitive information leakage to satisfy the demographic parity. Thus if all the subgroup satisfy the demographic parity, we achieve α-Demographic Parity across all α, indicating that we achieve demographic parity at a finer granularity.

To achieve α-demographic parity, we propose to directly encourage the independence of the distribution of the leaked sensitive information in non-sensitive attributes and prediction to gurantee the achievement of fairness. We formulate it as a cross-task knowledge distillation framework. Specifically, the sensitive teacher models the distribution of the sensitive information and the fair student models the distribution of the downstream task prediction. Then we encourage the independence between them by enforcing the Hilbert-Schmidt Independence Criterion to be 0. In addition, our model can naturally tackle the limited sensitive attribution scenario since the teacher models can be trained with partial samples with sensitive attributes. We highlight **main contributions** as follows:

- We investigate the distribution ($\mathrm{P}(s|\mathbf{x})$) of the sensitive information hidden in non-sensitive attributes and found the leakage of sensitive information from non-sensitive attributes are distinctly various over all samples. We also make an interesting finding that the violation of demographic parity for the high-sensitive-leak samples is worse than low-sensitive-leak samples, concluding that current fairness practice can not guarantee demographic parity for high-sensitive samples.
- We propose to leverage the distribution of the sensitive attributes to constrain the prediction for fairness via a cross-task [3] knowledge distillation framework, which includes a sensitive teacher and a fair student (STFS). Specifically, the sensitive teacher is designed to extract the sensitive information from non-sensitive attributes while the fair student makes fair predictions for downstream tasks. We guarantee the independence between distribution of the prediction and the sensitive information by enforcing the Hilbert-Schmidt Independence Criterion to be 0.
- We experimented on various datasets to validate the effectiveness of the proposed STFS. Since we can train the teacher model with partial samples, our proposed method is applicable to limited sensitive attributes scenarios. The experimental results show that our method can achieve comparable fairness performance with less than 20% training samples.

[3] By *cross-task*, we mean the teacher and student learn different tasks.

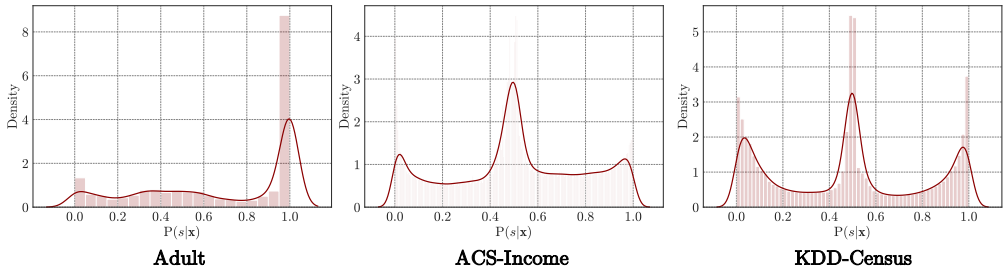

Figure 2: The distribution of $\mathrm{P}(s|\mathbf{x})$. The probability of $\mathrm{P}(s|\mathbf{x})$ is predicted by $\mathbf{x}$, which can be regarded as amount of *sensitive information*. The distribution shows that leakage of sensitive information in non-sensitive attributes are various over different samples. The sensitive attribute is sex.

2 MOTIVATION

In this section, we present the preliminary experiments to motivate our work. We first verify that the leakage of sensitive information for non-sensitive attributes are various over different samples and also found that the bias mainly stems from data samples with high sensitive information.

2.1 NOTATIONS AND DEMOGRAPHIC PARITY

For ease of exposition, we consider the binary classification and binary sensitive attribute.

Notations. The dataset is represented as $\{(\mathbf{x}_i, s_i, y_i)_{i=1}^N\}$, where $\mathbf{x} \in \mathbb{R}^d$ is the non-sensitive attributes, $s_i \in \{0, 1\}$ is sensitive attribute, and $y_i \in \{0, 1\}$ is the label of downstream task. We use $\hat{y}$ to denote the prediction probability of downstream task, which is obtained from the machine learning model $f(\mathbf{x}, \theta) : \mathbb{R}^d \to [0, 1]$ with trainable parameter θ.

Demographic Parity. and achieve demographic parity (DP). DP requires the predictions $\hat{y}$ to be independent of the sensitive attribute s, that is, $P(\hat{y}|s = 0) = P(\hat{y}|s = 1)$. The current practice to achieve algorithmic fairness is to leverage the explicit sensitive attributes (s) to debias the machine learning models with non-sensitive attributes ($\mathbf{x}$) as input. Given the difficulty of optimizing the independency constraints, Madras et al. (2018); Agarwal et al. (2018;?); Wei et al. (2019); Taskesen et al. (2020) propose the relaxed regulization $\Delta\mathrm{DP}(f) = |\mathbb{E}_{x\sim P_0} f(x) - \mathbb{E}_{x\sim P_1} f(x)|$ to penalize the cross entropy loss for downstream task, where $P_{0/1} = P(\cdot|s = 0/1)$.

2.2 PROBLEMS OF THE CURRENT PRACTICE FOR FAIRNESS

The current practice of fairness has a strong assumption that the non-sensitive attributes leak the sensitive information, however we argue that the sensitive information in the non-sensitive attributes could be various. To support our argument, we investigate the sensitive information in non-sensitive attributes. Specifically, we build a model to probe the sensitive information in the non-sensitive attributes, which take the non-sensitive attribute as input and the output of the model is $\mathrm{P}(s|\mathbf{x})$. The distribution of $\mathrm{P}(s|\mathbf{x})$ is presented in Figure 2. From the result, we observed: **Observation 1: leakage of sensitive information in non-sensitive attributes are various over different samples.**

We also conduct experiments to investigate the violation of demographic parity for data samples with different values of $\mathrm{P}(s|\mathbf{x})$ and the results are presented in Figure 3. We first split the data samples to high-sensitive samples and low-sensitive samples. We select the data samples which $\mathrm{P}(s|\mathbf{x}) < 0.25$ or > 0.75 and plot the distribution of the prediction probability of different demographic groups. From the results in Figure 3, in both the unfair and fair model, the violation of demographic parity for high-sensitive samples is more severe than low-sensitive attributes. **Observation 2: the overall bias mainly stems from data samples with high-sensitive information.**

From the results in Figure 3, in the fair model, the violation of demographic parity for high-sensitive samples is $todo$ while the the violation of demographic parity for high-sensitive samples for high-sensitive samples is $todo$. the violation of demographic parity for high-sensitive samples Thus we have the following observation, **Observation 3: the common fairness practice can not solve the current problem, the sample with high-sensitive information is still biased.**

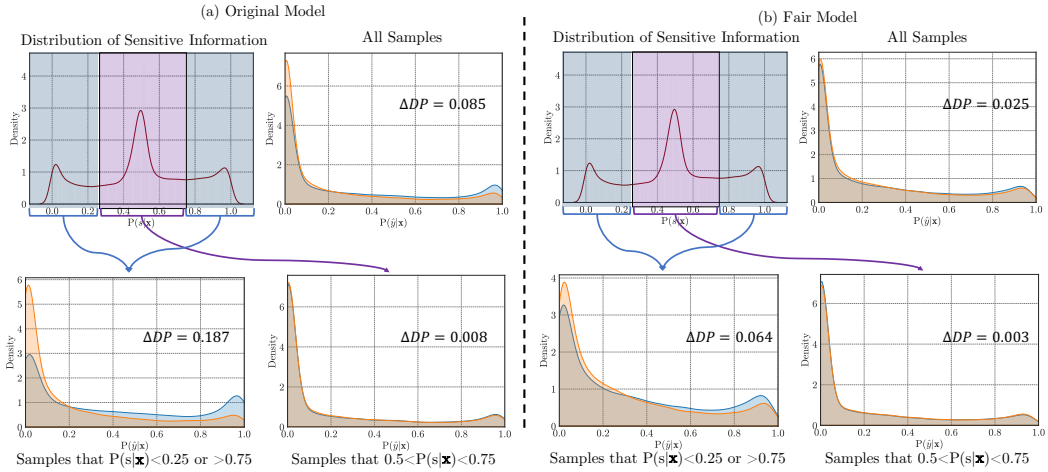

Figure 3: The ΔDP on subgroup with different levels of leaked sensitive attributes. The sensitive attribute is sex. The results show that ΔDP of high-sensitive samples is much larger than that of and low-sensitive samples. Especially, for low-sensitive samples, there even no fairness issue.

The reason is that violation of demographic parity does not take the level of the sensitive information leakage. To consider such information, we propose a new kind of fairness as well as the associated metric in the next section.

2.3 α-DEMOGRAPHIC PARITY

From the preliminary experiments, we conclude that *(i)* if there is no leakage of sensitive information, the prediction is not biased at all. *(ii)* current fairness methods do not achieve demographic parity for different levels of sensitive information leakage. However, the current demographic parity metric can not accurately measure the violation of the demographic parity when considering the levels of the sensitive information leakages. To further measure the fairness for different levels of sensitive information leakage, we first define α-Sensitive Information Leakage Group and then based on this group definition we define α-Demographic Parity.

Definition 1 **(α-Sensitive Information Leakage Group)** *An individual sample belong to α-sensitive information leakage group if* $P_{S|Z}(s|x) \in [0, \alpha] \cup [1-\alpha, 1]$.

Lower α means a higher sensitive information For example, if $\alpha = 0.1$, the samples in α-Sensitive Information Leakage Group is $\mathrm{P}_{S|Z}(s|x) \in [0, 0.1] \cup [0.9, 1]$ and they tend to leak more sensitive information.

Definition 2 **(α-Demographic Parity)** *A machine learning model satisfies α-Demographic Parity if $\forall\alpha \in [0, 0.5]$, α-Sensitive Information Leakage Group satisfy* $P_{\hat{Y},S|Z}(\hat{y}, s|x) = P_{\hat{Y}|X}(\hat{y}|x)P_{S|X}(s|x)$.

The basic idea behind this definition is that we split individuals into different groups based on different levels of sensitive information leakage, and then we guarantee the achievement of demographic party for each group. It is worthy to note that α-Demographic Parity is degraded to demographic parity when $\alpha = 0.5$, since all the samples will be considered the same if $\alpha = 0.5$. To measure the α-Demographic Parity, we propose α-ΔDP to evaluate the violation of α-Demographic Parity as follow:

$$\alpha\text{-}\Delta DP = \left| \frac{\sum_{i=1}^{N_0} P(\hat{y}_i|s_i = 0)}{N_0} - \frac{\sum_{i=1}^{N_1} P(\hat{y}_i|s_i = 1)}{N_1} \right| \quad \text{if} \quad \mathrm{P}_{S|Z}(s_i|x_i) \in [0, \alpha] \cup [1-\alpha, 1] \tag{1}$$

Where N_0/N_1 is the number of samples with the sensitive attribute $0/1$ while the smaple is in α-Sensitive Information Leakage Group.

In addition, we define the expectation of α-ΔDP ($\mathbb{E}_\alpha\alpha$-ΔDP) over $\alpha \in (0, 0.5]$ as a more strict version to measure the violation of α-Demographic Parity. And in the experiment, we use $\mathbb{E}_\alpha\alpha$-ΔDP as our fairness metric. Specifically, we use compute α-ΔDP for a series of αs and compute the average of them as the approximation of $\mathbb{E}_\alpha\alpha$-ΔDP.

3 METHODOLOGY

In this section, we introduce our proposed method STFS. In addition, we present the theoretical analysis for fairness guarantee, including the Hilbert-Schmidt Independence Criterion (HSIC), which is used to guarantee the independence between the distribution of the leaked sensitive information and downstream task prediction.

3.1 THE PROPOSED METHOD - STFS

Main Idea. As the results of preliminary experiments show that leaked sensitive information varies over different samples, these observations motivate us to seek more remedies. Thus we propose to directly encourage the distributions of leaked sensitive information and prediction to be independent.

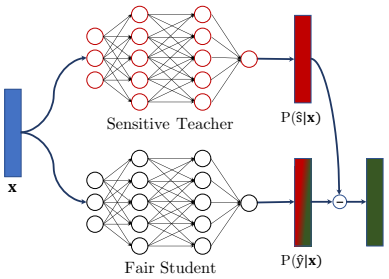

Figure 4: Overview of STFS.

The proposed STFS. We formalize this problem to a cross-task knowledge distillation, which includes a **S**ensitive **T**eacher model and a **F**air **S**tudent model. Since the distribution of the leaked sensitive information is unseen form the dataset explicitly, we use a teacher model to predict the sensitive information (*i.e.*, sensitive teacher). The sensitive teacher learns the distribution of leaked sensitive information from non-sensitive attributes and the fair student learns the distribution of the downstream task prediction. Then we encourage the independence between the outcome of sensitive teacher and fair student by enforcing the Hilbert-Schmidt Independence Criterion to be 0. Next, we elaborate our proposed method STFS as illustrated in Figure 4. Our goal is to achieve α-Demographic parity, which requires $\mathrm{P}_{\hat{Y},S|Z}(\hat{y}, s|x) = \mathrm{P}_{\hat{Y}|X}(\hat{y}|x)\mathrm{P}_{S|X}(s|x)$. To achieve this goal, the sensitive teacher model (red model in Figure 4) are utilized to model the distribution $\mathrm{P}_{S|X}(s|x)$ of sensitive information in the non-sensitive attributes, which is $\hat{s} = f_t(\mathbf{x}, \theta_t) = \mathrm{P}_{S|X}(s|x)$ where the trainable parameters is θ_s. The fair student are used to model the distribution of $\mathrm{P}_{\hat{Y}|X}(\hat{y}|x)$ of downstream task, which is $f_s(\mathbf{x}, \theta_s) = \mathrm{P}_{S|X}(s|x)$ where the trainable parameters is θ_s.

Training procedure We pre-train the sensitive teacher $\hat{s} = f_t(\mathbf{x}, \theta_t)$ with the Cross-Entropy loss function $\mathcal{L}_{CE}(f_t(\mathbf{x}, \theta_t), s)$, then we infer the leaked sensitive information $\hat{s}$ for all samples and use $\hat{s}$ to debias the fair student model. Next, we use the following objective function to optimize the fair student model:

$$\begin{aligned} Loss &= \mathcal{L}_{CE}(f_s(\mathbf{x}, \theta_s), y) + \lambda \cdot \mathcal{L}_{fair}(f_t(\mathbf{x}, \theta_t), f_s(\mathbf{x}, \theta_s)) \\ &= \mathcal{L}_{CE}(f_s(\mathbf{x}, \theta_s), y) + \lambda \cdot \mathrm{HSIC}(\hat{s}, f_s(\mathbf{x}, \theta_s)) \end{aligned} \tag{2}$$

where $f_s(\mathbf{x}, \theta_s) = \mathrm{P}_{S|X}(s|x)$ and $f_t(\mathbf{x}, \theta_t) = \mathrm{P}_{S|X}(s|x)$. The loss function $\mathcal{L}_{CE}$ will optimize the downstream task prediction and the HSIC$(\cdot)$ will encourage the distribution of sensitive information and prediction to be independent. The hyper-parameter λ is the balance parameters to balance between the performance and the fairness.

Independence Guarantee. Hereby we present the analysis of the independence guarantee of the distributions of leaked sensitive information and prediction. In our method, we minimize the HSIC to ensure the independence between the distributions of leaked sensitive information and prediction. Hilbert-Schmidt Independence Criterion (HSIC) is proposed to test if two random variables are independent only with the data samples from the random variables, and was introduced by Gretton et al. (2005b; 2008); Vepakomma et al. (2019). Consider two random variables X and Y, HSIC (Gretton et al., 2005b) is defined as Hilbert-Schmidt norm of the cross-covariance operator between

the distributions X and Y in Reproducing Kernel Hilbert Space (RKHS):

$$\begin{aligned}\text{HSIC}(\mathbb{P}_{XY}, \mathcal{H}, \mathcal{G}) &= \|C_{XY}\|^2 \\ &= \mathbb{E}_{XYX'Y'}[k_X(X, X')k_{Y'}(Y, Y')] \\ &+ \mathbb{E}_{XX'}[k_X(X, X')]\mathbb{E}_{Y'}[k_Y(Y, Y')] \\ &- 2\mathbb{E}_{XY}[\mathbb{E}_{X'}[k_X(X, X')]\mathbb{E}_{Y'}[k_Y(Y, Y')]],\end{aligned} \tag{3}$$

where k_X and k_Y are kernel functions, $\mathcal{H}$ and $\mathcal{G}$ are the Hilbert spaces, and $\mathbb{E}_{XY}$ is the expectation over X and Y. In practice, we can only observe the data samples while the exact distribution is unknown. Let $\mathcal{D} := \{(\mathbf{x}_1, \mathbf{y}_1), \cdots, (\mathbf{x}_m, \mathbf{y}_m)\}$ contain m i.i.d. samples drawn from $\mathbb{P}_{XY}$, where $\mathbf{x}_i \in \mathbb{R}^{d_x}$ and $\mathbf{y}_i \in \mathbb{R}^{d_y}$. Then equation 3 leads to the following empirical expression (Gretton et al., 2005a):

$$\text{HSIC}(\mathcal{D}, \mathcal{H}, \mathcal{G}) = (m-1)^{-2}\,\text{tr}(\mathbf{K}_X\mathbf{H}\mathbf{K}_Y\mathbf{H}) \tag{4}$$

where $\mathbf{K}_X \in \mathbb{R}^{m\times m}$ and $\mathbf{K}_Y \in \mathbb{R}^{m\times m}$ have entries $\mathbf{K}_{Xij} = k(\mathbf{x}_i, \mathbf{x}_j)$ and $\mathbf{K}_{Yij} = k(\mathbf{y}_i, \mathbf{y}_j)$, and $\mathbf{H} \in R^{m\times m}$ is the centering matrix $\mathbf{H} = \mathbf{I}_m - \frac{1}{m}\mathbf{1}_m\mathbf{1}_m^T$. With an appropriate kernel choice such as the Gaussian $k(\mathbf{x}, \mathbf{y}) \sim \exp(-\frac{1}{2}\|\mathbf{x} - \mathbf{y}\|^2/\sigma^2)$. HSIC is zero if and only if the random variables X and Y are independent, i.e., $P_{XY} = P_X P_Y$ (Sriperumbudur et al., 2010).

Theorem 1 (Independence Guarantee) *If STFS satisfies that HSCI($\hat{s}, \hat{y}$) = 0, then* $P_{\hat{Y},S|Z}(\hat{y}, s|x) = P_{\hat{Y}|X}(\hat{y}|x)P_{S|X}(s|x)$

Since we have HSIC is zero if and only if the random variables X and Y are independent, i.e., $P_{XY} = P_X P_Y$ (Sriperumbudur et al., 2010). The above theorem is easy to derive. This theorem suggests that if we minimize the HSIC regularization term to 0, the independence between the distributions of leaked sensitive information and prediction will be guaranteed.

3.2 Discussion

In this section, we provide discussions on the advantages of our proposed method STFS. we also discuss its potential limitations as well.

Achieve Fairness with Limited Sensitive Attributes In real-word scenarios, the sensitive attributes are typically very hard to collect, thus achieving fairness with limited sensitive attributes is urgently needed. Since our proposed method is formulated as a cross-task knowledge distillation framework, the sensitive teacher can be trained with partial training samples. This feature make our proposed method naturally applicable to limited sensitive attributes scenarios. We explore this more via an experiment with limited sensitive attributes in Section 4.4. The result shows that method enjoys the advantage that it can work well with a limited number of sensitive attributes.

Relation to Demographic Parity Our proposed α-Demographic Parity is closely related to the definition of demographic parity, thus we discuss the relation between our proposed method and demographic parity. Demographic Parity is a special case of α-Demographic Parity. The reason is straightforward that if we regard the whole sample as one group, the α-Demographic Parity will be degraded to regular Demographic Parity. Since our proposed α-Demographic Parity is a more strict fairness, we propose the following:

Proposition 1 *If a machine learning model satisfies the α-Demographic Parity, then it satisfies regular Demographic Parity.*

Proof Sketch. The binary sensitive attribute s can be decided by the distribution $\mathrm{P}_{S|X}(s|x)$, thus s is a function of $\hat{s}$, i.e., $s = f(\hat{s})$. If a machine learning model satisfies α-Demographic Parity, i.e., $\mathrm{P}_{\hat{Y},S|Z}(\hat{y}, s|x) = \mathrm{P}_{\hat{Y}|X}(\hat{y}|x)\mathrm{P}_{S|X}(s|x)$.

Limitations A possible limitation of our work is that model performance can be affected by the expressiveness of sensitive teacher models. Since the sensitive teacher is trained with the sensitive attributes as supervision, it could not be accurate to learn the distribution of sensitive information leaked from non-sensitive attributes. The results show that our proposed method performs well in the case of limited sensitive attributes (sensitive teacher models may be considered as undertrained).

4 EXPERIMENTS

We evaluate the performance of the proposed method in this section. First, we state the experimental setup, including the datasets, baselines and implementation details in Section 4.1. Then, we evaluate the accuracy-fairness trade off in Section 4.2. We also present the experimental results to demonstrate the applicability to limited sensitive attributes scenario. The major observations from the experimental results are highlighted with boldface.

4.1 EXPERIMENTAL SETUP

Datasets In the experiment, we consider the following datasets as our benchmark dataset,

- **UCI Adult** (Dua & Graff, 2017) contains clean information about $45,222$ individuals from the 1994 US Census. One instance is described with 15 attributes. The downstream task is to predict whether the income of a person is greater than $\$50$k, which is shown to bias to sex and race. We considered sex and race as sensitive attributes.
- **ACS-Income** (Ding et al., 2021) derives from the American Community Survey (ACS) Public Use Microdata Sample (PUMS). Like UCI Adult, the downstream task of this dataset is to predict whether an individual's income is above $\$50$k. The dataset contains $1,664,500$ data points. We choose sex and race as the sensitive attribute.
- **KDD Census** (Dua & Graff, 2017) contains $284,556$ clean instances with 41 attributes. The task on this dataset is also to predict whether the individual's income is above $\$50$k. The sensitive attributes are sex and race.

Baselines In our experiment, we use the following objective function $\mathcal{L}_{ce} + \lambda\mathcal{L}_{fair}$ to achieve fairness, where $\mathcal{L}_{ce}$ is the cross-entropy loss for downstream task and $\mathcal{L}_{fair}$ is demographic parity. We adopt three different regularizers as our baselines. The details of them are as follows:

- **DP-Gap** (Dua & Graff, 2017) is a kind of in-process method that adds the violation of demographic parity regularization term to the objective function Chuang & Mroueh (2020); Kamishima et al. (2012). This kind of method improves the fairness of the model with the regularization term simultaneously optimized during training. In our experiments, REG takes ΔDP_c as the regularization term.
- **Prejudice Remover** (Kamishima et al., 2012) This method is a prejudice remover regularizer, which enforces the independence between the prediction and sensitive attribute. Prejudice Remover leverages the mutual information to quantify the relation between the sensitive attribute and the prediction and minimize it.
- **HSIC** (Pérez-Suay et al., 2017; Quadrianto et al., 2019) used a HSIC are a regularization term to enforce the independence between model prediction and the sensitive attributes. Once HSIC equals zero, the mode prediction will be independent to sensitive attributes.

Implementation Details We run our experiments on the machine with NVIDIA RTX3090Ti GPU (24GB memory) and 256GB DDR4 memory to train the models. The code is implemented based on PyTorch (Paszke et al., 2019). The sensitive teacher and the fair student are both a two-layer MLP. The optimizer is Adam (Kingma & Ba, 2015) to train all the models.

Evaluation The evaluation of the performance of the downstream task performance is accuracy since we consider the binary classification in the experiments. We use the proposed fairness metric for α-Demographic Parity to evaluate the fairness.

4.2 WILL STFS ACHIEVE α-DEMOGRAPHIC PARITY?

In this section, we conducted experiments on various datasets to investigate the effectiveness of our proposed method to mitigate the bias and we present the results in Figure 5. We set the different λ in Equation (2) and plot the Pareto front for accuracy and fairness. From the results, we can see that our method obtain as better Pareto front than other baselines since Pareto front of STFS is at the outermost edge in most cases. Thus we have **Observation 4: our proposed STFS achieves the best trade off between prediction accuracy and α-Demographic Parity.**

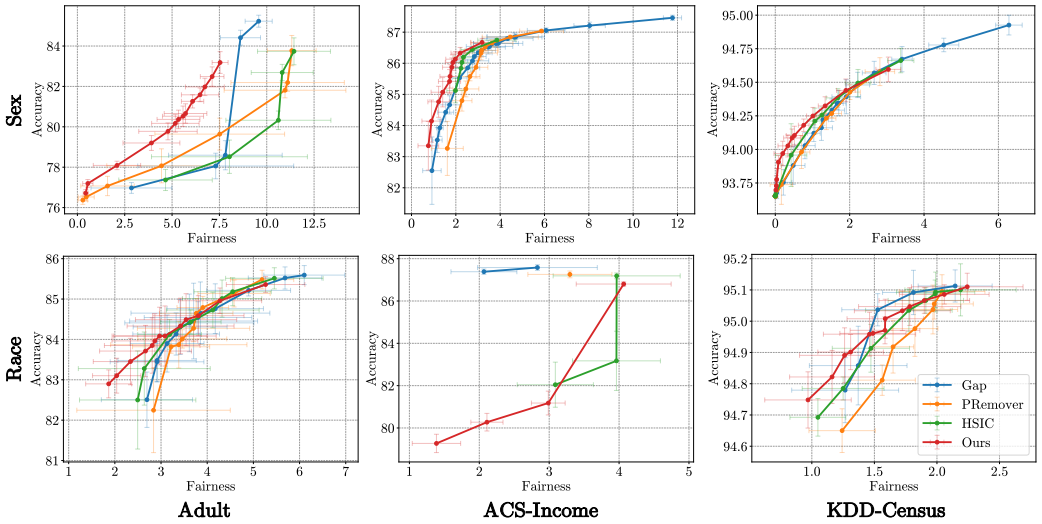

Figure 5: Pareto front for accuracy-fairness trade off. The fairness metric of the x-axis is the α-ΔDP. The results are based on 5 runs with different seeds.

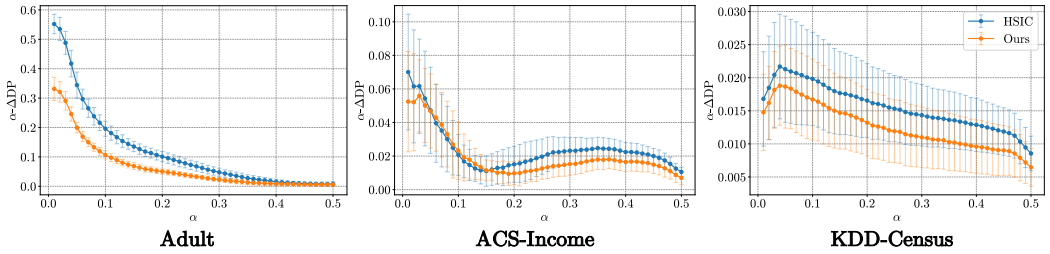

Figure 6: The performance of α-Demographic Parity with different αs. The sensitive attribute in the experiment is sex. A smaller α indicates a high level of leaked sensitive information. The results show that our proposed method generally achieves lower α-ΔDP over different αs.

4.3 Will STFS Perform with Different Alpha?

In this experiment, we conduct experiments to investigate the performance of our proposed model with different values of α and present the result in Figure 6. The baseline used in this experiment is using HSIC to enforce the independence between model prediction and sensitive attributes, which is the most similar method to ours. The results show that STFS generally obtains a lower δ-ΔDP than baseline HSIC. Thus we have **Observation 5: our proposed STFS achieves better α-Demographic Parity across various αs.**

4.4 How STFS Performs with Limited Sensitive Attribute?

In this section, we perform experiments to validate the applicability to limited sensitive attribute scenarios. Concretely, we leverage partial training samples (from 20% to 100%) to train the sensitive teacher and use the sensitive teacher to predict the leaked sensitive information for all training samples. Then we use predicted sensitive information to debias the fair student and reported the results in Figure 7.

From the results in Figure 7, we can see that on Adult and KDD-Census datasets, our method performs well with limited sensitive attributes. The Pareto fronts of model trained with partial training samples (from 20% to 100%) are nearly the same line, demonstrating that sensitive teachers in STFS can be trained by partial training samples. **Observation 6: Our proposed STFS can achieve fairness wiht limited sensitive attributes.**

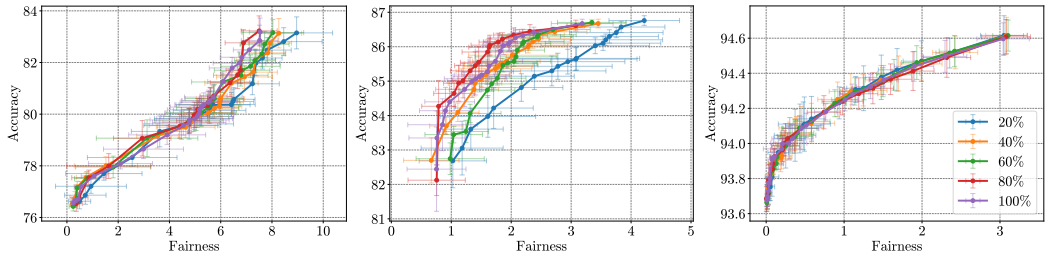

Figure 7: The Pareto front for accuracy-fairness trade off in limited sensitive attributes scenarios. Models are trained with partial training samples, from 20% to 100%. The sensitive attribute is sex.

5 Related Work

Fairness. Recently, lots of works from both academic and industrial have been proposed to address fairness issues. In this paper, we focus on in-processing fairness algorithms Agarwal et al. (2018); Elkan (2001); Jiang & Nachum (2020); Kamishima et al. (2012); Zafar et al. (2017); Zhang et al. (2018), which leverage the fairness constraint to enforce the fairness when training a model. Among the fairness constraint method, statistical independence between the model's outputs and groups (Kamishima et al., 2012; Zafar et al., 2017) are a major approach. Besides, adversarial learning technique is used to debias the model (Zhang et al., 2018), which make the sensitive attribute is unpredictable form the model by an adversary. In the computer vision domain, the discrimination problem has usually been tackled in facial analysis, such as face recognition Wang & Deng (2020); Wang et al. (2019). Wang *et al.* Wang et al. (2019) mitigated racial bias using the domain adaptation technique. Wang and Deng Wang & Deng (2020) utilized reinforcement learning. Their algorithms, however, have been specific only to the face recognition tasks.

Knowledge distillation. Knowledge distillation has been proposed to distill helpful information from teacher model to student model. Following the pioneer work (Hinton et al., 2015), where the teacher model distill the softmax output distribution to the student, various extensions have focused on how to exploit the learned features. The work of Romero *et al.* Romero et al. (2015) (FitNet) made the student mimic the features of the teacher through linear regression. Zagoruyko *et al.* Zagoruyko & Komodakis (2017) proposed attention transfer (AT) which transfers the knowledge using the attention map. Further, Yim *et al.* Yim et al. (2017) and Park *et al.* Park et al. (2019) studied approaches using the gram matrix and relation map respectively. Unlike the previous methods, In particular, Passalis *et al.* Passalis & Tefas (2018) suggested methods to reduce the distance between the teacher and the student feature distributions measured via *Kullback-Liebler* divergence.

HSIC (Hilbert-Schmidt Independence Criterion) is widely used as a independence measurement and is used for robustness learning (Greenfeld & Shalit, 2020). Recently HSIC are used to solve fairness issue. (Pérez-Suay et al., 2017; Quadrianto et al., 2019). For example, Wu et al. (2018) investigated the generalization properties of autoencoders using HSIC, while Lopez et al. (2018) uses HSIC to restrict the latent space search to constrain the aggregate variational posterior. Vepakomma et al. (2019) use distance correlation (an alternate formulation of HSIC) to remove unnecessary private information from medical training data.

6 Conclusion

In this paper, we focus on investigating the violation of demographic parity. We observes from preliminary experiments that the different samples have different sensitive information leakage and diverse levels of violation of demographic parity. Based on this interesting observation, we propose α-Demographic Parity to measure the violation for specific sensitive information leakage group. Additionally, we formulate a cross-task knowledge distillation framework to achieve α-Demographic Parity via chasing the independence of the distribution of the sensitive information in non-sensitive attributes and that of downstream task prediction. Naturally, the proposed framework can also tackle the situation with limited sensitive attribute.

REFERENCES

Alekh Agarwal, Alina Beygelzimer, Miroslav Dudik, John Langford, and Hanna Wallach. A reductions approach to fair classification. In *International Conference on Machine Learning*, pp. 60–69, 2018.

Muhammad Aurangzeb Ahmad, Arpit Patel, Carly Eckert, Vikas Kumar, and Ankur Teredesai. Fairness in machine learning for healthcare. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pp. 3529–3530, 2020.

Robert B Avery, Kenneth P Brevoort, and Glenn Canner. Does credit scoring produce a disparate impact? *Real Estate Economics*, 40:S65–S114, 2012.

Richard Berk, Hoda Heidari, Shahin Jabbari, Michael Kearns, and Aaron Roth. Fairness in criminal justice risk assessments: The state of the art. *Sociological Methods & Research*, 50(1):3–44, 2021.

Ching-Yao Chuang and Youssef Mroueh. Fair mixup: Fairness via interpolation. In *International Conference on Learning Representations*, 2020.

Frances Ding, Moritz Hardt, John Miller, and Ludwig Schmidt. Retiring adult: New datasets for fair machine learning. *NeurIPS*, 2021.

Dheeru Dua and Casey Graff. UCI machine learning repository, 2017. URL http://archive.ics.uci.edu/ml.

Charles Elkan. The foundations of cost-sensitive learning. In *International Joint Conference on Artificial Intelligence (IJCAI)*, 2001.

Daniel Greenfeld and Uri Shalit. Robust learning with the hilbert-schmidt independence criterion. In *International Conference on Machine Learning*, pp. 3759–3768. PMLR, 2020.

Arthur Gretton, Olivier Bousquet, Alex Smola, and Bernhard Schölkopf. Measuring statistical dependence with Hilbert-Schmidt norms. In *Proc. Int. Conf. Algorithmic Learning Theory*, pp. 63–77. Springer-Verlag, 2005a. ISBN 3-540-29242-X, 978-3-540-29242-5. doi: 10.1007/11564089_7.

Arthur Gretton, Olivier Bousquet, Alex Smola, and Bernhard Schölkopf. Measuring statistical dependence with hilbert-schmidt norms. In *International conference on algorithmic learning theory*, pp. 63–77. Springer, 2005b.

Arthur Gretton, Kenji Fukumizu, Choon H Teo, Le Song, Bernhard Schölkopf, and Alex J Smola. A kernel statistical test of independence. In *Advances in neural information processing systems*, pp. 585–592, 2008.

Nina Grgic-Hlaca, Elissa M Redmiles, Krishna P Gummadi, and Adrian Weller. Human perceptions of fairness in algorithmic decision making: A case study of criminal risk prediction. In *Proceedings of the 2018 World Wide Web Conference*, pp. 903–912, 2018.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

Heinrich Jiang and Ofir Nachum. Identifying and correcting label bias in machine learning. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2020.

Toshihiro Kamishima, Shotaro Akaho, Hideki Asoh, and Jun Sakuma. Fairness-aware classifier with prejudice remover regularizer. In *Joint European conference on machine learning and knowledge discovery in databases*, pp. 35–50. Springer, 2012.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *ICLR*, 2015.

Romain Lopez, Jeffrey Regier, Nir Yosef, and Michael I. Jordan. Information constraints on auto-encoding variational bayes. *CoRR*, abs/1805.08672, 2018. URL http://arxiv.org/abs/1805.08672.

David Madras, Elliot Creager, Toniann Pitassi, and Richard Zemel. Learning adversarially fair and transferable representations. *International Conference on Machine Learning*, 2018.

Wonpyo Park, Dongju Kim, Yan Lu, and Minsu Cho. Relational knowledge distillation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

Nikolaos Passalis and Anastasios Tefas. Learning deep representations with probabilistic knowledge transfer. In *European Conference on Computer Vision (ECCV)*, 2018.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. In *Advances in neural information processing systems*, pp. 8026–8037, 2019.

Adrián Pérez-Suay, Valero Laparra, Gonzalo Mateo-García, Jordi Muñoz-Marí, Luis Gómez-Chova, and Gustau Camps-Valls. Fair kernel learning. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pp. 339–355. Springer, 2017.

Kevin Petrasic, Benjamin Saul, James Greig, Matthew Bornfreund, and Katherine Lamberth. Algorithms and bias: What lenders need to know. *White & Case*, 2017.

Novi Quadrianto, Viktoriia Sharmanska, and Oliver Thomas. Discovering fair representations in the data domain. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 8227–8236, 2019.

Alvin Rajkomar, Michaela Hardt, Michael D Howell, Greg Corrado, and Marshall H Chin. Ensuring fairness in machine learning to advance health equity. *Annals of internal medicine*, 169(12):866–872, 2018.

Adriana Romero, Nicolas Ballas, Samira Ebrahimi Kahou, Antoine Chassang, Carlo Gatta, and Yoshua Bengio. Fitnets: Hints for thin deep nets. In *International Conference on Learning Representations (ICLR)*, 2015. URL http://arxiv.org/abs/1412.6550.

Bharath K Sriperumbudur, Kenji Fukumizu, and Gert Lanckriet. On the relation between universality, characteristic kernels and RKHS embedding of measures. In *International Conference on Artificial Intelligence and Statistics*, pp. 773–780, 2010.

Bahar Taskesen, Viet Anh Nguyen, Daniel Kuhn, and Jose Blanchet. A distributionally robust approach to fair classification. *arXiv preprint arXiv:2007.09530*, 2020.

Praneeth Vepakomma, Otkrist Gupta, Abhimanyu Dubey, and Ramesh Raskar. Reducing leakage in distributed deep learning for sensitive health data. In *ICLR AI for social good workshop*, 2019.

Mei Wang and Weihong Deng. Mitigating bias in face recognition using skewness-aware reinforcement learning. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

Mei Wang, Weihong Deng, Jiani Hu, Xunqiang Tao, and Yaohai Huang. Racial faces in the wild: Reducing racial bias by information maximization adaptation network. In *IEEE International Conference on Computer Vision (CVPR)*, 2019.

Dennis Wei, Karthikeyan Natesan Ramamurthy, and Flavio du Pin Calmon. Optimized score transformation for fair classification. *arXiv preprint arXiv:1906.00066*, 2019.

Denny Wu, Yixiu Zhao, Yao-Hung Hubert Tsai, Makoto Yamada, and Ruslan Salakhutdinov. "Dependency Bottleneck" in auto-encoding architectures: an empirical study. *CoRR*, abs/1802.05408, 2018. URL http://dblp.uni-trier.de/db/journals/corr/corr1802.html#abs-1802-05408.

Junho Yim, Donggyu Joo, Jihoon Bae, and Junmo Kim. A gift from knowledge distillation: Fast optimization, network minimization and transfer learning. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rogriguez, and Krishna P. Gummadi. Fairness constraints: Mechanisms for fair classification. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2017.

Sergey Zagoruyko and Nikos Komodakis. Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer. In *International Conference on Learning Representations (ICLR)*, 2017.

Brian Hu Zhang, Blake Lemoine, and Margaret Mitchell. Mitigating unwanted biases with adversarial learning. In *AAAI/ACM Conference on AI, Ethics, and Society (AIES)*, 2018.

A More Preliminary Experiments

In this appendix, we present more result to show the distribution of the leaked sensitive attribution. The results are consistent to the results discussed in Section 2 that the distribution shows that leakage of sensitive information in non-sensitive attributes are various over different samples.

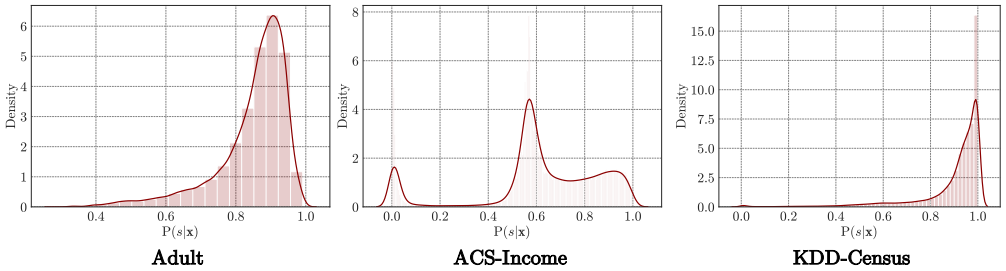

Figure 8: The distribution of $\mathrm{P}(s|\mathbf{x})$. The probability of $\mathrm{P}(s|\mathbf{x})$ is predicted by $\mathbf{x}$. The sensitive attribute is race. The distribution shows that leakage of sensitive information in non-sensitive attributes varies over different samples. More results are presented in Appendix A.