Joint Adversarial Learning for Cross-domain Fair Classification

Yueqing Liang\textsuperscript{1}, Canyu Chen\textsuperscript{1}, Tian Tian\textsuperscript{2}, and Kai Shu\textsuperscript{1}

\textsuperscript{1} Department of Computer Science, Illinois Institute of Technology
\{yliang40, cchen151\}@hawk.iit.edu, kshu@iit.edu
\textsuperscript{2} Stuart School of Business, Illinois Institute of Technology
ttian4@hawk.iit.edu

Abstract. Modern machine learning (ML) models are becoming increasingly popular and are widely used in decision-making systems. However, studies have shown critical issues of ML discrimination and unfairness, which hinder their adoption on high-stake applications. Recent research on fair classifiers has drawn significant attention to develop effective algorithms to achieve fairness and good classification performance. Despite the great success of these fairness-aware machine learning models, most of the existing models require sensitive attributes to preprocess the data, regularize the model learning or postprocess the prediction to have fair predictions. However, sensitive attributes are often incomplete or even unavailable due to privacy, legal or regulation restrictions. Though we lack the sensitive attribute for training a fair model in the target domain, there might exist a similar domain that has sensitive attributes. Thus, it is important to exploit auxiliary information from the similar domain to help improve fair classification in the target domain. Therefore, in this paper, we study a novel problem of exploring domain adaptation for fair classification. We propose a new framework that can simultaneously estimate the sensitive attributes while learning a fair classifier in the target domain. Extensive experiments on real-world datasets illustrate the effectiveness of the proposed model for fair classification, even when no sensitive attributes are available in the target domain.

Keywords: Fairness · Adversarial learning · Domain adaptation

1 Introduction

The recent development of machine learning models has been increasingly used for high-stake decision making such as filtering loan applicants [15], deploying police officers [27], etc. However, a prominent concern is when a learned model has bias against some specific demographic group such as race or gender. For example, a recent report shows that softwares used by schools to filter student applications may be biased to specific race group\textsuperscript{3}; and COMPAS [29], a tool

\textsuperscript{3} https://www.fastcompany.com/90342596/schools-are-quietly-turning-to-ai-to-help-pick-who-gets-in-what-could-go-wrong
for crime prediction, is shown to be more likely to assign a higher risk score to African-American offenders than to Caucasians with the same profile. Bias in algorithms can emanate from unrepresentative or incomplete training data, or the flawed information that reflects historical inequalities [8], which can lead to unfair decisions that have a collective and disparate impact on certain groups of people. Therefore, it is important to ensure fairness in machine learning models.

Recent research on fairness machine learning has drawn significant attention to develop effective algorithms to achieve fairness and maintain good prediction performance [3]. However, the majority of these models require sensitive attributes to preprocess the data, regularize the model learning or postprocess the prediction to have fair predictions. For example, Kamiran et al. propose to assign the weights of each training sample differently by reweighting to ensure fairness before model training [18]. In addition, Madras et al. utilize adversarial training on sensitive attributes and prediction labels for debiasing classification results [25]. However, sensitive attributes are often incomplete or even unavailable due to privacy, legal, or regulation restrictions. For example, by the law of the US Consumer Financial Protection Bureau (CFPB), creditors may not request or collect information about an applicant’s race, color, sex, etc.

Although we lack the sensitive attribute for training a fair model in the target domain, there might exist a similar domain that has sensitive attributes, which paves us a way to transfer the knowledge from the source domain to ensure the fairness in the target domain. For example, as shown in Fig. 1, when historical features and labels are available for credit risk assessment, while user demographics (e.g., gender) are unknown when expanding into new markets, the provider may learn models from its existing market data in the source domain to transfer the information of sensitive attributes to the new market in the target domain. Though there are extensive work on domain adaptation, they are overwhelmingly on classification or knowledge transfer [40,32]; while the work

---

4 [link]
Joint Adversarial Learning for Cross-domain Fair Classification

on fairness classification through domain adaptation is rather limited. Recent advancement of unsupervised domain adaptation [10,34] have shown promising results for adapting a classifier on source domain to the unlabeled data of the target domain. Although promising, unsupervised domain adaptation is not explored for fair classifiers.

Therefore, we propose to investigate an unsupervised domain adaptation framework to transfer the fairness without sensitive attributes in the target domain. In essence, we investigate the following challenges: (1) how to estimate the sensitive attributes in the target domain by transferring knowledge from the source domain; and (2) how to predict the labels accurately and satisfy fairness criteria. It is non-trivial to estimate the sensitive attributes in the target domain since there are domain discrepancies across domains and no anchor linkages are explicitly available for knowledge transferring. Our solutions to these challenges result in a novel framework called FairDA for fairness classification via domain adaptation. Our main contributions are summarized as follows:

– We study a novel problem of fair classification via domain adaptation.
– We provide a new framework FairDA that simultaneously transfers knowledge to estimate sensitive attributes and learns a fair classifier in the target domain with a dual adversarial learning approach.
– We conduct theoretic analysis demonstrating the fairness in the target domain can be achieved with estimated sensitive attributes derived from the source domain.
– We perform extensive experiments on real-world datasets to demonstrate the effectiveness of the proposed method for fair classification without sensitive attributes.

2 Problem Statement

We first introduce the notations of this paper, and then give the formal problem definition. Let \( D^1 = \{X^1, A^1, Y^1\} \) denote the data in the source domain, where \( X^1, A^1, \) and \( Y^1 \) represent the set of data samples, sensitive attributes, and corresponding labels. Let \( D^2 = \{X^2, Y^2\} \) be the target domain, where the sensitive attributes of the target domain are unknown. The data distributions in the source and target domains are similar but different with the domain discrepancy. Following existing work on fair classification [3,26], we evaluated fairness performance using metrics such as equal opportunity and demographic parity. Without loss of generality, we consider binary classification. Equal opportunity requires that the probability of positive instances with arbitrary sensitive attributes \( A \) being assigned to a positive outcome are equal: \( \mathbb{E}(\hat{Y} \mid A = a, Y = 1) = \mathbb{E}(\hat{Y} \mid A = b, Y = 1) \), where \( \hat{Y} \) is predicted label. Demographic parity requires the behavior of prediction model to be fair on different sensitive groups. Concretely, it requires that the positive rate across sensitive attributes are equal: \( \mathbb{E}(\hat{Y} \mid A = a) = \mathbb{E}(\hat{Y} \mid A = b), \forall a, b \). The problem of fair classification with domain adaptation is formally defined as follows:
Fig. 2: An illustration of the proposed framework FairDA. It consists of two major modules: (1) an adversarial domain adaptation module containing a sentiment attribute predictor and a domain classifier; and (2) an adversarial debiasing module containing a label predictor and a bias predictor.

**Problem Statement:** Given the training data $D^1$ and $D^2$ from the source and target domain, learn an effective classifier for the target domain while satisfying the fairness criteria such as demographic parity.

### 3 FairDA: Fair Classification with Domain Adaptation

In this section, we present the details of the proposed framework for fair classification with domain adaptation. As shown in Figure 2, our framework consists of two major modules: (1) an adversarial domain adaptation module that estimates sensitive attributes for the target domain; and (2) an adversarial debiasing module to learn a fair classifier in the target domain.

Specifically, first, the adversarial domain adaptation module contains a sensitive attribute predictor $f_{\theta_A}$ that describes the modeling of predicting sensitive attributes and a domain classifier $f_{\theta_d}$ that illustrates the process of transferring knowledge to the target domain for estimating sensitive attributes. In addition, the adversarial debiasing module consists of a label predictor $f_{\theta_Y}$ that models the label classification in the target domain, and a bias predictor $f_{\theta_a}$ to illustrate the process of differentiating the estimated sensitive attributes from the target domain data.
3.1 Estimating Target Sensitive Attributes

Since the sensitive attributes in the target domain are unknown, it is necessary to estimate the target sensitive attributes to build fair classifiers. Recent advancement of unsupervised domain adaptation has shown promising results to adapt a classifier on a source domain to the unlabeled data to the target domain [10]. Therefore, we propose to investigate an unsupervised domain adaptation framework to infer the sensitive attributes in the target domain. Specifically, to minimize the prediction error of sensitive attributes in the source domain, we aim to optimize the following objective function:

$$L_A = \min_{\theta_A} \mathbb{E}_{(X^1, A^1) \in D^1} \ell(A^1, f_{\theta_A}(h_1(X^1)))$$ \hspace{1cm} (1)

where $f_{\theta_A}$ is used to predict sensitive attributes, $\ell$ denotes the loss function to minimize prediction error, such as cross-entropy loss, and the embedding function $h_1(\cdot)$ is to learn the representation for predicting sensitive attributes. To ensure that $f_{\theta_A}$ trained on $D^1$ can well estimate the sensitive attributes of $D^2$. We further introduce a domain classifier $f_{\theta_d}$ to make the representation of source and target domains in the same feature space. Specifically, $f_{\theta_d}$ aims to differentiate if a representation is from the source or the target; while $h_1(\cdot)$ aims to learn a domain-invariant representation that can fool $f_{\theta_d}$, i.e.,

$$L_d = \min_{\theta_{h_1}} \max_{\theta_d} \mathbb{E}_{X^1 \in D^1} [\log(f_{\theta_d}(h_1(X^1)))] +$$
$$\mathbb{E}_{X^2 \in D^2} [\log(1 - f_{\theta_d}(h_1(X^2)))]$$ \hspace{1cm} (2)

where $\theta_d$ are the parameters of the domain classifier. The overall objective function for this module is a minmax game between the sensitive attributes predictor $f_{\theta_A}$ and the domain classifier $f_{\theta_d}$ as follows:

$$L_1 = \min_{\theta_A, \theta_{h_1}} \max_{\theta_d} \mathcal{L}_A - \alpha \mathcal{L}_d$$ \hspace{1cm} (3)

where $\alpha$ control the importance of the adversarial learning of the domain classifier. To estimate the sensitive attributes in the target domain, we will utilize the following function: $A^2 = f_{\theta_A}(h_1(X^2))$.

3.2 Adversarial Debiasing for Fair Classification

To learn a fair classifier for the target domain, we will leverage the derived sensitive attributes $A^2$ and the labels $Y^2$ in the target domain. Specifically, we aim to learn representation that can predict the labels accurately, while being irrelevant to sensitive attributes. To this end, we propose to feed the representation into a label predictor and a bias predictor. Specifically, the label predictor is to minimize the predict errors of the labels with the following objective function:

$$\mathcal{L}_Y = \min_{\theta_Y} \mathbb{E}_{(X^2, Y^2) \in D^2} \ell(Y^2, f_{\theta_Y}(h_2(X^2)))$$ \hspace{1cm} (4)
where $f_{\theta_Y}$ is to predict the labels, $h_2(\cdot)$ is an embedding layer to encode the features into a latent representation space, and $\ell$ is a cross-entropy loss. Furthermore, to learn fair representations and make fair predictions in the target domain, we incorporate a bias predictor to predict the sensitive attributes; while $h_2(\cdot)$ trying to learn the representation that can fool $f_{\theta_Y}$:

$$L_a = \min_{\theta_{h_2}} \max_{\theta_a} \mathbb{E}_{X^2 \sim p(X^2 | \hat{A}^2 = 1)} \left[ \log(f_{\theta_a}(h_2(X^2))) \right]$$

$$+ \mathbb{E}_{X^2 \sim p(X^2 | \hat{A}^2 = 0)} \left[ \log(1 - f_{\theta_a}(h_2(X^2))) \right]$$

(5)

where $\theta_a$ are the parameters for the adversary to predict sensitive attributes. Finally, the overall objective function of adversarial debiasing for fair classification is a minmax function:

$$L_2 = \min_{\theta_{h_2}} \max_{\theta_Y} \mathcal{L}_Y - \beta L_a$$

(6)

where $\beta$ control the importance of the bias predictor.

### 3.3 The Proposed Framework: FairDA

We have introduced how we can estimate sensitive attributes by adversarial domain adaptation, and how to ensure fair classification with adversarial debiasing. We integrate the components together and the overall objective function of the dual adversarial learning is as follows:

$$\mathcal{L} = \min_{\theta_{h_1}, \theta_{h_2}, \theta_Y, \theta_A} \max_{\theta_{d}, \theta_a} \mathcal{L}_1 + L_2$$

(7)

The parameters in the objective are learned through RMSprop, which is an adaptive learning rate method that divides the learning rate by an exponentially decaying average of squared gradients. We choose RMSprop as the optimizer because it is a popular and effective method to determine the learning rate adaptively, which is widely used for training neural networks. We will also prioritize the training of $f_{\theta_A}$ to ensure a good estimation of $\hat{A}^2$. Next, we will conduct a theoretic analysis on the dual adversarial learning for fairness guarantee.

### 4 A Theoretic Analysis on Fairness Guarantee

In this section, we perform a theoretic analysis of the fairness guarantee under the proposed framework FairDA with key assumptions. The model essentially contains two modules: (1) the adversarial domain adaptation for estimating sensitive attributes ($\mathcal{L}_1$ in Eqn. 3); and (2) the adversarial debiasing for learning a fair classifier ($\mathcal{L}_2$ in Eqn. 6). The performance of the second module is relying on the output (i.e., $\hat{A}^2$) of the first module.

To understand the theoretic guarantee for the first module, we follow the recent work on analyzing the conventional unsupervised domain adaptation [24,37].
The majority of them consider the model effectiveness under different types of domain shifts such as label shift, covariate shift, conditional shift, etc [24,21]. We assume that there exists a sensitive attribute shift across two domains, i.e., the prior distribution changes as \( p(A^1) \neq p(A^2) \). In addition, in our empirical application (see Table 1 in Section 5), we can assume that the condition probability holds, i.e., \( p(X^1|A^1) = p(X^2|A^2) \). For example, in the COMPAS data, the distribution of sensitive attributes may be different for different age groups, while the inherent relationship between the individual features and the sensitive attribute may remain the same. Under the assumption of the sensitive attribute shift, we can derive the risk in the target domain as [21]:

\[
R^2(h_1, \theta_A) = \sum_{A \in \mathcal{A}} \int_{X} \ell(A, f_{\theta_A}(h_1(X))) \frac{p(X^2|A^2)p(A^2)}{p(X^1|A^1)p(A^1)} p(X^1, A^1) \, dX
\]

\[
= \sum_{A \in \mathcal{A}} \int_{X} \ell(A, f_{\theta_A}(h_1(X))) \frac{p(A^2)}{p(A^1)} p(X^1, A^1) \, dX
\]

where the ratio \( p(A^2)/p(A^1) \) represents the change in the proportions in the sensitive attribute. Since we do not have samples with sensitive attributes from the target domain, we can use the samples drawn from the source distribution to estimate the target sensitive attributes distribution with mean matching [13] by minimizing the following function \( \|M^1p(A^1) - \mu^2\|_2^2 \): where \( M^1 \) is the vector empirical sample means from the source domain, i.e., \( \mu^1(f_{\theta_A}(h_1(X^1))|A^1 = 0), \mu^2(f_{\theta_A}(h_1(X^1))|A^1 = 1) \). \( \mu^2 \) is the encoded feature means for the target. We will incorporate the above strategy for estimating the target sensitive attribute proportions with gradient descent during the adversarial training [23].

The noise induced for estimating the target sensitive attributes \( \hat{A}^2 \) is non-negligible, which may influence the adversarial debiasing in the second module. Next we theoretically show that under mild conditions, we can satisfying fairness metrics such as demographic parity. First, we can prove that the global optimum of \( \mathcal{L}_2 \) can be achieved if and only if \( p(X^2|\hat{A}^2 = 1) = p(X^2|\hat{A}^2 = 0) \) according to the Proposition 1. in [12]. Next, we introduce the following theorem under two reasonable assumptions:

**Theorem 1.** Let \( \hat{Y}^2 \) denotes the predicted label of the target domain, if

1. The estimated sensitive attributes \( \hat{A}^2 \) and \( X^2 \) are independent conditioned on the true sensitive attributes, i.e., \( p(\hat{A}^2, X^2|A^2) = p(\hat{A}^2|A^2)p(X^2|A^2) \);
2. The estimated sensitive attributes are not random, i.e., \( p(A^2 = 1|\hat{A}^2 = 1) \neq p(|A^2 = 0|\hat{A}^2 = 1) \)

If \( \mathcal{L}_2 \) reaches the global optimum, the label prediction \( f_{\theta_V} \) will achieve demographic parity, i.e., \( p(\hat{Y}^2|A^2 = 0) = p(\hat{Y}^2|A^2 = 1) \)

We first explain the two assumptions: (1) since we use two separate embedding layers \( h_1(\cdot) \) and \( h_2(\cdot) \) to predict the target sensitive attributes, and learn the latent presentation, it generally holds that \( \hat{A}^2 \) is independent with the representation of \( X^2 \), i.e., \( p(\hat{A}^2, X^2|A^2) = p(\hat{A}^2|A^2)p(X^2|A^2) \); (2) Since we are using adversarial learning to learn an effective estimator \( f_{\theta_A} \) for sensitive attributes, it
is reasonable to assume that it does not produce random prediction results. We prove Theorem 1 as follows: since \( p(X^2, A^2 | A^2) = p(A^2 | A^2) p(X^2 | A^2) \), we have \( p(X^2, A^2 | A^2) = p(X^2 | A^2) \). In addition, when the algorithm converges, we have \( p(X^2 | A^2 = 1) = p(X^2 | A^2 = 0) \), which is equivalent with \( \sum_{A^2} p(X^2, A^2 | A^2 = 1) = \sum_{A^2} p(X^2, A^2 | A^2 = 0) \). Therefore,
\[
\sum_{A^2} p(X^2 | A^2) p(A^2 | A^2 = 1) = \sum_{A^2} p(X^2 | A^2) p(A^2 | A^2 = 0) \tag{9}
\]
Based on the above equation, we can get,
\[
\frac{p(X^2 | A^2 = 1)}{p(X^2 | A^2 = 0)} = \frac{p(A^2 = 0 | A^2 = 1) - p(A^2 = 0 | A^2 = 0)}{p(A^2 = 1 | A^2 = 0) - p(A^2 = 1 | A^2 = 1)} = 1 \tag{10}
\]
which shows that a global minimum achieves, i.e., \( p(X^2 | A^2 = 1) = p(X^2 | A^2 = 0) \). Since \( \hat{Y}^2 = f_{\theta_1}(X^2) \), we can get \( p(Y^2 | A^2 = 1) = p(Y^2 | A^2 = 0) \), which is the demographic parity.

5 Experiments

In this section, we present the experiments to evaluate the effectiveness of FairDA. We aim to answer the following research questions (RQs):

– **RQ1**: Can FairDA obtain fair predictions without accessing sensitive attributes in the target domain?

– **RQ2**: How can we transfer fairness knowledge from the source domain while effectively using it to regularize the target’s prediction?

– **RQ3**: How would different choices of the weights of the two adversarial components impact the performance of FairDA?

5.1 Datasets

We conduct experiments on two publicly available benchmark datasets for fair classification: COMPAS [17] and Adult [2].

– **COMPAS**: This dataset describes the task of predicting the recidivism of individuals in the U.S.

– **ADULT**: This dataset contains records of personal yearly income, and the label is whether the income of a specific individual exceed 50k or not.

For each of the original datasets, we first define a proper filter to group the dataset into the source domain and the target domain. For example, in the COMPAS data, we assume that the individuals whose ages less than 24 are not allowed to collect the sensitive attributes (e.g., sex and race). Similarly, for the ADULT data, we consider the sex of each individual is the sensitive attribute, and define two filters based on the working class and country to derive the source and target domains. The evaluation of all models are measured by the performance on target dataset. The statistics of the datasets are shown in Table 1.

5 https://github.com/propublica/compas-analysis
6 https://archive.ics.uci.edu/ml/machine-learning-databases/adult/
Joint Adversarial Learning for Cross-domain Fair Classification

Table 1: Statistics of datasets. SA refers to sensitive attributes. The filter defines the rule to group the original data to the source and target domains.

| Exp. | Ori. data | SA | Filter | Dataset | Domain | SA avail. | # Instance |
|------|-----------|----|--------|---------|---------|-----------|------------|
| 1    | COMPAS   | Sex | age < 24 | COMPAS 1 | Source  | Yes       | 881        |
| 2    | COMPAS   |      | age ≥ 24 | COMPAS 2 | Target  | No        | 4,397      |
| 3    | COMPAS   | Race | age < 24 | COMPAS 3 | Source  | Yes       | 881        |
| 4    | COMPAS   |      | age ≥ 24 | COMPAS 4 | Target  | No        | 4,397      |
| 5    | ADULT    | Sex | work class ≠ private | ADULT 1 | Source  | Yes       | 11,915     |
| 6    | ADULT    |      | work class = private | ADULT 2 | Target  | No        | 33,307     |
| 7    | ADULT    | Sex | country = US | ADULT 3 | Source  | Yes       | 41,292     |
| 8    | ADULT    |      | country ≠ US | ADULT 4 | Target  | No        | 3,930      |

5.2 Experimental Settings

Baselines Since there is no existing work on cross-domain fair classification, we compare our proposed FairDA with the following representative methods in fair classification without sensitive attributes. We also compare the vanilla model as the lower bound of fairness performance.

- **Vanilla**: This method directly train a base classifier without explicitly regularizing on the sensitive attributes.
- **ARL** [22]: This method utilize reweighting on under-represented regions detected by adversarial model to alleviate bias.
- **KSMOTE** [35]: It first derive pseudo groups, and then use them to design regularization to ensure fairness.
- **FairRF** [38]: It optimizes the prediction fairness without sensitive attribute but with some available features that are known to be correlated with the sensitive attribute.

Configurations For KSMOTE\(^7\) and FairRF\(^8\), we directly use the code provided by the authors. For all other approaches, we adopt a three-layer multi-layer-perceptron (MLP) as the base classifier. We set the two hidden dimensions of MLP as 64 and 32, and use Adam optimizer with 0.001 as initial learning rate.

Evaluation Metrics Following existing work on fair models, we measure the classification performance with Accuracy (ACC) and F1, and the fairness performance based on Demographic Parity and Equal Opportunity [26]. We consider the scenario when the sensitive attributes and labels are binary, which can be naturally extended to more general cases.

- Demographic Parity: A classifier is considered to be fair if the prediction $\hat{Y}$ is independent from the sensitive attribute $A$. In other words, demographic

---

\(^7\) https://imbalanced-learn.org/stable/
\(^8\) https://github.com/TianxiangZhao/fairlearn
parity requires each demographic group has the same chance for a positive outcome: $E(\hat{Y}|A = 1) = E(\hat{Y}|A = 0)$. We will report the difference of each sensitive group’s demographic parity ($\Delta_{DP}$):

$$\Delta_{DP} = |E(\hat{Y}|A = 1) - E(\hat{Y}|A = 0)|$$  (11)

- Equal Opportunity: Equal opportunity considers one more condition than demographic parity. A classifier is said to be fair if the prediction $\hat{Y}$ of positive instances is independent from the sensitive attributes. Concretely, it requires the true positive rate of different sensitive groups are equal: $E(\hat{Y}|A = 1, Y = 1) = E(\hat{Y}|A = 0, Y = 1)$. Similarly, in the experiments, we report the difference of each sensitive group’s equal opportunity ($\Delta_{EO}$):

$$\Delta_{EO} = |E(\hat{Y}|A = 1, Y = 1) - E(\hat{Y}|A = 0, Y = 1)|$$  (12)

Note that demographic parity and equal opportunity measure the fairness performance in different ways, and the smaller the values are, the better the performance of fairness.

5.3 Fairness Performances (RQ1)

To answer RQ1, we compare the proposed framework with the aforementioned baselines in terms of prediction and fairness. The parameters for the baselines are tuned through grid search, and for the setting of FairRF, we follow the original paper [38]. All experiments are conducted 3 times and the average performances (standard deviations results are shown in Appendix due to space limitation) are reported in Table 2 and Table 3 in terms of accuracy, $\Delta_{EO}$, $\Delta_{DP}$. We have the following observations:

- In general, the proposed framework FairDA can achieve a better performance of fairness, while not causing a significant drop of the prediction performance comparing with the Vanilla model. For example, in Exp. 4, the performance of $\Delta_{DP}$ of FairDA increases 20.6% compared to Vanilla, and the performance of accuracy only drops 1.7%.

- Comparing with all other fairness-aware baselines that are not utilizing sensitive attributes in the target domain, overall FairDA performs better in terms of fairness metrics $\Delta_{DP}$ and $\Delta_{EO}$. This is because by transferring sensitive information from the source domain, FairDA may make adversarial debiasing with the transferred sensitive information.

- We conduct experiments on various source and target datasets. No matter it is transferred from the source that has less or more samples than target, our proposed FairDA outperforms baselines on the target domain, which suggests the generality of our model.
Table 2: Fairness performance comparison in the COMPAS dataset.

| Model   | ACC  | F1   | ∆DP | ∆EO | ACC  | F1   | ∆DP | ∆EO |
|---------|------|------|-----|-----|------|------|-----|-----|
| Vanilla | 0.677| 0.639| 0.128| 0.137| 0.675| 0.644| 0.141| 0.143|
| ARL     | 0.672| 0.609| 0.109| 0.119| 0.663| 0.598| 0.165| 0.171|
| KSMOTE  | 0.686| 0.648| 0.136| 0.146| 0.682| 0.651| 0.150| 0.148|
| FairRF  | 0.641| 0.527| 0.086| 0.108| 0.631| 0.603| 0.103| 0.105|
| FairDA  | 0.661| 0.639| 0.050| 0.042| 0.611| 0.658| 0.085| 0.096|

Table 3: Fairness performance comparison in the ADULT dataset.

| Model   | ACC  | F1   | ∆DP | ∆EO | ACC  | F1   | ∆DP | ∆EO |
|---------|------|------|-----|-----|------|------|-----|-----|
| Vanilla | 0.857| 0.655| 0.100| 0.099| 0.839| 0.658| 0.107| 0.093|
| ARL     | 0.837| 0.668| 0.142| 0.122| 0.816| 0.673| 0.142| 0.079|
| KSMOTE  | 0.840| 0.511| 0.072| 0.138| 0.824| 0.589| 0.102| 0.140|
| FairRF  | 0.848| 0.594| 0.083| 0.092| 0.829| 0.611| 0.087| 0.068|
| FairDA  | 0.844| 0.595| 0.081| 0.085| 0.825| 0.599| 0.085| 0.070|

5.4 Ablation Study (RQ2)

In this section, we aim to analyse the effectiveness of each component in our proposed FairDA framework. As shown in Section 3, our model contains two adversarial components. The first one is designed for making better prediction on target domain’s sensitive attribute with the knowledge learned from source domain. The second one is designed for debiasing with the sensitive attribute predicted by the domain adaptive classifier. In order to answer RQ2, we investigate the effects of these components by defining two variants of FairDA:

- FairDA w/o DA: FairDA w/o DA is a variant of FairDA without domain adversary during the training of sensitive attribute predictor. It first trains a classifier with the source domain’s sensitive attribute and directly apply it to predict the target domain’s sensitive attribute.
- FairDA w/o Debiasing: FairDA w/o Debiasing is a variant of FairDA without debiasing adversary during the training on target domain. It trains a domain adaptive sensitive attribute predictor on the source domain, then use the trained predictor to estimate the sensitive attribute on the target domain. After obtaining target’s sensitive attribute, this variant just treats it as an additional normal feature to predict the label without fairness constraints.

Table 4 reports the average performances with standard deviations for each method. We make the following observations: (1) When we eliminate the domain adversary for learning a sensitive attribute predictor, the performances are reduced. It suggests the importance of capturing the domain invariant sensitive
Table 4: Effectiveness of each component in FairDA.

| Variants                  | ACC    | F1      | $\Delta_{DP}$ | $\Delta_{EO}$ |
|---------------------------|--------|---------|---------------|---------------|
| FairDA w/o DA             | 0.664±0.025 | 0.624±0.042 | 0.062±0.014 | 0.077±0.007 |
| FairDA w/o Debiasing      | 0.675±0.014 | 0.644±0.023 | 0.141±0.003 | 0.143±0.007 |
| FairDA                    | 0.661±0.023 | 0.639±0.016 | 0.050±0.014 | 0.042±0.008 |

We have a similar observation for FairDA w/o debiasing. For all the three metrics, when we remove the adversarial component of debiasing, both the prediction and fairness are reduced. The results demonstrate the significance of including the debiasing component.

Fig. 3: Parameter sensitivity analysis to assess the impact of domain adversary and debiasing adversary components.
5.5 Parameter Analysis (RQ3)

To answer RQ3, we explore the parameter sensitivity of the two important hyperparameters of our model. \( \alpha \) controls the impact of the adversary to the sensitive attribute predictor, while \( \beta \) controls the influence of the adversary to the debiasing. We vary both \( \alpha \) and \( \beta \) from \([0.0001, 0.001, 0.01, 0.1, 1]\), and other settings are the same as Exp.1. The results are shown in Fig. 3. It is worth noting that, for Fig. 3(a) and Fig. 3(b), lower values are better fairness performances, while for Fig. 3(c) and Fig. 3(d), higher values indicate better prediction performance. From Fig. 3(a) and Fig. 3(b) we could observe that: (1) Generally, larger \( \alpha \) and \( \beta \) will achieve fairer predictions, while smaller \( \alpha \) and \( \beta \) result in worse fairness; (2) When we increase the value of \( \alpha \), both \( \Delta_{DP} \) and \( \Delta_{EO} \) will first decrease and then increase when the value of \( \alpha \) is too large. This may be because the estimated target sensitive attributes are still noisy and strengthening the adversarial debiasing that optimizes the noisy sensitive attributes may not necessarily lead to better fairness performances; (3) For \( \beta \), the fairness performance is consistently better with a higher value. From Fig. 3(c), we observe that the classification performance is better when \( \alpha \) and \( \beta \) are balanced. Overall, we could make a conclusion that when \( \alpha \) and \( \beta \) are in the range of \([0.001, 0.1]\), FairDA can achieve relatively good performances of fairness and classification.

6 Related Work

In this section, we briefly review the related works on fairness machine learning and deep domain adaptation.

**Fairness Machine Learning** Recent research on fairness in machine learning has drawn significant attention to developing effective algorithms to achieve fairness and maintain good prediction performance. which generally focus on individual fairness [19] or group fairness [16,36]. Individual fairness requires the model to give similar predictions to similar individuals [6,5]. In group fairness, similar predictions are desired among multiple groups categorized by a specific sensitive attribute (e.g., gender). Other niche notions of fairness includes subgroup fairness [20] and Max-Min fairness[22], which aims to maximize the minimum expected utility across groups. In this work, we focus on group fairness. To improve group fairness, debiasing techniques have been applied at different stages of a machine learning model: (1) Pre-processing approaches [18] apply dedicated transformations on the original dataset to remove intrinsic discrimination and obtain unbiased training data prior modeling; (2) In-processing approaches [1,4] tackle this issue by incorporating fairness constraints or fairness-related objective functions to the design of machine learning models; and (3) Post-processing approaches [9] revise the biased prediction labels by debiasing mechanisms.

Although the aforementioned methods can improve group fairness, they generally require the access of sensitive attributes, which is often infeasible. Very few recent work study fairness with limited sensitive attributes. For example, [38] and [14] explore and use the related features as proxies of the sensitive attribute to achieve better fairness results. [22] proposes an adversarial reweighted method
to achieve the Rawlsian Max-Min fairness objective which aims at improving the accuracy for the worst-case protected group. However, these methods may require domain knowledge to approximate sensitive attributes or not suitable for ensuring group fairness.

In this paper, we study the novel problem of cross-domain fairness classification which aims to estimate the sensitive attributes of the target domain to achieve the group fairness.

**Deep Domain Adaption:** Domain adaptation [28] aims at mitigating the generalization bottleneck introduced from domain shift. With the rapid growth of deep neural networks, deep domain adaptation has drawn much attention lately. In general, deep domain adaptation methods aim to learn a domain-invariant feature space that can reduce the discrepancy between the source and target domains. This goal is accomplished either by transforming the features from one domain to be closer to the other domain, or projecting both domains into a domain-invariant latent space [31]. For instance, TLDA [39] is a deep autoencoder-based model for learning domain-invariant representations for classification. Inspired by the idea of Generative Adversarial Network (GAN) [12], researchers also propose to perform domain adaptation in an adversarial training paradigm [11,33,31]. By exploiting a domain discriminator to distinguish the domain labels while learning deep features to confuse the discriminator, DANN [11] achieves superior domain adaptation performance. ADDA [33] learns a discriminative representation using labeled source data and then map the target data to the same space through an adversarial loss. Recently, very few work apply transfer learning techniques for fair classification [30,7]. Although sensitive attributes are not available in the target domain, there may exist some publicly available datasets which can be used as auxiliary sources. However, these methods mostly consider a shallow scenario of transfer between different sensitive attributes in the same dataset.

In this paper, we propose a new domain adaptive approach based on dual adversarial learning to achieve fair classification in the target domain.

### 7 Conclusion and Future Work

In this paper, we study a novel and challenging problem of exploiting domain adaptation for fair and accurate classification for a target domain without the availability of sensitive attributes. We propose a new framework **FairDA** using a dual adversarial learning approach to achieve a fair and accurate classification. We provide a theoretic analysis to demonstrate that we can learn a fair model prediction under mild assumptions. Experiments on real-world datasets show that the proposed approach can achieve a more fair performance compared to existing approaches by exploiting the information from a source domain, even without knowing the sensitive attributes in the target domain. For future work, first, we can consider multiple source domains available and explore how to exploit domain discrepancies across multiple domains to enhance the performance of fair classifier in the target domain. Second, we will explore the fairness transfer
under different types of domain shifts, such as the conditional shift of sensitive attributes. Third, we will explore other domain adaption approach such as meta transfer learning to achieve cross-domain fair classification.

References

1. Alekh Agarwal, Alina Beygelzimer, Miroslav Dudík, John Langford, and Hanna Wallach. A reductions approach to fair classification. In International Conference on Machine Learning, pages 60–69, 2018.
2. Arthur Asuncion and David Newman. Uci machine learning repository, 2007.
3. Solon Barocas, Moritz Hardt, and Arvind Narayanan. Fairness in machine learning. NeurIPS tutorial, 1-2, 2017.
4. Yahav Bechavod and Katrina Ligett. Learning fair classifiers: A regularization-inspired approach. arXiv preprint arXiv:1707.00044, pages 1733–1782, 2017.
5. Simon Caton and Christian Haas. Fairness in machine learning: A survey. CoRR, abs/2010.04053, 2020.
6. Lu Cheng, Kush R Varshney, and Huan Liu. Socially responsible ai algorithms: Issues, purposes, and challenges. arXiv preprint arXiv:2101.02032, 2021.
7. Amanda Coston, Karthikeyan Natesan Ramamurthy, Dennis Wei, Kush R Varshney, Skyler Speakman, Zairah Mustahsan, and Supriyo Chakraborty. Fair transfer learning with missing protected attributes. In Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society, pages 91–98, 2019.
8. Mengnan Du, Fan Yang, Na Zou, and Xia Hu. Fairness in deep learning: A computational perspective. IEEE Intelligent Systems, 36(4):25–34, 2020.
9. Cynthia Dwork, Nicole Immorlica, Adam Tauman Kalai, and Max Leiserson. Decoupled classifiers for group-fair and efficient machine learning. In Conference on Fairness, Accountability and Transparency, pages 119–133, 2018.
10. Yaroslav Ganin and Victor Lempitsky. Unsupervised domain adaptation by backpropagation. In International conference on machine learning, pages 1180–1189. PMLR, 2015.
11. Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. Domain-adversarial training of neural networks. The Journal of Machine Learning Research, 2016.
12. Ian J Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. arXiv preprint arXiv:1406.2661, 2014.
13. Arthur Gretton, Alex Smola, Jiayuan Huang, Marcel Schmittfull, Karsten Borgwardt, and Bernhard Schölkopf. Covariate shift by kernel mean matching. Dataset shift in machine learning, 3(4):5, 2009.
14. Maya Gupta, Andrew Cotter, Mahdi Milani Fard, and Serena Wang. Proxy fairness. arXiv preprint arXiv:1806.11212, 2018.
15. Aboobyda Jafar Hamid and Tarig Mohammed Ahmed. Developing prediction model of loan risk in banks using data mining. Machine Learning and Applications: An International Journal (MLAI) Vol, 3(1), 2016.
16. Moritz Hardt, Eric Price, and Nathan Srebro. Equality of opportunity in supervised learning. arXiv preprint arXiv:1610.02413, 2016.
17. Surya Mattu Julia Angwin, Jeff Larson and Lauren Kirchner. Machine bias: There’s software used across the country to predict future criminals and it’s biased against blacks. ProPublica, 2016.
18. Faisal Kamiran and Toon Calders. Data preprocessing techniques for classification without discrimination. *Knowledge and information systems*, 33(1):1–33, 2012.
19. Jian Kang, Jingrui He, Ross Maciejewski, and Hanghang Tong. Inform: Individual fairness on graph mining. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 379–389, 2020.
20. Michael Kearns, Seth Neel, Aaron Roth, and Zhiwei Steven Wu. Preventing fairness gerrymandering: Auditing and learning for subgroup fairness. In *International Conference on Machine Learning*, pages 2564–2572, 2018.
21. Wouter M Kouw and Marco Loog. An introduction to domain adaptation and transfer learning. *arXiv preprint arXiv:1812.11806*, 2018.
22. Preethi Lahoti, Alex Beutel, Jilin Chen, Kang Lee, Flavien Prost, Nithum Thain, Xuezhi Wang, and Ed H Chi. Fairness without demographics through adversarially reweighted learning. *arXiv preprint arXiv:2006.13114*, 2020.
23. Yitong Li, Michael Murias, Samantha Major, Geraldine Dawson, and David Carlson. On target shift in adversarial domain adaptation. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 616–625. PMLR, 2019.
24. Xiaofeng Liu, Zhenhua Guo, Site Li, Fangxu Xing, Jane You, C-C Jay Kuo, Georges El Fakhri, and Jonghye Woo. Adversarial unsupervised domain adaptation with conditional and label shift: Infer, align and iterate. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10367–10376, 2021.
25. David Madras, Elliot Creager, Toniann Pitassi, and Richard Zemel. Learning adversarially fair and transferable representations. In *International Conference on Machine Learning*, pages 3384–3393. PMLR, 2018.
26. Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. A survey on bias and fairness in machine learning. *arXiv preprint arXiv:1908.09635*, 2019.
27. Jesus Mena. *Machine learning forensics for law enforcement, security, and intelligence*. CRC Press, 2011.
28. Sinno Jialin Pan and Qiang Yang. A survey on transfer learning. *IEEE Transactions on knowledge and data engineering*, 2009.
29. M Redmond. Communities and crime unnormalized data set. *UCI Machine Learning Repository*. In website: http://www.ics.uci.edu/mlrepository.html, 2011.
30. Candice Schumann, Xuezhi Wang, Alex Beutel, Jilin Chen, Hai Qian, and Ed H Chi. Transfer of machine learning fairness across domains. *arXiv preprint arXiv:1906.09688*, 2019.
31. Yang Shu, Zhangjie Cao, Mingsheng Long, and Jianmin Wang. Transferable curriculum for weakly-supervised domain adaptation. 2019.
32. Chuanqi Tan, Fuchun Sun, Tao Kong, Wenchang Zhang, Chao Yang, and Chunfang Liu. A survey on deep transfer learning. In *International conference on artificial neural networks*, pages 270–279. Springer, 2018.
33. Eric Tzeng, Judy Hoffman, Kate Saenko, and Trevor Darrell. Adversarial discriminative domain adaptation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7167–7176, 2017.
34. Hongliang Yan, Yukang Ding, Peihua Li, Qilong Wang, Yong Xu, and Wangmeng Zuo. Mind the class weight bias: Weighted maximum mean discrepancy for unsupervised domain adaptation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2272–2281, 2017.
35. Shen Yan, Hsien-te Kao, and Emilio Ferrara. Fair class balancing: Enhancing model fairness without observing sensitive attributes. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*, pages 1715–1724, 2020.

36. Lu Zhang, Yongkai Wu, and Xintao Wu. Achieving non-discrimination in data release. In *SIGKDD*, pages 1335–1344, 2017.

37. Yuchen Zhang, Tianle Liu, Mingsheng Long, and Michael Jordan. Bridging theory and algorithm for domain adaptation. In *International Conference on Machine Learning*, pages 7404–7413. PMLR, 2019.

38. Tianxiang Zhao, Enyan Dai, Kai Shu, and Suhang Wang. You can still achieve fairness without sensitive attributes: Exploring biases in non-sensitive features. In *WSDM ’22, The Fourteenth ACM International Conference on Web Search and Data Mining, Phoenix, March 8-12, 2022*. ACM, 2022.

39. Fuzhen Zhuang, Xiaohu Cheng, Ping Luo, Sinno Jialin Pan, and Qing He. Supervised representation learning: Transfer learning with deep autoencoders. In *IJCAI*, 2015.

40. Fuzhen Zhuang, Zhiyuan Qi, Keyu Duan, Dongbo Xi, Yongchun Zhu, Hengshu Zhu, Hui Xiong, and Qing He. A comprehensive survey on transfer learning. *Proceedings of the IEEE*, 109(1):43–76, 2020.

## A Additional Results

Table 5 show the standard deviations of performances to Table 2 and Table 3.

| Model       | Exp. 1: COMPAS 1 → COMPAS 2 | Exp. 2: COMPAS 3 → COMPAS 4 |
|-------------|------------------------------|-----------------------------|
|             | ACC | F1 | ∆DP | ∆EO | ACC | F1 | ∆DP | ∆EO |
| Vanilla     | ±0.015 | ±0.018 | ±0.015 | ±0.005 | ±0.014 | ±0.023 | ±0.003 | ±0.007 |
| ConstrainS  | ±0.024 | ±0.011 | ±0.006 | ±0.017 | ±0.016 | ±0.010 | ±0.012 | ±0.014 |
| ARL         | ±0.008 | ±0.032 | ±0.038 | ±0.057 | ±0.001 | ±0.036 | ±0.015 | ±0.026 |
| KSMOTE      | ±0.012 | ±0.018 | ±0.017 | ±0.018 | ±0.010 | ±0.020 | ±0.012 | ±0.013 |
| FairRF      | ±0.013 | ±0.064 | ±0.020 | ±0.048 | ±0.007 | ±0.012 | ±0.019 | ±0.022 |
| FairDA      | ±0.023 | ±0.016 | ±0.014 | ±0.008 | ±0.021 | ±0.006 | ±0.022 | ±0.012 |

| Model       | Exp. 3: ADULT 1 → ADULT 2 | Exp. 4: ADULT 3 → ADULT 4 |
|-------------|---------------------------|---------------------------|
|             | ACC | F1 | ∆DP | ∆EO | ACC | F1 | ∆DP | ∆EO |
| Vanilla     | ±0.003 | ±0.009 | ±0.005 | ±0.014 | ±0.001 | ±0.001 | ±0.003 | ±0.009 |
| ConstrainS  | ±0.004 | ±0.010 | ±0.007 | ±0.015 | ±0.002 | ±0.014 | ±0.004 | ±0.014 |
| ARL         | ±0.006 | ±0.005 | ±0.019 | ±0.016 | ±0.015 | ±0.004 | ±0.025 | ±0.023 |
| KSMOTE      | ±0.002 | ±0.010 | ±0.008 | ±0.055 | ±0.004 | ±0.040 | ±0.016 | ±0.017 |
| FairRF      | ±0.003 | ±0.005 | ±0.001 | ±0.015 | ±0.003 | ±0.011 | ±0.006 | ±0.013 |
| FairDA      | ±0.002 | ±0.012 | ±0.004 | ±0.011 | ±0.004 | ±0.017 | ±0.004 | ±0.003 |