

# Preserving Fairness in AI under Domain Shift

**Serban Stan**

SSTAN@USC.EDU

**Mohammad Rostami**

ROSTAMIM@USC.EDU

*University of Southern California  
Los Angeles, CA, USA*

## Abstract

Existing algorithms for ensuring fairness in AI use a single-shot training strategy, where an AI model is trained on an annotated training dataset with sensitive attributes and then fielded for utilization. This training strategy is effective in problems with stationary distributions, where both the training and testing data are drawn from the same distribution. However, it is vulnerable with respect to distributional shifts in the input space that may occur after the initial training phase. As a result, the time-dependent nature of data can introduce biases and performance degradation into the model predictions, even if the model is initially fair. Model retraining from scratch using a new annotated dataset is a naive solution that is expensive and time-consuming. We develop an algorithm to adapt a fair model to remain fair and generalizable under domain shift using solely new unannotated data points. We recast this learning setting as an unsupervised domain adaptation (UDA) problem. Our algorithm is based on updating the model such that the internal representation of data remains unbiased despite distributional shifts in the input space. We provide empirical validation on three common fairness datasets to show that the challenge exists in practical setting and to demonstrate the effectiveness of our algorithm.

## 1. Introduction

AI has been extensively utilized in automating heavy and electronic industry tasks such as logistics & transportation, retail & e-commerce, and entertainment and gaming. The effectiveness of AI in these practical domains has motivated its adoption in various societal decision-making processes that are more consequential in lives of people. These applications encompass tasks such loan application processing, parole decisions in prison systems, healthcare decision-making, public policy decisions, and police deployments (Chouldechova & Roth, 2018). The increasingly growth of adopting AI can be attributed to the advances in deep learning, which enable the training of complex and generalizable neural architectures using large manually annotated datasets in an end-to-end scheme. A notable advantage of the data-driven learning approach is that it reduces the requirement for laborious feature engineering through the end-to-end training procedure. However, deep learning methods also come with disadvantages, such as the time-consuming training procedure, lack of interpretability, requiring costly data annotation, and increasingly growing power consumption. One of the important challenges for adopting AI is existence of bias in trained models.

It is well documented that some of the best AI models are biased against certain racial or gender sub-groups (Eidinger et al., 2014; Zhang et al., 2017; Cirillo et al., 2020), particularly from underrepresented groups, and can produce adverse outcomes for disadvantaged groups. Hence, fairness is a major concern for using AI in societal decision-making processes. This concern is particularly important in deep learning because data-driven learning

can unintentionally lead to training unfair models due to the inherent biases that exist in annotating training datasets by human workers or skewed data distributions conditioned on certain sensitive attributes (Buolamwini & Gebru, 2018). As a result, training models by simply minimizing the empirical error on relevant datasets may lead to identifying correlations between majority subgroup features and positive outcomes for them as irrelevant causes for decision-making. For example, the correlation between gender and income may be lead to identifying “being male” a feature approving loan requests with a higher chance. This unwanted outcome happens because statistical learning primarily discovers correlations rather than causation. Thus, the decision boundary of AI models may be informed by group-specific characteristics that are irrelevant to the decision task (Mehrabi et al., 2021). For example, since the income level is generally correlated positively with the male gender in reality, it can lead to training models with unfair decisions against female loan applicants.

The crucial concern about fairness in AI and the need to overcome the resulting adverse effects have resulted in significant research interest from the AI community. The first attempt to address bias in AI is to arrive at a commonly agreed-upon definition of fairness such that it can also be quantified. Pioneer works in this area focused on defining quantitative notions for fairness based on commonsense intuition and using them to quantitatively demonstrate the presence and severity of bias in AI (Buolamwini & Gebru, 2018; Caliskan et al., 2017). The metrics then can be used for training fair models by considering them as loss functions during model training. Most existing fairness metrics consider that the input data points possess characteristics of protected subgroups (Feldman et al., 2015), e.g., gender and race, in addition to standard features that are used for model training based on empirical risk minimization (ERM). Based on subgroup membership, majority and minority populations emerge, or in general subgroups, which can be used to define fairness metrics. A model is then assumed to be fair if its predictions possess a notion of probabilistic independence from data membership into the subgroups (Mehrabi et al., 2021). For example, we train a model to guarantee that predictions are independent from the gender attribute. (see Section 5.1.3 for definitions of common fairness metrics).

There are several approaches for training fair AI models. A major approach for reinforcing fairness in an AI models is to map data into a latent space in which data representations are independent from the sensitive attributes such as gender and race. For example, we can benefit from adversarial learning for this purpose by training the generative subnetwork to extract agnostic features with respect to the subgroups (Zhang et al., 2018). Since the sensitive attributes are absent in the latent space, decision-making cannot identify sensitive attributes for output prediction. Despite being an effective approach, most existing fair model training algorithms consider that the data distribution will remain stationary after the training stage. This assumption is rarely true in practical settings, particularly when a model is used over extended periods. The reason is that societal applications are dynamic but fairness metrics are normally static. For example, approving or declining loan requests depends on macroeconomic factors that are dynamic and time-dependent. On the other hand, correlations with sensitive attributes are also time-dependent. For example, the gender-gap in income level can change over time. As a result, a fair model might fail to maintain its fairness under the input-space distributional shifts or when the model is used on differently sourced tasks (Pooch et al., 2020). When the input distribution under-

goes temporal shifts, the embedding space may not be attribute-agnostic anymore. The naive solution of retraining the model after distributional shifts requires annotating new data points to build datasets representative of the new input distribution. This process, however, is time consuming and expensive for deep learning and is challenging when data annotation becomes a persistent tasks. As a result, it is highly desirable to develop algorithms that can preserve model fairness under distribution shifts. Unfortunately, this problem has been marginally explored in the AI literature.

The negative effect of distributional shifts on the performance of AI models is well-known and the problem of model adaptation has been investigated extensively in the unsupervised domain adaptation (UDA) literature (Tzeng et al., 2017; Daumé III, 2007). The goal in UDA is to train a model such that it can have a good generalization performance on a target domain, where only unannotated data is available. In the absence of label information, which makes supervised learning infeasible, the idea is to transfer knowledge from a related source domain, where annotated data is accessible. The classic assumption for UDA is to consider two domains with the same classes to model cross-domain relations. A primary group of UDA algorithms achieves this goal by matching the source and the target distributions in a shared embedding space (Redko et al., 2017) such that the embedding space is domain-agnostic. In other words, the two domains become indistinguishable in the shared embedding space. As a result, a source-trained classifier that receives its input from the embedding space will generalize well in the target domain, despite being trained solely using the source domain annotated data. To align the two distributions in such an embedding, data points from both domains are mapped into a shared feature space that is modeled as the output space of a deep neural encoder. The deep encoder is then trained to minimize the distance between the two distributions, measured in terms of a suitable probability distribution metric. Despite being effective , existing UDA algorithms overlook model fairness and solely consider improving model performance in the target domain. Hence, the model fairness maybe compromised when the model is updated. In this work, we adopt the idea of domain alignment in UDA to preserve model fairness and mitigate model biases introduced by domain shift (Rostami & Galstyan, 2023; Luo et al., 2019).

**Contribution:** We address the problem of preserving the model fairness and the model generalization under distributional shifts. While domain shift can be instantiated in various form, we only consider distribution shifts in the input space. We also consider the setting when only unannotated data is accessible after an initial training stage. We model this problem within the classic unsupervised domain adaptation paradigm. We consider the initial annotated training dataset as the source domain and the collected unannotated data during model execution as the target domain. Our specific contributions include:

- We develop an algorithm that minimizes distributional mismatches that results from domain shift in a shared embedding space to maintain model fairness and model performance in non-stationary learning settings.
- We build three tasks using three standard fairness benchmarks and demonstrate that in addition to model performance, model fairness is compromised when domain shift exists in real-world applications and existing methods can't address this challenge.

- We conduct extensive empirical explorations and demonstrate that the existing methods for fairness in AI are vulnerable in our learning setting and show that the proposed algorithm is effective in maintaining both the model performance and its fairness.

Our implementation code is available at: <https://github.com/rostami-m/FairUDA/>.

## 2. Related Work

Our work is related to works both in UDA and fairness in AI.

### 2.1 Fairness in AI

There are various approaches for training a fair model for a single domain when the data distribution is stationary. The primary idea in existing works that is the most relevant to our work is based on mapping data points into an embedding space at which the sensitive attributes are entirely removed from the representative features, i.e., an attribute-agnostic space. As a result, a classifier that receives its input from this space will make unbiased decisions due to the independence of its decisions from the sensitive attributes. After training the model, fairness can also be measured at the classifier output using a desired fairness metric. To train a model with an agnostic output-space, we can either benefit from distribution metric minimization (Jiang et al., 2020) or adversarial machine learning (Beutel et al., 2017). DANN (Ajakan et al., 2014) is a pioneer work to benefit from adversarial learning to address domain adaptation. Saito et al. (2018) utilize the task-specific decision boundaries in domain alignment to improve the model generalization in the target domain. Shu et al. (2018) improved upon DANN by using virtual adversarial learning and refinement of decision boundaries. Our work benefits from the idea of using adversarial learning to find a latent representation that is independent of the sensitive attributes to address fairness.

Jiang et al. (2020) develop a fairness algorithm that induces probabilistic independence between the sensitive attributes and the classifier outputs by minimizing the optimal transport distance between the probability distributions conditioned on the sensitive attributes. Hence, the transformed probability in the embedding space then becomes independent (unconditioned) from the sensitive attributes. Celis et al. (2019a) study the possibility of using a meta-algorithm for fairness with respect to several disjoint sensitive attributes. Du et al. (2021) have followed a different approach. Instead of training an encoder that removes the sensitive attributes in a latent embedding space and then training a classifier, they propose to debias the classifiers by leveraging samples with the same ground-truth label yet having different sensitive attributes. The idea is to discourage undesirable correlation between the sensitive attribute and predictions in an end-to-end scheme, allowing for the emergence of attribute-agnostic representations in the hidden layers of the model. Agarwal et al. (2018) propose an approach that incrementally constructs a fair classifier by solving several cost-constrained classification problems and combining results. Zhang et al. (2018) train a deep model to produce predictions independent of sensitive attributes by training a classifier network to predict binary outcomes and then inputting the predictions to an adversary that attempts to guess their sensitive attribute. By optimizing the network to make this task harder for the adversary, their approach leads to fair predictions.

Beutel et al. (2017) benefit from removing sensitive attributes to train fair models by indirectly enforcing decision independence from the sensitive attributes in a latent representation using adversarial learning. They also amend the encoder model with a decoder to form an autoencoder. Since the representations are learned such that they can self-reconstruct the input, they contain a similar amount of information in a lower dimension which can be used for classification purposes. These work consider stationary settings. Our work builds upon using adversarial learning to preserve fairness when distribution shifts exist. In order to combat domain shift, our idea is to additionally match the target data distribution with the source data distribution in the latent embedding space, a process that ensures classifier generalization.

## 2.2 Unsupervised Domain Adaptation

UDA and zeros-shot learning (ZSL) are two learning settings for dealing with scarcity of labels in ML using auxiliary domains. While UDA and ZSL assumes access to unlabeled data from the target domain, ZSL enables predicting novel categories without any target domain data (Romera-Paredes & Torr, 2015; Xian et al., 2017; Rostami et al., 2022). ZSL enables a model to make predictions for unseen classes or tasks without any training examples, typically relying on auxiliary information such as high-level semantic attributes (Rostami et al., 2020). In contrast, UDA focuses on transferring knowledge from a labeled source domain to an unlabeled target domain with a similar task but different distribution using domain alignment. Works on domain alignment for UDA follow a diverse set of strategies. The goal of existing works in UDA is solely to improve the prediction accuracy in the target domain in the presence of domain shift without exploring the problem of fairness. A large groups of UDA methods rely on adversarial learning (Goodfellow et al., 2014) for indirect domain alignment (Ajakan et al., 2014; HassanPour Zonoozi & Seydi, 2023; Jian & Rostami, 2023; Mei et al., 2023). The closer line of research to our work addresses domain shift by minimizing a probability discrepancy measure between two distributions in a shared embedding space. Selection of the discrepancy measure is a critical task for these works. Several UDA methods simply match the low-order empirical statistics of the source and the target distributions as a surrogate for the distributions. For example, the Maximum Mean Discrepancy (MMD) metric is defined to match the means of two distributions for UDA (Long et al., 2015, 2017). Correlation alignment is another approach to include second-order moments (Sun & Saenko, 2016). Matching lower-order statistical moments overlooks the existence of discrepancies in higher-order statistical moments. In order to improve upon these methods, a suitable probability distance metric can be incorporated into UDA to consider higher-order statistics for domain alignment. A suitable metric for this purpose is the Wasserstein distance (WD) or the optimal transport metric (Courty et al., 2016; Bhushan Damodaran et al., 2018; Rostami, 2024b; Rostami et al., 2019). Since WD possesses non-vanishing gradients for two distributions with non-overlapping supports, it is a more suitable choice for deep learning than more common distribution discrepancy measures, e.g., KL-divergence. Optimal transport can be minimized as an objective using first-order optimization algorithms for deep learning. Using WD has led to a considerable performance boost in UDA (Bhushan Damodaran et al., 2018; Nananukul et al., 2024; Stan & Rostami, 2024b) compared to methods that rely on aligning the lower-order statistical

moments (Long et al., 2015; Sun & Saenko, 2016). Our work uses the sliced Wasserstein distance (SWD) (Lee et al., 2019; Rostami, 2024a; Stan & Rostami, 2024a) for domain alignment. SWD possesses the nice properties of WD, but it has a closed-form definition which makes its computation simpler.

### 2.3 Domain Adaptation in Fairness

Works on benefiting from knowledge transfer to maintain fairness are relatively limited. Madras et al. (2018a) benefit from adversarial learning to learn domain-agnostic transferable representations for fair model generalization. Their work does not extend to UDA with a target domain with unannotated data and the motivation is to improve generalization during training on a source domain. Interestingly, Mukherjee et al. (2022) show that by enforcing individual fairness during model training, the out-of-distribution generalization of the model can be improved under the covariate shift assumption. This assumption is limiting for UDA because often domain shift involves changes in the conditional distribution at the output space which is more challenging to address compared to covariate shift. Coston et al. (2019) consider a domain adaptation setting where the sensitive attributes for data points are accessible only in one of the source or the target domains. They also use covariate shift assumption. Their idea is to use a weighted average to compute the empirical risk and then tune the corresponding data point-specific weights to minimize covariate shifts. Their analysis also is limited to using logistic regression as the base model which is limiting for challenging tasks. Schumann et al. (2019) consider a setting, where they define the fairness distance of equalized odds, and then use it as a regularization term in addition to empirical risk, minimized for fair cross-domain generalization. Their approach is not benefiting from unlabeled target domain and mostly focuses on improving the model generalization. Hu et al. (2019) address fairness in a distributed learning setting, where the data exist in various servers with private demographic information. However, the data in these distributed servers is assumed to be labeled. Singh et al. (2021) consider that a causal graph for the source domain data and anticipated covariate shifts are given. They then use feature selection to estimate the fairness metric in the target domain for model adaptation. Zhang and Long (2021) explore the possibility of training fair models in the presence of missing data in a target domain using a source domain with complete data and find theoretical bounds for this purpose. Our learning setting is relevant yet different from the above settings. We consider a standard UDA setting that goes beyond addressing covariate shift. To address UDA, we also consider that the sensitive attributes are accessible in both domains. This constraint would limit our learning setting, but it is crucial to note that fair AI primarily concerns applications involving humans as data points, where sensitive attributes such as sex, gender, and age are typically accessible because these attributes can be easily collected without the need for additional data annotation, making them readily available for analysis. The challenge is to adapt the model to preserve fairness in the target domain without requiring data annotation when domain shift occurs.

## 3. Fair Model Training in Stationary Settings

We first describe how we can train a fair model, then explain how the problem extends to a non-stationary setting, and offer our solution in the next section.

Consider a source domain  $\mathcal{S}$ , where we are given an annotated training dataset  $\mathcal{D}^s = (X^s, A^s, Y^s) \in \mathbb{R}^{N \times d} \times \{0, 1\}^N \times \{0, 1\}^N$  for which  $X^s \in \mathbb{R}^n$  represents feature vectors with dimension  $d$  and  $Y^s$  represents the binary labels. Additionally,  $A^s$  represents a binary sensitive attribute for each data point, e.g., race, sex, age, etc. Each triplet  $(\mathbf{x}^s, \mathbf{a}^s, \mathbf{y}^s)$  is drawn from the source domain distribution  $P_{\mathcal{S}}(\mathbf{X}, \mathbf{A})$ , where the feature vector corresponds to characteristic features that are used for decision-making, e.g., occupation length, education years, credit history, etc. Additionally, we consider that each data point is annotated with sensitive attributes, e.g., sex, race, etc. Our goal is to train a fair model with respect to the sensitive attributes to perform binary decision making, e.g., approving for a loan, parole in prison system, etc, without any bias against one of the subgroups.

In classic parametric supervised learning, we select a family of predictive functions  $f_{\theta} : (X^s, A^s) \rightarrow Y^s$ , parameterized with learnable parameters  $\theta$ . We then search for the model with the optimal parameter based on ERM on the fully annotated dataset  $\mathcal{D}^s$ , as a surrogate for a model with the expected error on the unknown source domain distribution:

$$\hat{\theta} = \arg \min_{\theta} \mathcal{L}_{sl} = \arg \min_{\theta} \left\{ \frac{1}{N} \sum_{i=1}^N \mathcal{L}_{bce}(f_{\theta}(\mathbf{x}^s, \mathbf{a}^s), \mathbf{y}^s) \right\}, \quad (1)$$

where  $\mathcal{L}_{bce}$  is a suitable loss function such a binary cross-entropy loss. Under certain conditions, solving (1) leads to training a generalizable model during the testing stage. However, there is no guarantee to obtain a fair model because it is only optimized with respect to the prediction (1). Inherent bias in the training dataset, e.g., over/under-representation of subgroups, can lead to training a biased model. Note that even if the sensitive attributes are not used in (1), the sensitive attribute may still be highly correlated with the decision features due to data collection procedures. For example, a human operator might have subconsciously consider a sensitive attribute for annotation or a sensitive attribute maybe correlated with some elements of the feature vector.

As we discussed, an effective approach to train a fair model is to map the domain data into a latent embedding space such that the encoded data representations are fully independent from the sensitive attributes  $A$ . There are various approaches to implement this idea via training an appropriate encoding function. Inspired by adversarial learning, a group of fairness algorithms rely on solving a min-max optimization problem for this purpose (Beutel et al., 2017; Madras et al., 2018b). To this end, we first consider that the end-to-end binary predictive model  $f_{\theta}(\cdot) : \mathbb{R}^d \rightarrow \{0, 1\}$  can be decomposed into an encoder subnetwork  $e_u(\cdot) : \mathbb{R}^d \rightarrow \mathbb{R}^z$ , with learnable parameters  $u$ , followed by a classifier subnetwork  $g_v(\cdot) : \mathbb{R}^z \rightarrow \mathbb{R}^2$  with learnable parameters  $v$ , where  $f_{\theta}(\cdot) = (g_v \circ e_u)(\cdot)$  and  $\theta = (u, v)$ . The parameter  $z$  denotes the dimension of the latent embedding space that we want to be sensitive-agnostic which is modeled as the output space of the encoder subnetwork. To benefit from an adversarial machine learning scheme and induce “independence from the sensitive attribute” in the latent space, we consider an additional classification network  $h_w(\cdot) : \mathbb{R}^z \rightarrow \mathbb{R}^2$  with learnable parameters  $w$ . This classifier is tasked to predict the corresponding sensitive attribute  $\mathbf{a}^s$  using the latent space representations  $e_u(\mathbf{x}^s, \mathbf{a}^s)$ . If the second classifier fails to predict the sensitive attribute for a given data point, one can argue that data representations in the embedding space are independent from the attributes.

The core idea is to induce “probabilistic independence from sensitive attributes” by training  $e_u(\cdot)$  and  $h_w(\cdot)$  in an adversarial learning scheme, where  $e_u(\cdot)$  plays the role of

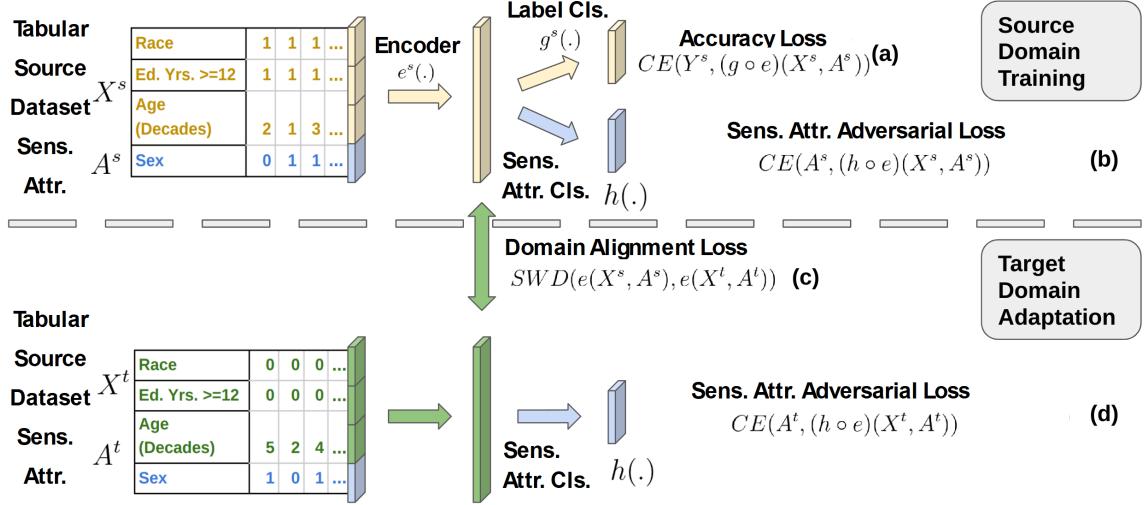


Figure 1: Block-diagram description of the proposed framework for preserving fairness under domain shift. First, a fair model is trained on a source domain using a fully labeled training dataset (in the figure, “CLS”, “Attr.”, and “Sens.” stand for “Classifier”, “Attribute”, and “Sensitive” respectively.): (a) minimizing binary cross entropy loss against the source labels (Eq. (1)) ensures the learnt embeddings are informative with respect to the classification task (b) adversarial optimization with respect to the sensitive attribute (Eq. (2)) makes the learnt embeddings conditionally independent from the sensitive attributes. During adaptation on the unlabeled target domain: (c) Sliced Wasserstein Distance is minimized between the target embedding distribution and the source embedding distribution (Eq. (4)) in order to maintain the relevance of the source classifier on the target domain, (d) the fairness loss is also minimized on the target domain to ensure conditional independence of the embeddings and sensitive attributes.

the generator network and  $h_w(\cdot)$  is the discriminator network. In other words, the goal is to train the models such that the classifier  $h(\cdot)$  is trained to perform poorly. To this end, consider the following loss function for predicting the sensitive attributes:

$$\mathcal{L}_{fair}^s = \mathcal{L}_{bce}((h_w \circ e_u)(\mathbf{x}^s, \mathbf{a}^s), \mathbf{a}^s). \quad (2)$$

To train an attribute-agnostic encoder, we solve the following alternating min-max optimization process to train a fair model based on adversarial learning scheme (Goodfellow et al., 2014):

1. We fix the encoder  $e_u(\cdot)$  and minimize the fairness loss  $\mathcal{L}_{fair}$  through updating the attribute classifier  $h_w(\cdot)$ .
2. We then fix the attribute classifier  $h_w(\cdot)$  and maximize the fairness loss  $\mathcal{L}_{fair}$  by updating the encoder  $e_u(\cdot)$ .

The first step will perform ERM for the attribute prediction classifier, conditioned on the encoder network being fixed. The second step will keep the classifier fixed and ensures that the latent data representations are as little informative as possible about the sensitive attribute  $A$ . Similar to vanilla adversarial learning, empirical explorations demonstrate that the above iterative alternations between the two optimization steps will lead to training an encoder that produces latent representations that are independent from the sensitive attribute when the attribute classifier fails to predict the sensitive attributes. To train a fair and generalizable model, we combine equations (1) and (2) and solve:

$$\hat{u}, \hat{w}, \hat{v} = \arg \min_{u,w,v} \mathcal{L}_{sl} + \mathcal{L}_{fair}^s, \quad (3)$$

to learn extracting features that are discriminative for performing the original classification task via  $g_v(\cdot)$ . The high-level description of this procedure is presented in top portion of Figure 1.

#### 4. Fair Model Training in Non-Stationary Settings

The above approach would suffice in practice if we have a single source domain, i.e., the data distribution is stationary and the testing data points are drawn from the source domain distribution. In our formulation, we consider that the test data is drawn from a second target domain  $\mathcal{T}$  with a different data distribution  $P_{\mathcal{T}}(\mathbf{X}, \mathbf{A}) \neq P_{\mathcal{S}}(\mathbf{X}, \mathbf{A})$ . The change in the input distribution can also violate the covariate shift assumption, i.e.,  $P_{y\mathcal{T}}(y|\mathbf{X}, \mathbf{A}) \neq P_{\mathcal{S}}(y|\mathbf{X}, \mathbf{A})$ , which makes techniques such reweighting or resampling inapplicable. The target domain may be result of drifts in the input space or can occur when we want to use the model in a different domain. We also assume that we only have access to the unannotated dataset  $\mathcal{D}^t = (\mathbf{X}^t, \mathbf{A}^t)$  in the target domain. Due to the distribution gap between the two domains, the source-trained model performance would degrade in the target domain. Hence, we need to update the model to remain fair in the target domain which would require annotating  $\mathcal{D}^t$ . Our goal is to make this process more practical by relaxing the need for data annotation.

To address fairness under domain shift, we formulate this problem in a UDA setting. UDA tackles the challenge of performance degradation under domain shift. The core idea in UDA is to improve generalization on the target domain via updating the encoder network such that the empirical distance between the distributions  $e_u(P_{\mathcal{S}}(\mathbf{X}, \mathbf{A}))$  and  $e_u(P_{\mathcal{T}}(\mathbf{X}, \mathbf{A}))$  is minimized, i.e., the two distributions are aligned such that the embedding space becomes domain agnostic. Under this restriction and when extrapolation on the target domain is not required, the classifier  $g_v(\cdot)$  that is trained on the source domain will generalize on the target domain. While this idea has been explored extensively in the UDA literature, it is insufficient to guarantee fairness after the adaptation phase because UDA approaches do not consider fairness metrics. Our goal is to extend UDA to preserve model fairness in the target domain in addition to maintaining model generalization.

##### 4.1 Proposed Algorithm

While adversarial learning has been used extensively to address UDA similar to training a fair model, extensions based on solely adversarial learning would be challenging. The

reason is that we would have two coupled adversarial learning problems and maintaining the optimization stability would be difficult. In our approach we still use adversarial learning to preserve fairness but we benefit from metric learning to maintain model generalization (Lee et al., 2019; Redko et al., 2017). The block-diagram description of our proposed approach is presented in Figure 1. We follow a two phase process. Initially, we train a fair model on the source domain dataset  $(X^s, A^s, Y^s)$  and then update it to work well on the target domain. To train a fair mode, we use the following three steps iteratively to solve (3):

1. We optimize the classifier  $f_\theta(\cdot) = (g_v \circ e_u)(\cdot)$  network in an end-to-end scheme by minimizing (1) using the annotated data. This process will generate informative and discriminative latent features for decision making.
2. We then fix the feature extractor encoder  $e_u(\cdot)$  and optimize the sensitive attribute classifier  $h_w(\cdot)$  by minimizing the loss in (2). This step will enforce the sensitive attribute classifier to extract information from the representations in the embedding space that can be used for predicting the sensitive attribute  $A$ .
3. We freeze the sensitive attribute classifier  $h_w(\cdot)$  and update the encoder subnetwork  $e_u(\cdot)$  in order to maximize the fairness loss function in (2). This step will force the encoder to produce representations that are independent from the sensitive attribute  $A$  to enforce the fairness constraint on the extracted features.

The above steps lead to training a model which is both fair and generalizable.

In the second phase, we update the model to remain fair and generalizable when used on the target domain. We first explain the classic UDA approach.

---

**Algorithm 1** FairAdapt (*thresh, ITR*)

---

```

for itr = 1, ..., ITR do
    Source Training:
        Optimize  $\mathcal{L}_{bce}$  via (1).
        Optimize  $\mathcal{L}_{fair}$  via (2) and freezing  $u$ .
        Optimize  $-\mathcal{L}_{fair}$  via (2) and freezing  $h$ .
    if itr > thresh then
        Target Adaptation:
            Optimize  $\mathcal{L}_{swd}$  via (4).
            Optimize  $\mathcal{L}_{fair}$  via (2) and freezing  $u$ .
            Optimize  $-\mathcal{L}_{fair}$  via (2) and freezing  $h$ .
    end if
end for
return  $u, g$ 

```

---

The classic adaptation process relies only on aligning the two distributions in the embedding space, i.e.,  $e(P_S(\mathbf{X}, \mathbf{A})) \approx e(P_T(\mathbf{X}, \mathbf{A}))$ . We follow metric minimization to enforce domain alignment (Lee et al., 2019; Redko et al., 2017). The idea is to select a suitable probability distribution distance  $d(\cdot, \cdot)$  and minimize it as a loss function at the encoder output, i.e.  $d(e(P_S(\mathbf{X}, \mathbf{A})), e(P_T(\mathbf{X}, \mathbf{A})))$ . As a result, the encoder is trained to guarantee

domain-agnostic embedding features at its output. Compared to using adversarial learning, this approach requires less hyperparameter tuning and the resulting optimization problem is more stable.

The choice of the distribution distance  $d(\cdot, \cdot)$  is a design choice and various metric have been used for domain alignment. In our work, we use the Sliced Wasserstein Distance (SWD) (Redko et al., 2017) for this purpose. SWD is defined based on optimal transport or the Wasserstein Distance (WD) metric to broaden its applicability in deep learning. The upside of using WD is that it has non-zero gradients even when the support for two distributions are non-overlapping which in turn makes it a suitable loss function for gradient descent. WD has been used successfully to address UDA but the downside of using WD is that it is defined in terms of an optimization problem. As a result, minimizing WD directly is a challenging task because often we need to solve another optimization problem to compute WD. The idea behind defining SWD is to develop a metric with closed-form solution by slicing two high-dimensional distributions to generate 1D projected distributions. Since WD has a closed-form solution for 1D distributions, SWD between the two high-dimensional distributions is computed as the average of these 1D WD slices (Wu et al., 2020; Rostami et al., 2023). In addition to having a closed-form solution, SWD can be computed using empirical samples from the two distributions as:

$$\mathcal{L}_{swd} = \frac{1}{K} \sum_{i=1}^K WD^1(\langle e(\mathbf{x}^s, \mathbf{a}^s), \gamma_i \rangle, \langle e(\mathbf{x}^t, \mathbf{a}^t), \gamma_i \rangle), \quad (4)$$

where,  $WD^1(\cdot, \cdot)$  denotes the 1D WD distance,  $K$  is the number of random 1D projections we are averaging over and  $\gamma_i$  is one such projection direction. We use random projections to estimate averaging over all possible projections. We can then solve the following problem to maintain model generalization on the source domain:

$$\mathcal{L}_{sl} + \mathcal{L}_{swd}. \quad (5)$$

If we only align the two distributions using (5), the model fairness can be compromised because when the encoder is updated to maintain model generalization, there is no guarantee that the embedding space remains independent from the sensitive attributes. Hence, the model can become biased. To preserve fairness in the target domain under distributional shifts, we augment the iterative steps (1) – (3) described above with the following two steps:

4. We minimize the empirical SWD distance between  $e(P_S(\mathbf{X}, \mathbf{A}))$  and  $e(P_T(\mathbf{X}, \mathbf{A}))$  via (4). This step ensures the source-trained classifier  $g(\cdot)$  will generalize on the target domain samples from  $e(P_T(\mathbf{X}, \mathbf{A}))$ .
5. We repeat steps (2) and (3) using solely the sensitive attributes of the target domain.

The additional steps will update the model on the target domain to preserve both fairness and generalization accuracy. Following steps (1)-(5), the total loss function that we minimize would become:

$$\mathcal{L}_{bce}(\hat{y}, y_s) + \mathcal{L}_{fair}^s + \mathcal{L}_{fair}^t + \mathcal{L}_{swd}, \quad (6)$$

Algorithm 1 summarizes the above described training process for our proposed algorithm, named FairAdapt.

## 5. Empirical Validation

We adopt existing common datasets in the AI fairness literature and tailor them for our formulation. Our implementation code is available as a supplement.

### 5.1 Experimental Setup

Our implementation code is available at: <https://github.com/rostami-m/FairUDA/>. For details beyond what we describe in this section, please refer to the implementation. We first describe our empirical exploration setting.

#### 5.1.1 DATASETS AND TASKS

Common datasets in the fairness literature pose binary decision-making problems, e.g., approval of a credit application, alongside relevant features used for decision-making by professionals, e.g., employment history, credit history etc., and group-related sensitive attributes, e.g., sex, race, nationality, etc. Based on sensitive group membership, data points can be part of privileged or unprivileged subgroups. For example, with respect to sex, men are part of the privileged group while women are part of the unprivileged group. Choice of privileged group is often dependant on the specific prediction task. When learning a model that aims to predict employability, the privileged group may consist of young applicants, however if the same feature set is used to predict mortgage application success, older age may be considered as the privileged group. We perform experiments on three datasets widely used by the AI fairness community. We consider *sex* as our sensitive attribute because it is recorded for all three datasets. These datasets are:

The **UCI Adult dataset**<sup>1</sup> is part of the UCI database (Dua & Graff, 2017) and consists of 1994 US Census data. The task associated with the dataset is predicting whether annual income exceeds 50k. After data cleaning, the dataset consists of more than 48,000 entries. Possible sensitive attributes for this dataset include *sex* and *race*.

The **UCI German credit dataset**<sup>2</sup> contains financial information for 1000 different people applying for credit and is also part of the UCI database. The predictive task involves categorizing individuals as acceptable or non-acceptable credit risks. *Sex* and *age* are possible sensitive attributes for the German dataset.

The **Correctional Offender Management Profiling for Alternative Sanctions (COMPAS) recidivism dataset**<sup>3</sup> maintains information of over 5,000 individuals' criminal records. Models trained on this dataset attempt to predict people's two year risk of recidivism. For the COMPAS dataset, *sex* and *race* may be used as sensitive attributes.

#### 5.1.2 EVALUATION PROTOCOL

Experiments on these datasets have primarily considered random 70/30 splits for the training and test splits. While such data splits are useful in evaluating overfitting for fairness algorithms, features for training and test sets will be sampled from the same data distribution. As a result, randomly splitting the datasets is not suitable for our learning setting

---

<sup>1</sup><https://archive.ics.uci.edu/ml/datasets/Adult>

<sup>2</sup>[https://archive.ics.uci.edu/ml/datasets/statlog+\(German+credit+data\)](https://archive.ics.uci.edu/ml/datasets/statlog+(German+credit+data))

<sup>3</sup><https://github.com/propublica/COMPAS-analysis/>

because domain shift will not exist between the training and the testing splits. Instead, we consider natural data splits obtained from sub-sampling the three datasets along different criteria to generate the training and testing splits. We show that compared to random splits, where learning a model that guarantees fairness on the source domain is often enough to guarantee fairness on the target domain predictions, domain discrepancy between the source and target domains can lead to biased or degenerate predictions on the target domain, even if the model is initially trained to be fair. For details about the splits for each dataset, please refer to Appendix A. In short, these splits introduce domain gap between the testing and training splits to generate appropriate tasks for our setting.

Next, for each of the three datasets, we will generate source/target data splits where ignoring domain discrepancy between the source and target can negatively impact model fairness. Per dataset, we produce three such splits. We characterize the label distributions and sensitive attribute conditional distributions for the Adult dataset in Table 1. We provide similar analysis for the German and COMPAS datasets in the supplementary material.

**Adult Dataset.** We use age, education and race to generate the source and target domains. These domains can be a natural occurrence in practice, as gathered census information may differ along these axes geographically. For example, urban population is on average more educated than rural population <sup>4</sup>, and more ethnically diverse <sup>5</sup>. Thus, a fair model trained on one of the two populations will need to overcome distribution shift when evaluated on the other population. The source/target splits we consider are as follows:

1. **Source Domain:** White, +12 education years. **Target Domain:** Non-white, Less than 12 education years.
2. **Source Domain:** White, Older than 30. **Target Domain:** Non-white, younger than 40.
3. **Source Domain:** Younger than 70, +12 education years. **Target Domain:** Older than 70, less than 12 education years.

In Table 1, we analyze the conditional distributions of the labels and sensitive attribute for the above data splits. For the random split (A), we see that the conditional distributions of the sensitive attributes are identical in both domains which is expected due to absence of domain shift. For the three splits that we generated, we observe all three distributions:  $P(Y)$ ,  $P(A|Y = 0)$ ,  $P(A|Y = 1)$  differ between the source and the target domains. We also note the label distribution becomes more skewed towards  $Y = 0$ . Table 1 suggests that common UDA methods would fail to preserve fairness.

### 5.1.3 FAIRNESS METRICS

Because fairness does not have a universally accepted definition, there exist a multitude of criteria developed for evaluating algorithmic fairness (Mehrabi et al., 2021). The goal is to define fairness intuitively and then come up with a computable quantitative metric based on a notion of independence. In the context of datasets presenting a privileged and

---

<sup>4</sup><https://www.ers.usda.gov/topics/rural-economy-population/employment-education/rural-education/>

<sup>5</sup><https://www.ers.usda.gov/data-products/chart-gallery/gallery/chart-detail/?chartId=99538>

Split	Source				Target			
	Size	Y=0	A=0 Y=0	A=0 Y=1	Size	Y=0	A=0 Y=0	A=0 Y=1
A	34120	0.76	0.39	0.15	14722	0.76	0.39	0.15
A1	12024	0.53	0.41	0.16	5393	0.91	0.49	0.18
A2	29466	0.66	0.34	0.14	2219	0.97	0.48	0.30
A3	11887	0.52	0.42	0.16	778	0.89	0.39	0.17

Table 1: Data split statistics corresponding to the Adult dataset: the row with no number, i.e., “A”, corresponds to random data splits. The numbered rows, i.e., A1,A2,A3 correspond to statistics for specific splits that we prepared. The columns represent the probabilities of specific outcomes for specific splits, e.g.,  $P(Y = 0)$ , when using *sex* as sensitive attribute.

unprivileged group, these metrics rely on ensuring predictive parity between the two groups under different constraints.

The most common fairness metric employed is demographic parity (DP)  $P(\hat{Y} = 1|A = 0) = P(\hat{Y} = 1|A = 1)$ , which is optimized when predicted label probability is identical across the two groups. However, DP only ensures similar representation between the two groups, while ignoring actual label distribution. Equal opportunity (EO) (Hardt et al., 2016) conditions the fairness value on the true label  $Y$ , and is optimized when  $P(\hat{Y} = 1|A = 0, Y = 1) = P(\hat{Y} = 1|A = 1, Y = 1)$ . EO is preferred when the label distribution is different across privilege classes, i.e.,  $P(Y|A = 0) \neq P(Y|A = 1)$ . A more constrained fairness metric is averaged odds (AO), which is minimized when outcomes are the same conditioned on both labels and sensitive attributes, i.e.,  $P(\hat{Y}|A = 0, Y = y) = P(\hat{Y}|A = 1, Y = y), y \in \{0, 1\}$ . EO is a less constrained case of AO, for the case where  $y = 1$ . Following the AI fairness literature, we report the “left hand side and right hand side difference  $\Delta$ ” for each of the above measures. Under this format,  $\Delta$  values that are close to 0 will signify that the model maintains fairness, while values close to 1 signify a lack of fairness. Tuning a model to optimize fairness may incur accuracy trade offs (Madras et al., 2018a; Kleinberg et al., 2016; Wick et al., 2019). For example, a classifier which predicts every element to be part of the same group, e.g.,  $P(\hat{Y} = 0) = 1$  will obtain  $\Delta EO = \Delta AO = 0$ , without providing informative predictions. Our approach has the advantage that the regularizers of the three employed losses  $\mathcal{L}_{CE}, \mathcal{L}_{fair}, \mathcal{L}_{swd}$  can be tuned in accordance with the importance of accuracy against fairness for a specific task.

#### 5.1.4 BASELINE METHODS FOR COMPARISON

To the best of our knowledge, no prior method has exactly addressed our learning setting. For this reason, we evaluate our work against four popular fairness preserving algorithms: Meta-Algorithm for Fair Classification (MC) (Celis et al., 2019b), Adversarial Debiasing (AD) (Zhang et al., 2018), Reject Option Classification (Kamiran et al., 2012), and Exponentiated Gradient Reduction (Agarwal et al., 2018). Comparison against these methods reveals the weakness of existing methods and strength of our method. The shortcoming of these methods provides the motivation behind developing our algorithm. We additionally report as baseline (Base) version where we only minimize  $\mathcal{L}_{bce}$  without optimizing fairness or

minimizing distributional distance. This baseline corresponds to the performance of a naive source-trained classifier and serves as a lowerbound to show the amount of performance boost we obtain.

### 5.1.5 TRAINING AND MODEL SELECTION

Implementation of our approach is done using the PyTorch (Paszke et al., 2019) deep learning library. We model our encoder  $e_u$  as a one layer neural network with output space  $z \in \mathbb{R}^{20}$ . Classifiers  $g$  and  $h$  are also one layer networks with output space  $\in \mathbb{R}^2$ . We train our model for 45,000 iterations, where the first 30,000 iterations only involve source training. For the first 15,000 we only perform minimization of the binary cross entropy loss  $\mathcal{L}_{bce}$ . We introduce source fairness training at iteration 15,000, and train the fair model, i.e. with respect to both  $\mathcal{L}_{bce}$  and  $\mathcal{L}_{fair}$ , for 15,000 more iterations. In the last 15,000 iterations we perform adaptation, where we optimize  $\mathcal{L}_{bce}, \mathcal{L}_{fair}$  on the source domain,  $\mathcal{L}_{fair}$  on the target domain, and  $\mathcal{L}_{swd}$  between the source and target embeddings  $e_u((x^s, a^s)), e_u((x^t, a^t))$  respectively. We use a learning rate for  $\mathcal{L}_{bce}, \mathcal{L}_{fair}$  of  $1e - 4$ , and learning rate for  $\mathcal{L}_{swd}$  of  $1e - 5$ . Model selection is done by considering the difference between accuracy on the validation set, and demographic parity on the test set. Given equalized odds and averaged opportunity require access to the underlying labels on the test set we cannot use these metrics for model selection. Additionally, models corresponding to degenerate predictions i.e. test set predicted labels being either all 0s or all 1s are not included in result reporting.

## 5.2 Comparison Results

We report balanced accuracy (Acc.), demographic parity ( $\Delta DP$ ), equalized odds ( $\Delta EO$ ) and averaged opportunity ( $\Delta AO$ ) in our comparison results to study both accuracy and fairness. Desirable accuracy values are close to 1, while desirable fairness metric values should be close to 0. Prior studies have shown that there is a trad-off between the performance accuracy and the model fairness because the fairness and the ERM loss can potentially work against each other. Results are averaged over 7 runs to make comparisons statistically meaningful. We use  $sex$  as the  $A$  because it is a shared attribute across all datasets.

Alg.	Adult				German				COMPAS			
	Acc.	$\Delta DP$	$\Delta EO$	$\Delta AO$	Acc.	$\Delta DP$	$\Delta EO$	$\Delta AO$	Acc.	$\Delta DP$	$\Delta EO$	$\Delta AO$
Base	0.74	0.35	0.30	0.28	0.64	0.03	0.16	0.05	0.68	0.22	0.27	0.18
MC	0.71	0.13	0.09	0.08	0.63	0.22	0.18	0.20	0.65	0.22	0.22	0.20
AD	0.67	0.11	0.13	0.08	0.53	0.35	0.46	0.38	0.62	0.28	0.25	0.29
ROC	0.71	0.05	0.01	0.01	0.55	0.11	0.04	0.09	0.52	0.02	0.03	0.02
EGR	0.65	0.06	0.02	0.01	0.51	0.01	0.04	0.02	0.63	0.02	0.02	0.02
Ours	0.70	0.00	0.07	0.08	0.64	0.00	0.05	0.01	0.65	0.00	0.02	0.03

Table 2: Results for random data splits.

As a sanity check experiment, we first report performance results for the standard random splits that are commonly used in the fairness literature in Table 2. Since for the standard splits, the source and the target are sampled from the same distribution, there is no domain shift. We observe in Table 2 that the baseline approach obtains the highest

accuracy across all datasets, but does not lead to fair predictions according to the three fairness metrics. The rest of method preserve fairness significantly better than the baseline but their performance accuracies are less than the baseline. This observation aligns well with what has been reported in the fairness literature. Importantly, our method leads to the best accuracy performance compared to the methods that maintain fairness. Our method leads to the minimum demographic parity which indicates that the embedding space is fully independent from the sensitive attributes. We also see that our method matches the best averaged opportunity on the German dataset, and best equalized odds on the COMPAS dataset, despite the fact that our method is not directly minimizing these metrics. These observations are critical because as it can be seen from Table 2, methods that maintain fairness, pay a cost in terms of performance accuracy. But our method is more robust in this aspect. We conclude that our algorithm successfully learns a competitively fair model when domain shift does not exist while leading to the best performance accuracy.

We then present results for the three data splits for each of the considered datasets that we proposed to study the effect of domain shift. These are custom splits for each dataset such that domain shift exists during the testing phase by design.

**Adult dataset** We report results on the three splits of the Adult dataset in Table 3. On the first split, MC obtains the highest accuracy of 0.68, and AD obtains accuracy of 0.63 which is higher accuracy than our method. However, none of the methods maintains fairness better than our method, as can be seen by the large  $\Delta DP, \Delta EO, \Delta AO$  values. On the remaining splits, our method is able to maintain fairness after adaptation while being competitive in terms of accuracy performance. We conclude that existing fairness-preserving methods struggle with domain shift between the source and target, while our method is positioned to overcome the challenge of domain shift.

Alg.	Race, Education				Race, Age				Age, Education			
	Acc.	$\Delta DP$	$\Delta EO$	$\Delta AO$	Acc.	$\Delta DP$	$\Delta EO$	$\Delta AO$	Acc.	$\Delta DP$	$\Delta EO$	$\Delta AO$
Base	0.63	0.34	0.53	0.42	0.60	0.24	0.25	0.24	0.59	0.90	0.92	0.91
MC	0.68	0.28	0.32	0.28	0.63	0.05	0.26	0.15	0.63	1.00	1.00	1.00
AD	0.63	0.21	0.33	0.25	0.60	0.25	0.25	0.25	0.51	0.16	0.15	0.16
ROC	0.59	0.34	0.25	0.31	0.62	0.02	0.20	0.11	0.50	0.00	0.00	0.00
EGR	0.62	0.06	0.16	0.10	0.59	0.02	0.19	0.11	0.56	0.43	0.40	0.42
Ours	0.62	0.01	0.05	0.01	0.62	0.00	0.19	0.10	0.52	0.01	0.06	0.03

Table 3: Performance results for the three splits of the Adult dataset

**COMPAS dataset** results for the COMPAS dataset are reported in Table 4. We first note that performances on this dataset are close to random label assignment. The reason is that after creating the biased datasets, they become small in size and biased. Hence, the performance of all models, including our, degrades on this dataset. Nevertheless, this dataset still can be used to compare the methods against each other in terms of fairness. On the first data split, MC again achieves the best accuracy. However, none of the methods we compare against besides EGR are able to preserve fairness. We are able to obtain higher accuracy than EGR while also obtaining improved fairness scores. On the second data split, our method is able to achieve the highest accuracy and also the lowest fairness scores amongst the methods. EGR, AD, and Base are not able to maintain fairness, while

MC and ROC provide degenerate results because their performances is similar to random assignment. On the third data split, our method achieves the best results when both the accuracy and fairness are considered together. We can see that AD, ROC, and EGR lead to degenerate models that works similar to random label assignment. Note that a fair model is not helpful if it assigns labels randomly. We conclude that our method works effectively for this dataset and is able to maintain both accuracy and fairness under domain shift.

Alg.	Age, Priors				Race, Age, Priors				Age, Priors, Charge			
	Acc.	$\Delta DP$	$\Delta EO$	$\Delta AO$	Acc.	$\Delta DP$	$\Delta EO$	$\Delta AO$	Acc.	$\Delta DP$	$\Delta EO$	$\Delta AO$
Base	0.54	0.29	0.27	0.28	0.49	0.33	0.56	0.43	0.58	0.52	0.54	0.52
MC	0.58	0.33	0.36	0.33	0.50	0.00	0.00	0.00	0.53	0.53	0.49	0.52
AD	0.52	0.62	0.73	0.66	0.47	0.70	0.72	0.70	0.49	0.77	0.77	0.77
ROC	0.53	0.28	0.09	0.21	0.50	0.00	0.00	0.00	0.50	0.00	0.00	0.00
EGR	0.49	0.05	0.10	0.06	0.53	0.27	0.34	0.26	0.51	0.10	0.17	0.11
Ours	0.53	0.00	0.05	0.02	0.65	0.15	0.17	0.19	0.54	0.16	0.18	0.18

Table 4: Performance results for the three splits of the COMPAS dataset

**German dataset** in Table 5, we present the results on the German dataset. In the first data split, our approach once again demonstrates competitive performance in terms of accuracy while achieving the best fairness performance. This highlights the ability of our method to strike a balance between accuracy and fairness, making it a compelling choice for domain adaptation tasks. Moving on to the second data split, our method outperforms all other approaches by obtaining the highest accuracy and demonstrating the best fairness performance across all fairness metrics. Note that the solution of ROC is degenerate. On the last data split, our method achieves the best fairness performance while still maintaining a decent accuracy. This proves the robustness of our approach, even in challenging scenarios, where fairness is a critical concern. On a wholistic view, we observe that ROC yields good fairness scores but fails to provide informative predictions. On the other hand, MC, AD, an EGR do not adequately maintain fairness.

Alg.	Employment				Credit hist., Empl.				Credit hist., Empl.			
	Acc.	$\Delta DP$	$\Delta EO$	$\Delta AO$	Acc.	$\Delta DP$	$\Delta EO$	$\Delta AO$	Acc.	$\Delta DP$	$\Delta EO$	$\Delta AO$
Base	0.67	0.09	0.05	0.07	0.58	0.07	0.10	0.06	0.56	0.35	0.35	0.32
MC	0.67	0.06	0.12	0.03	0.56	0.15	0.34	0.22	0.55	0.30	0.34	0.30
AD	0.52	0.53	0.58	0.55	0.53	0.40	0.56	0.46	0.52	0.44	0.52	0.46
ROC	0.50	0.00	0.00	0.00	0.50	0.00	0.00	0.00	0.62	0.13	0.09	0.11
EGR	0.57	0.22	0.36	0.27	0.50	0.43	0.44	0.43	0.50	0.01	0.00	0.00
Ours	0.62	0.01	0.05	0.02	0.58	0.02	0.01	0.01	0.55	0.01	0.02	0.01

Table 5: Performance results for the three splits of the German dataset

From Tables 3–5, we conclude that existing algorithms for training fair models are vulnerable in our setting. FairAdapt is effective and well-suited for preserving model fairness and performance accuracy on tasks associated with domain shift. Its ability to achieve competitive accuracy while ensuring fairness makes it a promising choice for real-world applications where domain adaptation and fairness are crucial considerations.

### 5.3 Analytic and Ablative Experiments

To provide a more intuitive understanding of our method, we visualize the impact of domain shift by generating 2D embeddings of the source and target domain features in the shared embedding space. For this purpose, we employ the UMAP (McInnes et al., 2020) visualization tool, which helps us create meaningful visual representations that encode the geometry of high dimensions. The resulting visualizations are presented in Figure 2. We have compared the source and target features resulting from a random split of the Adult dataset (Figure 2 (a)) with our first custom split (Figure 2 (b)). Upon examining the visualization of the random split, we notice that the source and target samples exhibit a considerable degree of similarity. However, when using a custom split, we observe a substantial discrepancy between the two distributions, indicating the existence of distributional mismatch. This disparity can have a significant impact on the model’s ability to generalize effectively. Our numerical results align with this observation, indicating that in the presence of domain shift, maintaining both generalization and fairness becomes a challenging task.

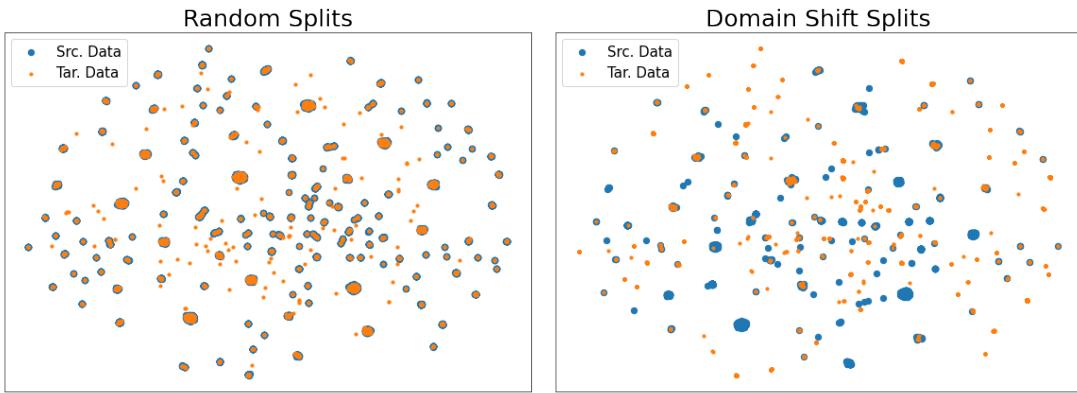


Figure 2: UMAP embeddings of the source and target feature spaces for random and custom splits of the Adult dataset

We additionally provide ablative experiments to investigate the impact of the different components of our approach on the downstream performance. We have compared the performances on the COMPAS dataset in Table 6 for four variants of our algorithm: (1) Base, similar to the main experiments, where no fairness or distributional minimization metric is used, (2) SWD, only the loss  $\mathcal{L}_{swd}$  is minimized (3) Fair, training is performed only with respect to  $\mathcal{L}_{fair}$  on the source and target domains (4) Our complete pipeline using both fairness and adaptation objectives. We can see that on all data splits, utilizing all losses leads to the best performance in terms of fairness. On the first split, the Fair only model is able to achieve competitive fairness results at the cost of accuracy. The SWD only approach achieves better accuracy but at the cost of fairness. Combining the two losses leads to improved accuracy over the Fair only model, and also improved fairness. Due to  $\mathcal{L}_{swd}$  being minimized at the encoder output space, both classifier and fairness head benefit from a shared source-target feature space. On the second split we observe the SWD only model has poorest performance, and the Fair Only and combined model have similar

fairness performance, with the combined model obtaining higher accuracy. This signifies the adversarial fair training process can act as a proxy task during training, improving model generalization. Finally, SWD achieves the best performance in the best performance but at the cost of fairness. We obtain better performance over Fair baseline while achieving similar fairness. We conclude from the ablative experiments that all components that we use in our method are crucial to obtain good performances.

In the previous experiments, we only considered *sex* as the sensitive attribute. We assess the performance of our proposed algorithm when using a different sensitive attribute. For this purpose, we utilize the German dataset and designate *age* as the sensitive attribute. The results of these experiments are presented in Table 7. Similar to our previous experiments where *sex* was chosen as the sensitive attribute, FairAdapt continues to exhibit outstanding performance by achieving the best demographic parity score among all the methods considered. Moreover, it outperforms other fairness-preserving approaches while maintaining accuracy values close to those of the Base model which does not maintain fairness. This observation demonstrates the robustness of our approach in terms of the choice of sensitive attribute. Regardless of the choice of the sensitive attribute, our algorithm consistently provides strong fairness results and competitive accuracy. This versatility indicates that our method can adapt to various fairness settings and has the potential to cover a wide range of domain adaptation tasks where fairness is a critical consideration.

Alg.	Age, Priors				Race, Age, Priors				Age, Priors, Chrg.			
	Acc.	$\Delta DP$	$\Delta EO$	$\Delta AO$	Acc.	$\Delta DP$	$\Delta EO$	$\Delta AO$	Acc.	$\Delta DP$	$\Delta EO$	$\Delta AO$
Base	0.68	0.22	0.27	0.18	0.54	0.29	0.27	0.28	0.58	0.52	0.54	0.52
SWD	0.56	0.29	0.38	0.32	0.45	0.44	0.33	0.40	0.59	0.64	0.69	0.64
Fair	0.50	0.01	0.08	0.04	0.64	0.15	0.17	0.19	0.52	0.17	0.19	0.20
Ours	0.53	0.00	0.05	0.02	0.65	0.15	0.17	0.19	0.54	0.16	0.18	0.18

Table 6: Ablative experiments using a subset of losses on the COMPAS dataset

Alg.	Empl.				Credit hist., Empl.				Credit hist., Empl.			
	Acc.	$\Delta DP$	$\Delta EO$	$\Delta AO$	Acc.	$\Delta DP$	$\Delta EO$	$\Delta AO$	Acc.	$\Delta DP$	$\Delta EO$	$\Delta AO$
Base	0.63	0.46	0.33	0.40	0.61	0.19	0.15	0.16	0.58	0.23	0.25	0.19
MC	0.59	0.32	0.29	0.30	0.67	0.35	0.18	0.27	0.61	0.12	0.18	0.11
AD	0.51	0.54	0.61	0.55	0.50	0.53	0.52	0.54	0.52	0.50	0.60	0.53
ROC	0.54	0.09	0.02	0.07	0.51	0.05	0.03	0.04	0.59	0.24	0.18	0.21
EGR	0.50	0.03	0.03	0.03	0.56	0.12	0.23	0.16	0.50	0.02	0.01	0.01
Ours	0.62	0.02	0.09	0.02	0.59	0.02	0.16	0.06	0.62	0.02	0.18	0.04

Table 7: Results on the German dataset when optimizing fairness metrics with respect to the *age* sensitive attribute

We performed another analytic experiment to study the effect of model training on the important loss terms and metric. In Figure 3, we analyze the effect of the adaptation process on target domain accuracy, validation accuracy, demographic parity on the source

domain, and demographic parity on the target domain for the Adult dataset. We compare two scenarios: (1) running the algorithm when  $\mathcal{L}_{swd}$  is not enforced (bottom), and (2) running the algorithm using both fairness and domain alignment (top). For the first 30,000 iterations, we only perform source-training, where the first half of iterations is spent optimizing  $\mathcal{L}_{bce}$ , and the second half is spent jointly optimizing  $\mathcal{L}_{bce}$  and the source fairness objective. We note once optimization with respect to  $\mathcal{L}_{fair}$  starts, demographic parity decreases until adaptation start, i.e., iterations 15,000 to 30,000. The validation accuracy in this interval also slightly decreases, as improving fairness may affect accuracy performance. During adaptation, i.e., after iteration 30,000, we observe that in the scenario where we use  $\mathcal{L}_{swd}$ , the target domain accuracy increases, while demographic parity on both the source and target domains remains relatively unchanged. In the scenario where no optimization of  $\mathcal{L}_{swd}$  is performed, there is still improvement with respect to target accuracy. However, target domain demographic parity becomes on average larger. These observations imply that the distributional alignment at the output of the encoder has beneficial effects both for the classification as well as the fairness objective and our algorithm gradually leads to the desired effects.

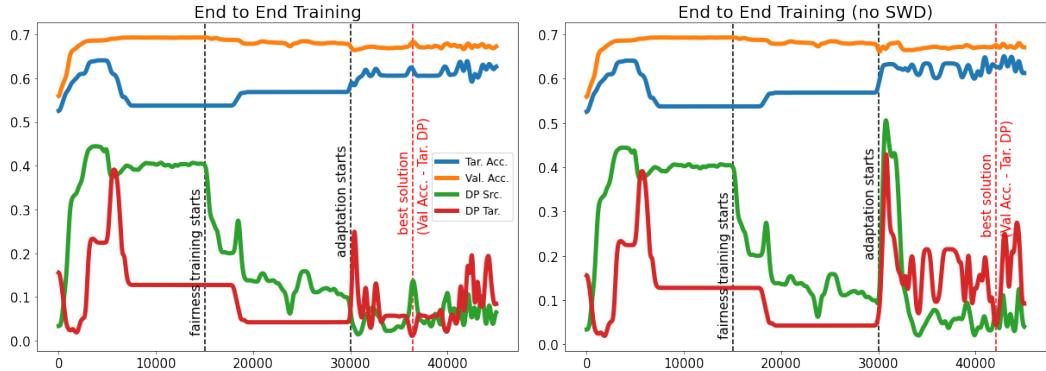


Figure 3: Learning behavior during training when using both  $\mathcal{L}_{fair}$  and  $\mathcal{L}_{swd}$  (top) versus when only using  $\mathcal{L}_{fair}$  (bottom)

We further investigate the different components present in our algorithm. In Figure 3, we analyze the training and adaptation process with respect to target accuracy, validation accuracy, demographic parity on the source domain, and demographic parity on the target domain. Performance plots are reported for the Adult dataset. We compare two scenarios: running the algorithm when  $\mathcal{L}_{swd}$  is not enforced (bottom), and running the algorithm using both fairness and domain alignment (top). For the first 30,000 iterations we only perform source training, where the first half of iterations is spent optimizing  $\mathcal{L}_{bce}$ , and the second half is spent jointly optimizing  $\mathcal{L}_{bce}$  and the source fairness objective. We note once optimization with respect to  $\mathcal{L}_{fair}$  starts, demographic parity decreases until adaptation start, i.e. between iterations 15,000 – 30,000. The validation accuracy in this interval also slightly decreases, as improving fairness may affect accuracy performance. During adaptation, i.e. after iteration 30,000, we observe that in the scenario where we use  $\mathcal{L}_{swd}$ , the target accuracy increases, while demographic parity on both source and target domains

remains relatively unchanged. In the scenario where no optimization of  $\mathcal{L}_{swd}$  is performed, there is still improvement with respect to target accuracy, however target demographic parity becomes on average larger. This implies that the distributional alignment loss done at the output of the encoder has beneficial effects both for the classification as well as the fairness objective.

In summary, we observe that the target accuracy consistently increases while demographic parity on both the source and target domains remains relatively unchanged, i.e., fairness is maintained. These observations validate that our algorithm leads to desired effects on the model performance.

## 6. Limitations

Since our approach relies on adversarial learning for domain alignment, it faces the limitations observed in adversarial learning. As described by Shu et al. (2018), adversarial learning is a suitable option for domain alignment when our base model can achieve low generalization error in both source and target domains if we have annotated labels in both domains. Only if this constraint is met, we can safely use adversarial learning for domain alignment. If such a classifier does not exist, when we train the base classifier model to achieve good generalization on the source domain using its annotated data points, its performance on the target domain is going to degrade because the two domains are not related enough to be learned jointly by the same classifier.

Moreover, when the support sets of the distributions in the source and target domains are disjoint prior to aligning the two domains, the constraint of domain invariance may become weak, thus impeding the feasibility of domain alignment. This observation was elucidated by Shu et al. (2018), who used virtual adversarial learning (Miyato et al., 2018) to could offer a potential solution to these challenges. In essence, virtual adversarial learning introduces perturbations to the input data that are maximally effective in inducing misclassification, thus encouraging the model to produce robust predictions across both domains, even when their supports are disjoint.

## 7. Conclusions and Future Work

We study the problem of fairness under domain shift. Fairness preserving methods have overlooked the problem of domain shift when deploying a source trained model to a target domain. Our first contribution is providing different data splits for common datasets employed in fairness tasks which present significant domain shift between the source and target. We show that as the distribution of data changes between the two domains, existing state-of-the-art fairness-preserving algorithms cannot match the performance they have on random data splits, where the source and target features are sampled from the same distribution. This observation demonstrates that model fairness is not naturally preserved under domain shift. Second, we present a novel algorithm that addresses domain shift when a fair outcome is of concern by combining fair model training via adversarial learning and producing a shared domain-agnostic latent feature space for the source and target domains by minimizing the distance between the source and target embedding distributions. Through empirical evaluation, we show that combining our algorithms maintains fairness

effectively under domain shift and also mitigates the effect of domain shift on the performance accuracy. Future extensions of this work includes considering scenarios where in addition to maintaining fairness under domain shift, the target domain maybe encountered sequentially, necessitating source-free model updating.

## Appendix A. Generating Data splits with Domain Shift

Split	Source				Target			
	Size	$Y=0$	$A=0 Y=0$	$A=0 Y=1$	Size	$Y=0$	$A=0 Y=0$	$A=0 Y=1$
A	34120	0.76	0.39	0.15	14722	0.76	0.39	0.15
A1	12024	0.53	0.41	0.16	5393	0.91	0.49	0.18
A2	29466	0.66	0.34	0.14	2219	0.97	0.48	0.30
A3	11887	0.52	0.42	0.16	778	0.89	0.39	0.17
C	3701	0.52	0.77	0.86	1577	0.54	0.76	0.84
C1	2886	0.58	0.74	0.82	1096	0.67	0.78	0.86
C2	903	0.47	0.80	0.80	96	0.74	0.70	0.92
C3	2031	0.45	0.80	0.85	162	0.58	0.60	0.79
G	697	0.70	0.28	0.37	303	0.70	0.30	0.34
G1	573	0.66	0.34	0.45	427	0.76	0.23	0.20
G2	388	0.61	0.36	0.49	196	0.84	0.20	0.16
G3	439	0.62	0.35	0.45	159	0.87	0.21	0.19

Table 8: Data split statistics. A,C,G correspond to the Adult, COMPAS and German dataset respectively. The rows with no number i.e. A,C,G correspond to random data splits. The numbered rows e.g. A1,A2,A3 correspond to statistics for specific splits. The columns represent the probabilities of specific outcomes for specific splits e.g.  $P(Y = 0)$ . Results when using *sex* as sensitive attribute.

The data splits employed in our approach are as follows:

**Adult Dataset.** We will use age, education and race to generate source and target domains. This can be a natural occurrence in practice, as gathered census information may differ along these axes geographically. For example, urban population is on average more educated than rural population<sup>6</sup>, and more ethnically diverse<sup>7</sup>. Thus, a fair model trained on one of the two populations will need to overcome distribution shift when evaluated on the other population. Besides differences in the feature distributions, we also note the Adult dataset is both imbalanced in terms of outcome,  $P(Y = 1) = 0.34$ , and sensitive attribute of positive outcome,  $P(A = 1|Y = 1) = 0.85$ , i.e. only a fraction of participants are earning more than 50k/year, and 85% of them are male.

The source/target splits we consider are as follows:

1. Source data: White, More than 12 education years. Target data: Non-white, Less than 12 education years.

<sup>6</sup><https://www.ers.usda.gov/topics/rural-economy-population/employment-education/rural-education/>

<sup>7</sup><https://www.ers.usda.gov/data-products/chart-gallery/gallery/chart-detail/?chartId=99538>

2. Source data: White, Older than 30. Target data: Non-white, younger than 40.
3. Source data: Younger than 70, More than 12 education years. Target data: Older than 70, less than 12 years of education.

In Table 8 we analyze the label and sensitive attribute conditional distributions for the above data splits. For the random split (A), the training and test label and conditional sensitive attribute distributions are identical, which is to be expected. For the three custom splits we observe all three distributions:  $P(Y)$ ,  $P(A|Y = 0)$ ,  $P(A|Y = 1)$  differ between training and test. We also note the label distribution becomes more skewed towards  $Y = 0$ .

**COMPAS Dataset** Compared to the Adult dataset, the COMPAS dataset is balanced in terms of label distribution, however is imbalanced in terms of the conditional distribution of the sensitive attribute. We will split the dataset along age, number of priors, and charge degree, i.e. whether the person committed a felony or misdemeanor. Considered splits are as follows:

1. Source data: Younger than 45, Less than 3 prior convictions. Test data: Older than 45, more than 3 prior convictions.
2. Source data: Younger than 45, White, At least one prior conviction. Target data: Older than 45, Non-white, No prior conviction.
3. Source data: Older than 25, At least one prior conviction, Convicted for a felony. Target data: Younger than 25, No priors, Convicted for a misdemeanor.

The first split tests whether a young population with limited number of convictions can be leveraged to fairly predict outcomes for an older population with more convictions. The second split introduces racial bias in the sampling process. In the third split we additionally consider the type of felony committed when splitting the dataset. For all splits, the test datasets become more imbalanced compared to the random split.

**German Credit Dataset** The dataset is smallest out of the three considered. For splitting we consider credit history and employment history. Similar to the Adult dataset, the label distribution is skewed towards increased risk i.e.  $P(Y = 0) = 0.7$ , and individuals of low risk are also skewed towards being part of the privileged group i.e.  $P(A = 1|Y = 1) = 0.63$ . We consider the following splits:

1. Source data: Employed up to 4 years. Test data: Employed long term (4+ years).
2. Source data: Up to date credit history, Employed less than 4 years. Target data: un-paid credit, Long term employed.
3. Source data: Delayed or paid credit, Employed up to 4 years. Target data: Critical account condition, Long term employment.

Compared to random data splits, the custom splits reduce label and sensitive attribute imbalance in the source domain, and increase these in the target domain.

## References

- Agarwal, A., Beygelzimer, A., Dudik, M., Langford, J., & Wallach, H. (2018). A reductions approach to fair classification. In Dy, J., & Krause, A. (Eds.), *Proceedings of the 35th International Conference on Machine Learning*, Vol. 80 of *Proceedings of Machine Learning Research*, pp. 60–69. PMLR.
- Ajakan, H., Germain, P., Larochelle, H., Laviolette, F., & Marchand, M. (2014). Domain-adversarial neural networks. *stat*, 1050, 15.
- Beutel, A., Chen, J., Zhao, Z., & Chi, E. H. (2017). Data decisions and theoretical implications when adversarially learning fair representations. *arXiv preprint arXiv:1707.00075*, 1.
- Bhushan Damodaran, B., Kellenberger, B., Flamary, R., Tuia, D., & Courty, N. (2018). Deepjdot: Deep joint distribution optimal transport for unsupervised domain adaptation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 447–463.
- Buolamwini, J., & Gebru, T. (2018). Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on fairness, accountability and transparency*, pp. 77–91. PMLR.
- Caliskan, A., Bryson, J. J., & Narayanan, A. (2017). Semantics derived automatically from language corpora contain human-like biases. *Science*, 356(6334), 183–186.
- Celis, L. E., Huang, L., Keswani, V., & Vishnoi, N. K. (2019a). Classification with fairness constraints: A meta-algorithm with provable guarantees. In *Proceedings of the conference on fairness, accountability, and transparency*, pp. 319–328.
- Celis, L. E., Huang, L., Keswani, V., & Vishnoi, N. K. (2019b). Classification with fairness constraints: A meta-algorithm with provable guarantees. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, FAT\* ’19, p. 319–328 New York, NY, USA. Association for Computing Machinery.
- Chouldechova, A., & Roth, A. (2018). The frontiers of fairness in machine learning. *arXiv preprint arXiv:1810.08810*, 1.
- Cirillo, D., Catuara-Solarz, S., Morey, C., Guney, E., Subirats, L., Mellino, S., Gigante, A., Valencia, A., Rementeria, M. J., Chadha, A. S., et al. (2020). Sex and gender differences and biases in artificial intelligence for biomedicine and healthcare. *NPJ digital medicine*, 3(1), 81.
- Coston, A., Ramamurthy, K. N., Wei, D., Varshney, K. R., Speakman, S., Mustahsan, Z., & Chakraborty, S. (2019). Fair transfer learning with missing protected attributes. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, pp. 91–98.

- Courty, N., Flamary, R., Tuia, D., & Rakotomamonjy, A. (2016). Optimal transport for domain adaptation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(9), 1853–1865.
- Daumé III, H. (2007). Frustratingly easy domain adaptation. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pp. 256–263.
- Du, M., Mukherjee, S., Wang, G., Tang, R., Awadallah, A., & Hu, X. (2021). Fairness via representation neutralization. *Advances in Neural Information Processing Systems*, 34.
- Dua, D., & Graff, C. (2017). UCI machine learning repository..
- Eidinger, E., Enbar, R., & Hassner, T. (2014). Age and gender estimation of unfiltered faces. *IEEE Transactions on information forensics and security*, 9(12), 2170–2179.
- Feldman, M., Friedler, S. A., Moeller, J., Scheidegger, C., & Venkatasubramanian, S. (2015). Certifying and removing disparate impact. In *proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining*, pp. 259–268.
- Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., & Bengio, Y. (2014). Generative adversarial nets. *Advances in neural information processing systems*, 27.
- Hardt, M., Price, E., & Srebro, N. (2016). Equality of opportunity in supervised learning. *Advances in neural information processing systems*, 29.
- HassanPour Zonoozi, M., & Seydi, V. (2023). A survey on adversarial domain adaptation. *Neural Processing Letters*, 55(3), 2429–2469.
- Hu, H., Liu, Y., Wang, Z., & Lan, C. (2019). A distributed fair machine learning framework with private demographic data protection. In *2019 IEEE International Conference on Data Mining (ICDM)*, pp. 1102–1107. IEEE.
- Jian, D., & Rostami, M. (2023). Unsupervised domain adaptation for training event-based networks using contrastive learning and uncorrelated conditioning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 18721–18731.
- Jiang, R., Pacchiano, A., Stepleton, T., Jiang, H., & Chiappa, S. (2020). Wasserstein fair classification. In *Uncertainty in Artificial Intelligence*, pp. 862–872. PMLR.
- Kamiran, F., Karim, A., & Zhang, X. (2012). Decision theory for discrimination-aware classification. In *2012 IEEE 12th International Conference on Data Mining*, pp. 924–929.
- Kleinberg, J., Mullainathan, S., & Raghavan, M. (2016). Inherent trade-offs in the fair determination of risk scores. *arXiv preprint arXiv:1609.05807*, 1.
- Lee, C.-Y., Batra, T., Baig, M. H., & Ulbricht, D. (2019). Sliced wasserstein discrepancy for unsupervised domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

- Long, M., Cao, Y., Wang, J., & Jordan, M. (2015). Learning transferable features with deep adaptation networks. In *Proceedings of International Conference on Machine Learning*, pp. 97–105.
- Long, M., Zhu, H., Wang, J., & Jordan, M. I. (2017). Deep transfer learning with joint adaptation networks. In *Proceedings of the 34th International Conference on Machine Learning- Volume 70*, pp. 2208–2217. JMLR. org.
- Luo, Y., Zheng, L., Guan, T., Yu, J., & Yang, Y. (2019). Taking a closer look at domain shift: Category-level adversaries for semantics consistent domain adaptation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 2507–2516.
- Madras, D., Creager, E., Pitassi, T., & Zemel, R. (2018a). Learning adversarially fair and transferable representations. In *International Conference on Machine Learning*, pp. 3384–3393. PMLR.
- Madras, D., Creager, E., Pitassi, T., & Zemel, R. (2018b). Learning adversarially fair and transferable representations. In Dy, J., & Krause, A. (Eds.), *Proceedings of the 35th International Conference on Machine Learning*, Vol. 80 of *Proceedings of Machine Learning Research*, pp. 3384–3393. PMLR.
- McInnes, L., Healy, J., & Melville, J. (2020). Umap: Uniform manifold approximation and projection for dimension reduction..
- Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2021). A survey on bias and fairness in machine learning. *ACM Computing Surveys (CSUR)*, 54(6), 1–35.
- Mei, Z., Ye, P., Ye, H., Li, B., Guo, J., Chen, T., & Ouyang, W. (2023). Automatic loss function search for adversarial unsupervised domain adaptation. *IEEE Transactions on Circuits and Systems for Video Technology*, 33(10), 5868–5881.
- Miyato, T., Maeda, S.-i., Koyama, M., & Ishii, S. (2018). Virtual adversarial training: a regularization method for supervised and semi-supervised learning. *IEEE transactions on pattern analysis and machine intelligence*, 41(8), 1979–1993.
- Mukherjee, D., Petersen, F., Yurochkin, M., & Sun, Y. (2022). Domain adaptation meets individual fairness. and they get along.. *Advances in Neural Information Processing Systems*, 35, 28902–28913.
- Nananukul, N., Soltanian-Zadeh, H., & Rostami, M. (2024). Multi-source data integration for segmentation of unannotated mri images. *IEEE Journal of Biomedical and Health Informatics*, 25.
- Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Kopf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J., & Chintala, S. (2019). Pytorch: An imperative style, high-performance deep learning library. In

- Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E., & Garnett, R. (Eds.), *Advances in Neural Information Processing Systems 32*, pp. 8024–8035. Curran Associates, Inc.
- Pooch, E. H., Ballester, P., & Barros, R. C. (2020). Can we trust deep learning based diagnosis? the impact of domain shift in chest radiograph classification. In *Thoracic Image Analysis: Second International Workshop, TIA 2020, Held in Conjunction with MICCAI 2020, Lima, Peru, October 8, 2020, Proceedings 2*, pp. 74–83. Springer.
- Redko, I., Habrard, A., & Sebban, M. (2017). Theoretical analysis of domain adaptation with optimal transport. In Ceci, M., Hollmén, J., Todorovski, L., Vens, C., & Džeroski, S. (Eds.), *Machine Learning and Knowledge Discovery in Databases*, pp. 737–753 Cham. Springer International Publishing.
- Romera-Paredes, B., & Torr, P. (2015). An embarrassingly simple approach to zero-shot learning. In *International conference on machine learning*, pp. 2152–2161. PMLR.
- Rostami, M. (2024a). Continuous unsupervised domain adaptation using stabilized representations and experience replay. *Neurocomputing*, 597, 128017.
- Rostami, M. (2024b). Improving unsupervised domain adaptation through class-conditional compact representations. *Neural Computing and Applications*, 36, 1–18.
- Rostami, M., Bose, D., Narayanan, S., & Galstyan, A. (2023). Domain adaptation for sentiment analysis using robust internal representations. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 11484–11498.
- Rostami, M., & Galstyan, A. (2023). Cognitively inspired learning of incremental drifting concepts. In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence*, pp. 3058–3066.
- Rostami, M., Isele, D., & Eaton, E. (2020). Using task descriptions in lifelong machine learning for improved performance and zero-shot transfer. *Journal of Artificial Intelligence Research*, 67, 673–704.
- Rostami, M., Kolouri, S., Eaton, E., & Kim, K. (2019). Sar image classification using few-shot cross-domain transfer learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pp. 0–0.
- Rostami, M., Kolouri, S., Murez, Z., Owechko, Y., Eaton, E., & Kim, K. (2022). Zero-shot image classification using coupled dictionary embedding. *Machine Learning with Applications*, 8, 100278.
- Saito, K., Watanabe, K., Ushiku, Y., & Harada, T. (2018). Maximum classifier discrepancy for unsupervised domain adaptation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3723–3732.
- Schumann, C., Wang, X., Beutel, A., Chen, J., Qian, H., & Chi, E. H. (2019). Transfer of machine learning fairness across domains. *arXiv preprint arXiv:1906.09688*, 1.

- Shu, R., Bui, H. H., Narui, H., & Ermon, S. (2018). A dirt-t approach to unsupervised domain adaptation. In *Proc. 6th International Conference on Learning Representations*.
- Singh, H., Singh, R., Mhasawade, V., & Chunara, R. (2021). Fairness violations and mitigation under covariate shift. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pp. 3–13.
- Stan, S., & Rostami, M. (2024a). Source-free domain adaptation for semantic image segmentation using internal representations. *Frontiers in Big Data*, 7, 1359317.
- Stan, S., & Rostami, M. (2024b). Unsupervised model adaptation for source-free segmentation of medical images. *Medical Image Analysis*, 95, 103179.
- Sun, B., & Saenko, K. (2016). Deep coral: Correlation alignment for deep domain adaptation. In *European conference on computer vision*, pp. 443–450. Springer.
- Tzeng, E., Hoffman, J., Saenko, K., & Darrell, T. (2017). Adversarial discriminative domain adaptation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 7167–7176.
- Wick, M., Tristan, J.-B., et al. (2019). Unlocking fairness: a trade-off revisited. *Advances in neural information processing systems*, 32.
- Wu, H., Yan, Y., Ng, M. K., & Wu, Q. (2020). Domain-attention conditional wasserstein distance for multi-source domain adaptation. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 11(4), 1–19.
- Xian, Y., Schiele, B., & Akata, Z. (2017). Zero-shot learning—the good, the bad and the ugly. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4582–4591.
- Zhang, B. H., Lemoine, B., & Mitchell, M. (2018). Mitigating unwanted biases with adversarial learning. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, pp. 335–340.
- Zhang, Y., & Long, Q. (2021). Assessing fairness in the presence of missing data. *Advances in Neural Information Processing Systems*, 34.
- Zhang, Z., Song, Y., & Qi, H. (2017). Age progression/regression by conditional adversarial autoencoder. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 5810–5818.