ECAI 2023
K. Gal et al. (Eds.)

doi:10.3233/FAIA230556

Fair Few-Shot Learning with Auxiliary Sets

Song Wang[a], Jing Ma[a], Lu Cheng[b] and Jundong Li[a]

[a]University of Virginia
[b]University of Illinois Chicago
ORCiD ID: Song Wang https://orcid.org/0000-0003-1273-7694, Jing Ma https://orcid.org/0000-0003-4237-6607,
Lu Cheng https://orcid.org/0000-0002-2503-2522, Jundong Li https://orcid.org/0000-0002-1878-817X

Abstract. Recently, there has been a growing interest in developing machine learning (ML) models that can promote fairness, i.e., eliminating biased predictions towards certain populations (e.g., individuals from a specific demographic group). Most existing works learn such models based on well-designed fairness constraints in optimization. Nevertheless, in many practical ML tasks, only very few labeled data samples can be collected, which can lead to inferior fairness performance. This is because existing fairness constraints are designed to restrict the prediction disparity among different sensitive groups, but with few samples, it becomes difficult to accurately measure the disparity, thus rendering ineffective fairness optimization. In this paper, we define the fairness-aware learning task with limited training samples as the *fair few-shot learning* problem. To deal with this problem, we devise a novel framework that accumulates fairness-aware knowledge across different meta-training tasks and then generalizes the learned knowledge to meta-test tasks. To compensate for insufficient training samples, we propose an essential strategy to select and leverage an *auxiliary set* for each meta-test task. These auxiliary sets contain several labeled training samples that can enhance the model performance regarding fairness in meta-test tasks, thereby allowing for the transfer of learned useful fairness-oriented knowledge to meta-test tasks. Furthermore, we conduct extensive experiments on three real-world datasets to validate the superiority of our framework against the state-of-the-art baselines.

1 Introduction

Machine learning (ML) tools have been increasingly utilized in high-stake tasks such as credit assessments [26] and crime predictions [22]. Despite their success, the data-driven nature of existing machine learning methods makes them easily inherit the biases buried in the training data and thus results in predictions with discrimination against some sensitive groups [33]. Here, sensitive groups are typically defined by certain sensitive attributes such as race and gender [35, 3, 4, 19, 45]. For example, a criminal risk assessment model can unfavorably assign a higher crime probability for specific racial groups [33]. In fact, such undesirable biases commonly exist in various real-world applications such as toxicity detection [6], recommendation systems [21], loan approval predictions [29], and recruitment [11].

In response, a surge of research efforts in both academia and industry have been made for developing fair machine learning models [9, 7]. These models have demonstrated their ability to effectively mitigate unwanted bias in various applications [1, 47]. Many fair ML methods [8, 10] incorporate fairness constraints to penalize predictions with statistical discrepancies among different sensitive groups. These methods often rely on sufficient training data from each sensitive group (e.g., collecting data from a specific region with an imbalanced population composition [49]). However, in many scenarios, only very few data samples can be collected, especially for those from the minority group. This could render existing fair ML methods ineffective or even further amplify discrimination against the minority group. To enhance the applicability of fair ML in practice [49], this work aims to address the crucial and urgent problem of *fair few-shot learning*: promoting fairness in few-shot learning tasks with a limited number of samples.

One feasible solution to address fair few-shot learning is to incorporate fairness techniques into few-shot learning methods. Particularly, we first learn from *meta-training tasks* with adequate samples [32, 18, 39], and then leverage the learned knowledge and fine-tune the model on other disjoint *meta-test tasks* with few samples based on fairness constraints. We define such a step of fine-tuning as *fairness adaptation*. However, there still remain two primary challenges for our problem. First, the insufficiency of samples in meta-test tasks can result in unsatisfactory fairness adaptation performance. Although the model can adapt to meta-test tasks with limited samples via fine-tuning for classification, these samples may not be sufficient to ensure fairness performance. Many fairness constraints are designed to restrict the prediction disparity among different sensitive groups. However, in fair few-shot learning, the lack of samples in each sensitive group inevitably increases the difficulties in measuring the prediction disparity. Moreover, in meta-test sets, the sensitive attributes of data samples can often be extremely imbalanced (e.g., a majority of individuals belonging to the same race, while other sensitive groups have very few, or even no samples). In these cases, the conventional fairness constraints are often ineffective, or completely inapplicable. Second, the generalization gap between meta-training tasks and meta-test tasks hinders the efficacy of fairness adaptation. Similar to other few-shot learning studies, the key point of fair few-shot learning is to leverage the learned knowledge from meta-training tasks to facilitate the model performance on meta-test tasks with few samples. In our problem, it is essential to leverage the learned knowledge for fairness adaptation. However, models that manage to reduce disparities on meta-training tasks do not necessarily achieve the same performance in fairness on meta-test tasks [10], due to the fact that fairness constraints are data-dependent and thus lack generalizability [8]. As a result, it remains challenging to extract and leverage the learned knowledge that is beneficial for fairness adaptation.

To tackle these challenges, we devise a novel framework for fair few-shot learning, named FEAST (**F**air f**E**w-shot learning with **A**uxiliary **S**e**T**s). Specifically, we propose to leverage an *auxiliary set* for each meta-test task to promote fair adaptation with few samples while addressing the issues caused by insufficient samples. The auxiliary set is comprised of several samples from meta-training data and is specific to each meta-test task. By incorporating these auxiliary sets via a novel *fairness-aware mutual information loss*, the model can be effectively adapted to a meta-task with few samples while preserving the fairness knowledge learned during training. Furthermore, to effectively leverage the learned knowledge from meta-training tasks for fairness adaptation, our proposed framework selects the auxiliary sets based on the *fairness adaptation direction*. This ensures that the selected auxiliary sets share similar fairness adaptation directions and thus can provide beneficial learned knowledge. We summarize our main contributions as follows:

- **Problem.** We study the crucial problem of fair few-shot learning. We introduce the importance of this problem, analyze the challenges, and point out the limitations of existing studies. To the best of our knowledge, this is the first work that addresses these unique challenges in fair few-shot learning.
- **Method.** We develop a novel fair few-shot learning framework that (1) can leverage auxiliary sets to aid fairness adaptation with limited samples, and (2) can select auxiliary sets with similar optimization directions to promote fairness adaptation.
- **Experiments.** We conduct extensive experiments on three real-world fairness datasets under the few-shot scenario and demonstrate the superiority of our proposed framework in terms of fairness compared with a couple of state-of-the-art baselines.

2 Problem Statement

In this section, we provide a formal definition for the problem of fair few-shot learning that we study in this paper. Denote $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$ as the input space, where $\mathcal{X} \subset \mathbb{R}^n$ is the input space with n different features and $\mathcal{Y} = \{1, 2, \ldots, N\}$ is the label space with N discrete classes. We consider inputs $X \in \mathcal{X}$, labels $Y \in \mathcal{Y}$, and sensitive attribute $A \in \{0, 1\}$. In the few-shot setting, the dataset $\mathcal{D}$ is comprised of two different smaller datasets: meta-training data $\mathcal{D}_{tr}$ and meta-test data $\mathcal{D}_{te}$. Moreover, $\mathcal{D} = \mathcal{D}_{tr} \cup \mathcal{D}_{te}$ and $\mathcal{D}_{tr} \cap \mathcal{D}_{te} = \emptyset$, i.e., $|\mathcal{D}_{tr}| + |\mathcal{D}_{te}| = |\mathcal{D}|$. In general, few-shot settings assume that there exist sufficient samples in $\mathcal{D}_{tr}$, while samples in $\mathcal{D}_{te}$ are generally scarce [18, 34].

The proposed framework is built upon the prevalent paradigm of episodic meta-learning [34, 32], which has demonstrated superior performance in the field of few-shot learning [18, 39]. The process of episodic meta-learning consists of meta-training on $\mathcal{D}_{tr}$ and meta-test on $\mathcal{D}_{te}$. During meta-training, the model is trained on a series of *meta-training tasks* $\{\mathcal{T}_1, \mathcal{T}_2, \ldots, \mathcal{T}_T\}$, where each meta-training task contains support set $\mathcal{S}$ as the reference and a query set $\mathcal{Q}$ to be classified. T is the number of meta-training tasks. More specifically, $\mathcal{S} = \{(x_1, y_1), (x_2, y_2), \ldots, (x_{N\times K}, y_{N\times K})\}$ contains N classes and K samples for each of these N classes (i.e., the N-way K-shot setting). Meanwhile, the query set $\mathcal{Q} = \{(x_1^q, y_1^q), (x_2^q, y_2^q), \ldots, (x_{|\mathcal{Q}|}^q, y_{|\mathcal{Q}|}^q)\}$ consists of $|\mathcal{Q}|$ different samples to be classified from these N classes. Subsequently, our goal is to develop a machine learning model that can accurately and fairly predict labels for samples in $\mathcal{D}_{te}$ with limited labeled samples after training on $\mathcal{D}_{tr}$. Formally, the studied problem of fair few-shot learning can be formulated as follows.

Definition 1. ***Fair few-shot learning:*** *Given meta-training data $\mathcal{D}_{tr}$ and a meta-test task $\mathcal{T} = \{\mathcal{S}, \mathcal{Q}\}$ sampled from meta-test data $\mathcal{D}_{te}$, our goal is to develop a fair learning model such that after meta-training on samples in $\mathcal{D}_{tr}$, the model can accurately and fairly predict labels for samples in the query set $\mathcal{Q}$ when the only available reference is the limited samples in the support set $\mathcal{S}$.*

Note that the support sets and the query sets are sampled from meta-training data $\mathcal{D}_{tr}$. That is, for any sample (x_i, y_i) in a meta-training task, $(x_i, y_i) \sim P_{tr}(X, Y)$, where $P_{tr}(X, Y)$ is the meta-training task distribution from meta-training data $\mathcal{D}_{tr}$. We then evaluate the model on a series of meta-test tasks, which share the same structure as meta-training tasks, except that the samples are now from meta-test data $\mathcal{D}_{te}$. In other words, for any sample (x_i, y_i) during meta-test, we have $(x_i, y_i) \sim P_{te}(X, Y)$, where $P_{te}(X, Y)$ is the meta-test task distribution from meta-test data $\mathcal{D}_{te}$. Under the meta-learning framework [18, 51, 20], the model needs to be first fine-tuned for several steps (i.e., fairness adaptation) using the support set, and then performs fair classification for samples in the query set.

3 Proposed Framework

We formulate the problem of *fair few-shot learning* in the N-way K-shot meta-learning framework. The meta-training process typically involves a series of randomly sampled meta-training tasks, each of which contains K samples for each of the N classes as the support set, along with several query samples to be classified. Under the few-shot scenario, it is challenging to conduct fairness adaptation on the support set due to the insufficiency of samples and the generalization gap between meta-training tasks and meta-test tasks. Therefore, as illustrated in Fig. 1, we propose the use of auxiliary sets that can enhance fairness adaptation for each meta-test task. In this section, we first introduce the process of conducting fairness adaptation with auxiliary sets and then discuss the strategy to select auxiliary sets.

3.1 Fairness Adaptation with Auxiliary Sets

To alleviate the issue of ineffective fairness adaptation to meta-test tasks caused by insufficient samples, we propose to leverage the samples in meta-training tasks for fairness adaptation. Specifically, considering a target meta-test task $\mathcal{T} = (\mathcal{S}, \mathcal{Q})$, our goal is to utilize an auxiliary set $\mathcal{A}$ obtained from meta-training data that can compensate for inadequate samples in $\mathcal{S}$. However, due to the distribution difference between meta-training tasks and meta-test tasks, it remains non-trivial to leverage the auxiliary set $\mathcal{A}$, which follows a different distribution from $\mathcal{S}$. Since the data distribution in $\mathcal{A}$ differs from that in $\mathcal{S}$, directly conducting fairness adaptation on $\mathcal{A}$ can be ineffective for fairness in $\mathcal{S}$. Therefore, to enhance fairness adaptation with the help of the auxiliary set $\mathcal{A}$, we propose to maximize the mutual information (MI) between the support set $\mathcal{S}$ and the auxiliary set $\mathcal{A}$. In consequence, the fairness adaptation on $\mathcal{S}$ will benefit from $\mathcal{A}$.

Generally, the support set $\mathcal{S}$ in $\mathcal{T}$ can be expressed as $\mathcal{S} = \{(x_1, y_1), (x_2, y_2), \ldots, (x_{N\times K}, y_{N\times K})\}$, which contains K samples for each of N classes. x_i is an input sample, and y_i is the corresponding label. We use $a_i \in \{0, 1\}$ to denote its sensitive attribute. In particular, we propose to construct an auxiliary set that shares the same structure as the support set. In this way, the auxiliary set $\mathcal{A}$ can be represented as $\mathcal{A} = \{(x_1^*, y_1^*), (x_2^*, y_2^*), \ldots, (x_{|\mathcal{A}|}^*, y_{|\mathcal{A}|}^*)\}$. Here $|\mathcal{A}|$, i.e., the size of the auxiliary set, is set as a controllable hyper-parameter. Moreover, based on the classification model $f(\cdot)$, we can obtain the sample embedding $\mathbf{x}_i \in \mathbb{R}^d$, and the classification probabilities $\mathbf{p}_i = f(x_i) \in \mathbb{R}^N$ for x_i. Here d denotes the embedding

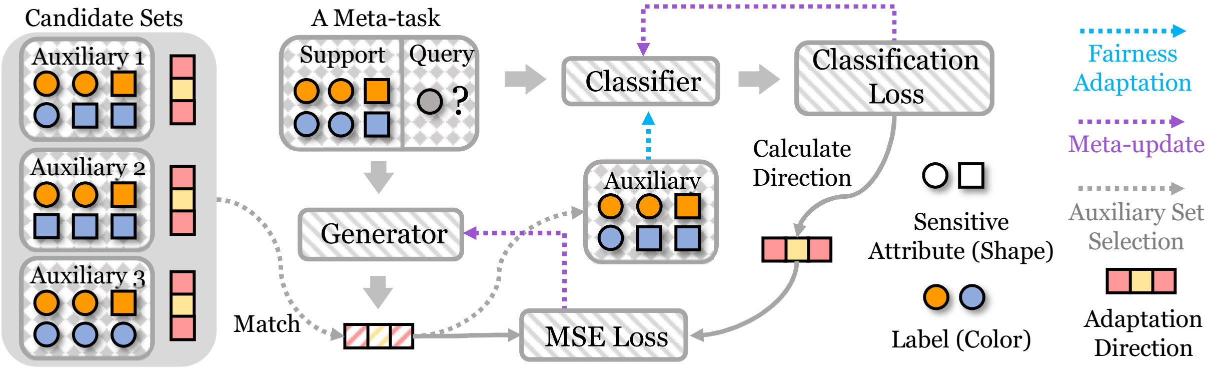

Figure 1: The overall framework of FEAST. Here different shapes denote different sensitive attributes, and colors represent sample classes. Given a meta-task, the generator will output the estimated fairness adaptation direction, which is used to select an auxiliary set with the most similar direction from the candidate set. Then we conduct fairness adaptation with the auxiliary set on the current meta-task and perform predictions. The resulting fairness adaptation will be used to update the generator. Note that during training, the meta-task will be incorporated into the candidate auxiliary sets after the optimization of one episode.

dimension of samples, and N is the number of classes in $\mathcal{T}$. Particularly, we maximize the fairness-aware MI between $\mathcal{S}$ and $\mathcal{A}$ by

$$\max_\theta I(\mathcal{S};\mathcal{A}) = \max_\theta \sum_{i=1}^{|\mathcal{S}|}\sum_{j=1}^{|\mathcal{A}|} p(x_i, x_j^*;\theta)\log\frac{p(x_i|x_j^*;\theta)}{p(x_i;\theta)}, \tag{1}$$

where θ denotes the parameters of classification model $f(\cdot)$. Since the MI term $I(\mathcal{S};\mathcal{A})$ is difficult to obtain and also intractable, it is infeasible to directly maximize it [27]. Therefore, we first re-formulate the MI term to make it computationally tractable based on the property of conditional probabilities:

$$\begin{aligned} I(\mathcal{S};\mathcal{A}) &= \sum_{i=1}^{|\mathcal{S}|}\sum_{j=1}^{|\mathcal{A}|} p(x_i|x_j^*;\theta)p(x_j^*;\theta)\log\frac{p(x_i|x_j^*;\theta)}{p(x_i;\theta)} \\ &= \sum_{i=1}^{|\mathcal{S}|}\sum_{j=1}^{|\mathcal{A}|} p(x_j^*|x_i;\theta)p(x_i;\theta)\log\frac{p(x_i|x_j^*;\theta)}{p(x_i;\theta)}. \end{aligned} \tag{2}$$

Since the support set $\mathcal{S}$ is randomly sampled, we can assume that the prior probability $p(x_i;\theta)$ follows a uniform distribution and set it as a constant: $p(x_i;\theta) = 1/|\mathcal{S}|$, which thus can be ignored in optimization. Therefore, it remains to estimate $p(x_i|x_j^*;\theta)$ and $p(x_j^*|x_i;\theta)$ to obtain the value of $I(\mathcal{S};\mathcal{A})$.

3.1.1 Estimation of $p(x_i|x_j^*;\theta)$

We first denote $\mathcal{S}_0$ and $\mathcal{S}_1$ as the sets of samples with sensitive attributes of 0 and 1, respectively[1]. In other words, $\mathcal{S} = \mathcal{S}_0 \cup \mathcal{S}_1$ and $\mathcal{S}_0 \cap \mathcal{S}_1 = \emptyset$. Similarly, we define sets $\mathcal{A}_0$ and $\mathcal{A}_1$ for the auxiliary set $\mathcal{A}$. Then we propose to estimate $p(x_i|x_j^*;\theta)$ as follows:

$$p(x_i|x_j^*;\theta) = \begin{cases} \dfrac{\mathbf{p}_i(y_j^*)}{\sum_{x_k\in\mathcal{S}_{a_i}}\mathbf{p}_k(y_j^*)} & \text{if } a_i = a_j^*, \\ 0 & \text{else.} \end{cases} \tag{3}$$

Here $\mathbf{p}_i(y_j^*) \in \mathbb{R}$ denotes the classification probability of x_i regarding y_j^*, which is the label of x_j^*. Intuitively, the probability measures the alignment of the classification between the support sample x_i and the auxiliary sample x_j^*, which (1) shares the same sensitive attribute with x_i and (2) is also similar to x_i regarding the classification output. In other words, maximizing $p(x_i|x_j^*;\theta)$ can increase the fairness adaptation consistency between sample x_i and auxiliary samples that are specifically beneficial for the fairness adaptation with x_i, thus promoting the fairness adaptation performance.

3.1.2 Estimation of $p(x_j^*|x_i;\theta)$

The term $p(x_j^*|x_i;\theta)$ in Eq. (2) is conditioned on x_i and denotes the probability of x_j^* inferred by x_i. Moreover, since the value of $p(x_i|x_j^*;\theta)$ becomes zero when the sensitive attributes of x_i and x_j^* are different, we only need to estimate $p(x_j^*|x_i;\theta)$ when x_i and x_j^* share the same sensitive attributes, i.e., $a_i = a_j^*$. Therefore, since x_i and x_j^* maintain the same sensitive attributes, we can estimate the probability $p(x_j^*|x_i;\theta)$ based on the squared Euclidean distance between their embeddings without explicitly considering their fairness-aware correlation. In particular, we further normalize the probability with a softmax function to formulate term $p(x_j^*|x_i;\theta)$ as follows:

$$p(x_j^*|x_i;\theta) = \frac{\exp\left(-\|\mathbf{x}_i - \mathbf{x}_j^*\|_2^2\right)}{\sum_{x_k^*\in\mathcal{A}_{a_j^*}}\exp\left(-\|\mathbf{x}_i - \mathbf{x}_k^*\|_2^2\right)}. \tag{4}$$

Furthermore, to ensure the consistency of sample representations in meta-training and meta-test data, we apply the ℓ_2 normalization on both $\mathbf{x}_i$ and $\mathbf{x}_j^*$, which results in $\|\mathbf{x}_i - \mathbf{x}_j^*\|_2^2 = 2 - 2\mathbf{x}_i^\top \cdot \mathbf{x}_j^*$. In this manner, the logarithmic term $\log p(x_j^*|x_i;\theta)$ becomes:

$$\begin{aligned} \log\left(p(x_j^*|x_i;\theta)\right) &= \log\left(\frac{\exp\left(-2 + 2\mathbf{x}_i^\top\cdot\mathbf{x}_j^*\right)}{\sum_{x_k^*\in\mathcal{A}_{a_j^*}}\exp\left(-2 + 2\mathbf{x}_i^\top\cdot\mathbf{x}_k^*\right)}\right) \\ &= 2\mathbf{x}_i^\top\cdot\mathbf{x}_j^* - \log\sum_{x_k^*\in\mathcal{A}_{a_j^*}}\exp\left(2\mathbf{x}_i^\top\cdot\mathbf{x}_k^*\right). \end{aligned} \tag{5}$$

Finally, the MI loss $\mathcal{L}_{MI}$ can be derived as follows:

$$\begin{aligned} \mathcal{L}_{MI} = &\frac{1}{|\mathcal{A}|}\sum_{j=1}^{|\mathcal{A}|}\sum_{x_i\in\mathcal{S}_{a_j^*}} -\frac{\mathbf{p}_i(y_j^*)}{\sum_{x_k\in\mathcal{S}_{a_i}}\mathbf{p}_k(y_j^*)}\Bigg(2\mathbf{x}_i^\top\cdot\mathbf{x}_j^* \\ &- \log\sum_{x_k^*\in\mathcal{A}_{a_j^*}}\exp\left(2\mathbf{x}_i^\top\cdot\mathbf{x}_k^*\right)\Bigg). \end{aligned} \tag{6}$$

[1] For the sake of simplicity, we focus on tasks with only binary sensitive attributes in this paper. Nevertheless, our work can be easily generalized to tasks with multiple types of sensitive attributes.

The overall fairness adaptation loss can be represented as the combination of fairness regularization terms on the support set $\mathcal{S}$ and the auxiliary set $\mathcal{A}$ along with the MI loss between $\mathcal{S}$ and $\mathcal{A}$:

$$\mathcal{L}_{FA} = \mathcal{L}_R(\mathcal{S}) + \gamma \left(\mathcal{L}_R(\mathcal{A}) + \mathcal{L}_{MI}\right), \tag{7}$$

where γ is an adjustable weight hyper-parameter to control the importance of the auxiliary set. Specifically, $\mathcal{L}_R$ denotes the regularized optimization loss:

$$\mathcal{L}_R(S) = \frac{1}{|\mathcal{S}|} \sum_{(x,y)\in\mathcal{S}} \ell(f(x), y) + \lambda R(\mathcal{S}), \tag{8}$$

where ℓ is the classification loss, and $R(\mathcal{S})$ denotes the fairness regularization term.

3.2 Auxiliary Sets Selection

The second problem of the generalization gap between meta-training and meta-test in fair few-shot learning can also pose a significant challenge in fairness adaptation. To address this issue, we propose to select the auxiliary set based on its similarity in fairness adaptation directions to the target meta-test task. In this way, incorporating the auxiliary set with a similar fairness adaptation direction can potentially leverage beneficial learned knowledge in meta-training to enhance fairness adaptation in the target meta-task. However, it is difficult to identify the fairness adaptation direction of the auxiliary set that aligns with the target meta-task. It is possible that the auxiliary set holds a different or even opposite fairness adaptation direction from the target meta-task. As such, the incorporation of such an auxiliary set can even harm the fairness adaptation performance. Therefore, to select the auxiliary set with a similar fairness adaptation direction to the target meta-test task, we introduce a *dynamic dictionary*, $\mathcal{A}_{can}$, which stores all candidate auxiliary sets for selection, with the keys being their corresponding fairness adaptation directions. This allows us to efficiently identify and select an auxiliary set with a similar adaptation direction for the target meta-test task, thereby improving the fairness adaptation performance in the presence of the generalization gap.

Notably, this dictionary will be dynamically updated by adding a new auxiliary set after each meta-training step and meanwhile removing the oldest auxiliary set, of which the fairness adaptation direction is the most outdated. In this manner, the dictionary also acts like a queue, which means that the size can be flexible and independent to fit various scenarios. Specifically, after each step on a meta-training task $\mathcal{T} = \{\mathcal{S}, \mathcal{Q}\}$, we will enqueue the support set $\mathcal{S}$ as a candidate auxiliary set[2] into $\mathcal{A}_{can}$ and remove the oldest auxiliary set. The key of enqueued $\mathcal{S}$, which is the fairness adaptation direction of $\mathcal{S}$, is set as the gradient of $\mathcal{L}_R(\mathcal{S})$, i.e., $\nabla_\theta \mathcal{L}_R(\mathcal{S})$, where θ denotes the model parameters of $f(\cdot)$.

Algorithm 1 Detailed training process of our framework.

```
Input: Meta-training task distribution $P_{tr}$ from the meta-training data $\mathcal{D}_{tr}$, number of meta-training tasks $T$, number of fine-tuning steps $\tau$.
Output: A trained fairness-aware classification model $f(\cdot)$ and a generator model $g(\cdot)$.
Randomly initialize the dictionary queue $\mathcal{A}_{can}$;
for $i = 1, 2, \ldots, T$ do
    Sample a meta-training task $\mathcal{T}_i = \{\mathcal{S}, \mathcal{Q}\} \sim P_{tr}$;
    Obtain the fairness adaptation direction via Eq. (10);
    Select an auxiliary set $\mathcal{A}$ from the candidate auxiliary set dictionary $\mathcal{A}_{can}$ based on Eq. (11);
    for $t = 1, 2, \ldots, \tau$ do
        Conduct one step of fairness adaptation according to Eq. (7) and Eq. (12);
    end for
    Meta-optimize classification model $f(\cdot)$ and generator $g(\cdot)$ based on Eq. (13) and Eq. (14), respectively;
    Enqueue support set $\mathcal{S}$ into the dictionary queue $\mathcal{A}_{can}$ and remove the oldest candidate auxiliary set in $\mathcal{A}_{can}$;
end for
```

Identifying the true fairness adaptation direction. With the help of the dynamic dictionary as a queue during meta-training, it may still remain difficult to obtain the fairness adaptation direction of the target meta-test task $\mathcal{T}$. This is because the fairness adaptation direction of $\mathcal{S}$ cannot faithfully reveal the true direction due to potentially imbalanced sensitive attributes. Therefore, to identify the true fairness adaptation direction without directly conducting fairness adaptation on the support set $\mathcal{S}$, we propose the use of a generator $g(\cdot)$, parameterized by ϕ, to estimate the fairness adaptation results for each meta-test task. In particular, the generator $g(\cdot)$ takes the support set $\mathcal{S}$ as input and outputs an estimation of the gradient of $\mathcal{L}_R(\mathcal{S})$, i.e., $\nabla_\theta \mathcal{L}_R(\mathcal{S})$. To optimize the generator $g(\cdot)$, we introduce the Mean Squared Error (MSE) loss as the objective function as follows:

$$\mathcal{L}_E = \|g(\mathcal{S}) - \nabla_\theta \mathcal{L}_R(\mathcal{S})\|_2^2, \tag{9}$$

where $g(\mathcal{S}) \in \mathbb{R}^{d_\theta}$ is the generator output, and d_θ is the size of the classification model parameter θ. It is worth mentioning that the input of the generator $g(\cdot)$ is an entire support set $\mathcal{S}$, which means that the generator should be able to capture the contextual information within the support set. For this reason, we propose to leverage the transformer encoder architecture [38] followed by a Multiple Layer Perceptron (MLP) as the implementation of the generator. In specific, the output of the generator can be expressed as:

$$g(\mathcal{S}) = \text{MLP}\left(\text{Mean}\left(\text{Transformer}\left(\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_{|\mathcal{S}|}\right)\right)\right). \tag{10}$$

In this manner, the generator can estimate the corresponding fairness adaptation direction from $\mathcal{S}$, where the result can be used for selecting an auxiliary set.

After the meta-training process on a series of meta-training tasks $\{\mathcal{T}_1, \mathcal{T}_2, \ldots, \mathcal{T}_T\}$, we can obtain a dictionary of candidate auxiliary sets in $\mathcal{A}_{can} = \{\mathcal{A}_1, \mathcal{A}_2, \ldots, \mathcal{A}_{|\mathcal{A}_{can}|}\}$ along with their fairness adaptation directions as keys. Here we denote their corresponding keys as $\mathbf{k}(\mathcal{A}) \in \mathbb{R}^{d_\theta}$. Then given a new meta-test task $\mathcal{T}_{\text{test}} = \{\mathcal{S}_{\text{test}}, \mathcal{Q}_{\text{test}}\}$, the corresponding selected auxiliary set $\mathcal{A}^*$ can be selected via the following criterion:

$$\mathcal{A}^* = \operatorname*{argmin}_{\mathcal{A}\in\mathcal{A}_{can}} \text{dist}\left(g(\mathcal{S}_{\text{test}}), \mathbf{k}(\mathcal{A})\right), \tag{11}$$

where $\text{dist}(\cdot, \cdot)$ is a function to measure the distance between two vectors. In the experimentation, we implement it as the Euclidean distance. We can then efficiently select an auxiliary set from a significantly large dictionary based on the keys. It is noteworthy that to keep consistency between meta-training and meta-test, we will also select an auxiliary set for each meta-training task for optimization.

3.3 Meta-optimization

Our framework is optimized under the episodic meta-learning paradigm [18]. Specifically, let θ denote the total parameters of the

[2] Note that the auxiliary set size is controllable via randomly removing samples in $\mathcal{S}$ or incorporating new samples before enqueuing.

classification model $f(\cdot)$. In order to perform fairness adaptation, we first initialize the model parameters as $\theta_0 \leftarrow \theta$. After that, given a specific meta-task $\mathcal{T} = \{\mathcal{S}, \mathcal{Q}\}$, we conduct τ steps of gradient descent based on the fairness adaptation loss $\mathcal{L}_{FA}$ calculated on the support set $\mathcal{S}$. Thus, the fairness adaptation process in $\mathcal{T}$ can be formulated as follows:

$$\theta_t \leftarrow \theta_{t-1} - \alpha \nabla_{\theta_{t-1}} \mathcal{L}_{FA}\left(\mathcal{S}; \theta_{t-1}\right), \tag{12}$$

where $t \in \{1, 2, \ldots, \tau\}$ and $\mathcal{L}(\mathcal{S}; \theta_{t-1})$ denotes the loss calculated based on the support set $\mathcal{S}$ with the parameters θ_{t-1}. τ is the number of fine-tuning steps applied, and α is the learning rate in each fine-tuning step. After conducting τ steps of fine-tuning, we will meta-optimize the classification model $f(\cdot)$ with the loss calculated on the query set $\mathcal{Q}$. In specific, we meta-optimize the model parameters θ with the following update function:

$$\theta =: \theta - \beta_1 \nabla_\theta \mathcal{L}_{FA}(\mathcal{Q}; \theta_\tau), \tag{13}$$

where β_1 is the meta-learning rate for the classification model $f(\cdot)$.

For the optimization of the generator $g(\cdot)$, parameterized by ϕ, the update can be formulated as follows:

$$\phi =: \phi - \beta_2 \nabla_\phi \mathcal{L}_E(\mathcal{S}; \theta_\tau), \tag{14}$$

where $\mathcal{L}_E$ is the MSE loss introduced in Eq. (9), and β_2 is the meta-learning rate for the generator $g(\cdot)$. In this way, the model parameters ϕ of $g(\cdot)$ will be updated based on loss $\mathcal{L}_E$ after the fairness adaptation of the classification model $f(\cdot)$. The detailed training process of our framework is demonstrated in Algorithm 1.

4 Experimental Evaluations

4.1 Datasets

In this subsection, we introduce the datasets used in our experiments. To evaluate the performance of FEAST on fair few-shot learning, we conduct experiments on three prevalent real-world datasets: Adult [15], Crime [22], and Bank [26]. The detailed dataset statistics are provided in Table 1.

- The Adult dataset contains information from 48,842 individuals from the 1994 US Census, where each instance is represented by 14 features and a binary label. Here the label indicates whether the income of a person is higher than 50K dollars. Following the data split setting in PDFM [49], we split the dataset into 34 subsets based on the country information of instances. We consider gender as the sensitive attribute.
- The Crime dataset includes information on 2,216 communities from different states in the U.S., where each instance consists of 98 features. Following [31], the binary label of each instance is obtained by converting the continuous crime rate based on whether the crime rate of a community is in the top 50% within the state. The sensitive attribute is whether African-Americans are among the highest or second highest populations in each community. We further split this dataset into 46 subsets by considering each state as a subset.
- The Bank dataset consists of 41,188 individual instances in total. Specifically, each instance maintains 20 features along with a binary label that indicates whether the individual has subscribed to a term deposit. Here, we consider marital status as the binary sensitive attribute. Moreover, the dataset is split into 50 subsets based on the specific date records of instances.

Table 1: Statistics of three real-world datasets.

Dataset	Adult	Crime	Bank
Sensitive Attribute	Gender	Race	Marital Status
Label	Income	Crime Rate	Deposit
# Instances	48,482	2,216	41,188
# Features	12	98	17
# Subsets	34	46	50
# Training Subsets	22	30	40
# Validation Subsets	6	8	5
# Test Subsets	6	8	5

4.2 Experimental Settings

To achieve a fair comparison of FEAST with competitive baselines, we conduct experiments with the state-of-the-art fair few-shot learning methods and other few-shot learning methods with fairness constraints. The details are provided below.

- MAML [18]: This method utilizes a classic meta-learning framework to deal with the fair few-shot learning problem without explicitly applying fairness constraints.
- M-MAML [18]: This method uses the same framework as MAML while modifying datasets by removing the sensitive attribute of each instance to enhance fairness during optimization.
- Pretrain [49]: This method learns a single model on all meta-training data without episodic training. Moreover, a fairness constraint is added to the training objective.
- F-MAML [50]: This method applies a fairness constraint in each episode and tunes a Lagrangian multiplier shared across different episodes for fair few-shot learning tasks.
- FM-dp and FM-eop (Fair-MAML) [31]: These two baselines provide a regularization term for each episode based on demographic parity (DP) and equal opportunity (EOP), respectively.
- PDFM [49]: This method leverages a primal-dual subgradient approach to ensure that the learned model can be fast adapted to a new episode in fair few-shot learning.

Particularly, we use the average classification accuracy (ACC) over T_{test} meta-test tasks to evaluate the prediction performance. For fairness performance, we propose to utilize demographic parity (DP) and equalized odds (EO), which are commonly used in existing works [8, 48, 16, 44]. Since we consider the binary classification datasets, the output $f(x) \in \mathbb{R}$ denotes the prediction score of a specific sample x. In this manner, the metrics can be calculated over T_{test} meta-test tasks sampled from the meta-test task distribution P_{te} as follows:

$$\Delta\text{DP} = \mathbb{E}_{\mathcal{T}\sim P_{te}} \left| \frac{1}{|\mathcal{Q}_0|} \sum_{x \in \mathcal{Q}_0} f(x) - \frac{1}{|\mathcal{Q}_1|} \sum_{x \in \mathcal{Q}_1} f(x) \right|, \tag{15}$$

$$\Delta\text{EO} = \mathbb{E}_{\mathcal{T}\sim P_{te}} \sum_{y \in \{0,1\}} \left| \frac{1}{|\mathcal{Q}_0^y|} \sum_{x \in \mathcal{Q}_0^y} f(x) - \frac{1}{|\mathcal{Q}_1^y|} \sum_{x \in \mathcal{Q}_1^y} f(x) \right|, \tag{16}$$

where $\mathcal{Q}_0$ and $\mathcal{Q}_1$ denote the query samples with a sensitive attribute of 0 and 1, respectively. Similarly, $\mathcal{Q}_0^y$ (or $\mathcal{Q}_1^y$) denotes the query samples in $\mathcal{Q}_0$ (or $\mathcal{Q}_1$) with label y. P_{te} is the meta-test task distribution of meta-test sets $\mathcal{D}^n$. Our code is released at https://github.com/SongW-SW/FEAST.

Table 2: Results w.r.t. fairness and prediction performance of FEAST and baselines under different settings for all three datasets.

Dataset	Adult						Crime						Bank					
Setting	5-shot			10-shot			5-shot			10-shot			5-shot			10-shot		
Metric	ΔDP	ΔEO	ACC	ΔDP	ΔEO	ACC	ΔDP	ΔEO	ACC	ΔDP	ΔEO	ACC	ΔDP	ΔEO	ACC	ΔDP	ΔEO	ACC
MAML	0.473	0.706	0.801	0.409	0.584	**0.886**	0.558	0.952	0.718	0.443	0.832	0.792	0.214	0.573	**0.603**	0.185	0.496	0.619
M-MAML	0.447	0.689	**0.826**	0.381	0.555	0.857	0.359	0.732	0.711	0.300	0.569	0.757	0.214	0.544	0.600	0.175	0.459	0.619
F-MAML	0.339	0.432	0.825	0.310	0.353	0.840	0.503	0.871	0.719	0.463	0.707	0.762	0.207	0.585	0.575	0.181	0.528	**0.650**
FM-dp	0.313	0.502	0.814	0.241	0.438	0.844	0.385	0.722	0.741	0.329	0.604	0.771	0.238	0.614	0.586	0.187	0.553	0.604
FM-eop	0.430	0.703	0.812	0.370	0.601	0.846	0.352	0.706	0.739	0.311	0.591	0.804	0.289	0.683	0.581	0.245	0.600	0.640
Pretrain	0.365	0.513	0.806	0.310	0.450	0.885	0.390	0.692	**0.746**	0.354	0.582	0.776	0.248	0.659	0.594	0.208	0.539	0.642
PDFM	0.261	0.461	0.815	0.276	0.401	0.869	0.402	0.784	0.722	0.325	0.669	**0.816**	0.210	0.585	0.589	0.180	0.493	0.645
FEAST	**0.258**	**0.355**	0.820	**0.235**	**0.256**	0.861	**0.203**	**0.309**	0.739	**0.164**	**0.217**	0.797	**0.190**	**0.524**	0.583	**0.154**	**0.414**	0.641

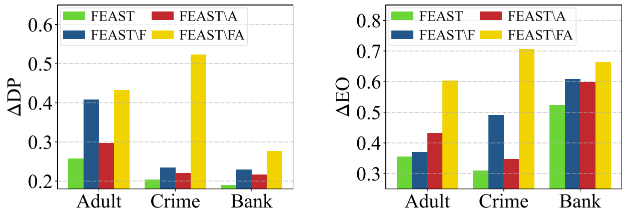

Figure 2: Ablation study on our framework FEAST on three datasets under the 5-shot setting.

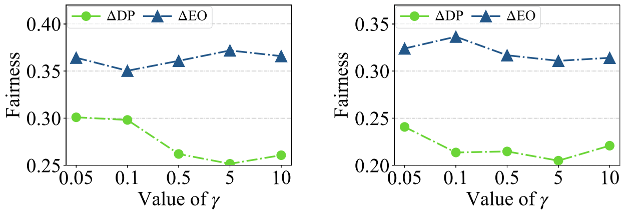

Figure 3: Results of FEAST on Adult (left) and Crime (right) with different values of γ.

4.3 Performance Comparison

Table 2 presents the fairness and prediction performance comparison of FEAST and all other baselines on fair few-shot learning. Specifically, we report the results of ΔDP, ΔEO, and classification accuracy over 500 meta-test tasks for 10 repetitions. We conduct experiments on both 5-shot and 10-shot settings (i.e., $K = 5$ and $K = 10$). From Table 2, we can have following observations:

- Our framework FEAST consistently outperforms other baselines in terms of fairness in all datasets under both 5-shot and 10-shot settings. These results provide compelling evidence for the effectiveness of our framework FEAST in fair few-shot learning.
- The performance improvement of FEAST over other baselines is more significant on the Crime dataset. This is due to that in this dataset, each subset consists of fewer samples. Consequently, the learned fairness-aware meta-knowledge will be more difficult to be transferred in baselines. Nevertheless, our proposed fairness adaptation strategy based on mutual information can effectively deal with this scenario.
- The accuracy of FEAST is comparable with other baselines, demonstrating that FEAST can substantially reduce biases without sacrificing its classification capability. This is because our framework FEAST can select the auxiliary set with similar fairness adaptation directions and thus will not harm model performance regarding accuracy.
- FEAST is more robust to the changes of the number of support samples per class, i.e., when the number decreases from 10 to 5, FEAST has the least performance drop in comparison to other baselines. We believe this is primarily because, with fewer support samples, the problem of insufficient samples becomes more significant. Nevertheless, FEAST can effectively address this issue with the incorporation of auxiliary sets into fairness adaptation.

4.4 Impact of Each Component in FEAST

In this subsection, we conduct an ablation study on three datasets under the 5-shot setting to evaluate the effectiveness of different components in our framework by comparing FEAST with three degenerate versions: (1) FEAST without fairness adaptation based on MI, referred to as FEAST\F. In this variant, the fairness adaptation process is simplified such that only fairness constraints are applied. (2) FEAST without auxiliary set selection, i.e., the auxiliary set is randomly sampled. We refer to this variant as FEAST\A. (3) FEAST without both fairness adaptation and auxiliary set selection, referred to as FEAST\FA. The results, as presented in Fig. 2, show that FEAST outperforms all other variants, validating the importance of both fairness adaptation and auxiliary set selection components in fair few-shot learning. Of particular interest is that the removal of the MI fairness adaptation has a more significant adverse impact on the Crime dataset, which contains significantly fewer meta-training samples. This result highlights the crucial role of this component in addressing the issue of insufficient training samples. In addition, when the two components are both removed, the fairness performance drops greatly. Such results indicate that the mutual impact brought by these two components is also critical for our proposed framework FEAST.

4.5 Effect of Loss Weight γ

Given the significance of the auxiliary sets in the fairness adaptation, in this subsection, we further examine in-depth how the auxiliary sets will influence the performance of FEAST. Specifically, we vary the value of γ, which controls the importance of the auxiliary set loss during fairness adaptation. A higher value of γ implies a larger importance weight on the auxiliary set and a smaller importance weight on the target task. Due to the limitation of space, we

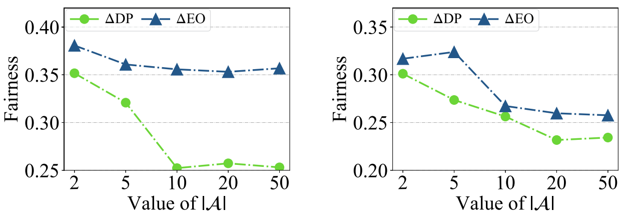

Figure 4: Results of FEAST on Adult under 5-shot (left) and 10-shot (right) settings with different values of $|\mathcal{A}|$.

evaluate the model's performance on two datasets, Adult and Crime, using various values of γ (similar results on the Bank dataset) on the 5-shot setting. The results, as shown in Fig. 3, indicate that a value around 0.5 for γ generally yields better fairness performance for both datasets. This is mainly because a small γ can be insufficient to leverage the fairness-aware meta-knowledge in auxiliary sets, while an excessively large value of γ can result in the loss of crucial fairness information in the target meta-task. Moreover, the effect of different γ values is more significant on the Adult dataset. The reason is that this dataset contains a larger number of samples in meta-training data. As a result, the learned fairness-aware knowledge is richer in the auxiliary sets, thus propagating the benefits from auxiliary sets.

4.6 *Effect of Auxiliary Set Size*

In this section, we conduct experiments to evaluate the impacts brought by varying the size of the auxiliary set $\mathcal{A}$. Intuitively, the auxiliary set size $|\mathcal{A}|$ should be at least comparable with the support set, since an excessively small auxiliary set can be potentially insufficient for fairness adaptation. Specifically, we conduct experiments on dataset Adult under both 5-shot and 10-shot settings to evaluate the effect of auxiliary set size $|\mathcal{A}|$. From the results presented in Fig. 4, we can make the following observations: (1) The fairness results are less satisfactory with a smaller value of $|\mathcal{A}|$, indicating that the capacity of $\mathcal{A}$ can be important in FEAST. With a small auxiliary set $\mathcal{A}$, the fairness adaptation effect will be reduced due to insufficient knowledge in $\mathcal{A}$. (2) When further increasing the size of $\mathcal{A}$, the fairness performance does not accordingly increase. This demonstrates that knowledge in a larger auxiliary set may not be helpful for fairness adaptation. (3) When the number of shots increases from 5 to 10, the best value of $|\mathcal{A}|$ also increases, implying that with a larger support set, the auxiliary set should also be expanded to provide more knowledge for fairness adaptation. In consequence, the fairness performance can be further improved.

5 Related Work

5.1 *Few-shot Learning*

Few-shot learning aims to obtain satisfactory classification performance with only a few labeled samples as references [37, 36]. The typical approach is to accumulate transferable knowledge from meta-training tasks, which contain abundant labeled samples. Then such knowledge is generalized to meta-test tasks with limited labeled samples. Particularly, existing few-shot learning methods can be divided into two main categories: (1) *Metric-based* methods propose to learn a metric function that matches samples in the query set with the support samples to conduct classification [23, 34, 42, 41]. For example, Prototypical Networks [32] learn a prototype (i.e., the average embedding of samples in the same class) for each class and then classify query samples according to the Euclidean distances between query samples and each prototype. Matching Networks [39] output predictions for query samples via the similarity between query samples and each support sample. (2) *Optimization-based* methods aim to first fine-tune model parameters based on gradients calculated on support samples and then conduct meta-optimization on each meta-task [25, 28, 43, 40]. As a classic example, MAML [18] learns a shared model parameter initialization for various meta-tasks with the proposed meta-optimization strategy. LSTM-based meta-learner [28] proposes an adjustable step size to update model parameters.

5.2 *Fairness-aware Machine Learning*

Various fairness-aware algorithms have been proposed to mitigate the unwanted bias in machine learning models. Generally, there are two categories of statistical fairness notions: *individual fairness* and *group fairness*. In particular, individual fairness requires that the model results for similar individuals should also be similar [16, 44, 13, 12]. Here, the similarity between individuals can be measured via specific metrics (e.g., Euclidean distance) learned during training or from prior knowledge. On the other hand, group fairness refers to the statistical parity between subgroups (typically defined by sensitive attributes, e.g., gender and race) via specific algorithms [46, 24, 19, 14]. Common fairness learning tasks include fair classification [45, 17], regression [2, 5], and recommendations [30]. Although these methods have demonstrated satisfactory performance in mitigating unfairness, it is noteworthy that existing works mainly focus on the settings where sufficient labeled samples are provided. As a result, it is challenging for these methods to accommodate few-shot scenarios with limited labeled samples.

More recently, several methods are proposed to deal with the fair few-shot learning problem [31, 50]. For example, PDFM [49] utilizes a primal-dual subgradient approach to ensure fast adaptation to a novel meta-task. In [48], the authors propose to address fairness in supervised few-shot meta-learning models that are sensitive to discrimination in historical data by detecting and controlling the dependency effect of sensitive attributes on target prediction. Moreover, F-MAML [50] provides a fairness constraint for each episode and tunes a Lagrangian multiplier shared across different episodes based on a meta-learning mechanism. However, these methods cannot effectively solve the problem of insufficient samples and the generalization gap.

6 Conclusion

In this paper, we propose a novel problem of fair few-shot learning, which focuses on accurately and fairly predicting labels for samples in unseen data while using limited labeled samples as references. To tackle the challenges posed by insufficient samples and the generalization gap between meta-training and meta-test, we propose an innovative framework FEAST that utilizes learned fairness-aware meta-knowledge by incorporating auxiliary sets. In particular, our framework maximizes the mutual information between meta-tasks and the auxiliary sets to enhance fairness adaptation. Moreover, we select auxiliary sets based on the estimated fairness adaptation direction of meta-tasks to improve the fairness performance. We conduct extensive experiments on three real-world datasets, and the results validate the superiority of FEAST over the state-of-the-art baselines. For future work, it is important to consider expanding the candidate auxiliary set with external knowledge, since samples in the dataset can be insufficient. In this case, incorporating external information for fairness adaptation can be crucial.

7 Acknowledgements

The work in this paper is supported by the National Science Foundation under grants (IIS-2006844, IIS-2144209, IIS-2223769, CNS2154962, and BCS-2228534), the Commonwealth Cyber Initiative awards (VV-1Q23-007 and HV-2Q23-003), the JP Morgan Chase Faculty Research Award, the Cisco Faculty Research Award, the Jefferson Lab subcontract 23-D0163, and the UVA 4-VA collaborative research grant.

References

[1] Maria Barrett, Yova Kementchedjhieva, Yanai Elazar, Desmond Elliott, and Anders Søgaard, 'Adversarial removal of demographic attributes revisited', in *EMNLP*, (2019).

[2] Richard Berk, Hoda Heidari, Shahin Jabbari, Matthew Joseph, Michael Kearns, Jamie Morgenstern, Seth Neel, and Aaron Roth, 'A convex framework for fair regression', *arXiv:1706.02409*, (2017).

[3] Tolga Bolukbasi, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai, 'Man is to computer programmer as woman is to homemaker? debiasing word embeddings', in *NeurIPS*, (2016).

[4] Joy Buolamwini and Timnit Gebru, 'Gender shades: Intersectional accuracy disparities in commercial gender classification', in *FAccT*, (2018).

[5] Toon Calders, Asim Karim, Faisal Kamiran, Wasif Ali, and Xiangliang Zhang, 'Controlling attribute effect in linear regression', in *ICDM*, (2013).

[6] Lu Cheng, Ahmadreza Mosallanezhad, Yasin N Silva, Deborah L Hall, and Huan Liu, 'Bias mitigation for toxicity detection via sequential decisions', in *SIGIR*, (2022).

[7] Lu Cheng, Kush R Varshney, and Huan Liu, 'Socially responsible ai algorithms: Issues, purposes, and challenges', *JAIR*, (2021).

[8] Ching-Yao Chuang and Youssef Mroueh, 'Fair mixup: Fairness via interpolation', in *ICLR*, (2021).

[9] Sam Corbett-Davies, Emma Pierson, Avi Feller, Sharad Goel, and Aziz Huq, 'Algorithmic decision making and the cost of fairness', in *SIGKDD*, (2017).

[10] Andrew Cotter, Maya Gupta, Heinrich Jiang, Nathan Srebro, Karthik Sridharan, Serena Wang, Blake Woodworth, and Seungil You, 'Training well-generalizing classifiers for fairness metrics and other data-dependent constraints', in *ICML*, (2019).

[11] Jeffrey Dastin, 'Amazon scraps secret ai recruiting tool that showed bias against women', in *Ethics of Data and Analytics*, (2018).

[12] Yushun Dong, Jing Ma, Song Wang, Chen Chen, and Jundong Li, 'Fairness in graph mining: A survey', *TKDE*, (2023).

[13] Yushun Dong, Song Wang, Jing Ma, Ninghao Liu, and Jundong Li, 'Interpreting unfairness in graph neural networks via training node attribution', in *AAAI*, (2023).

[14] Yushun Dong, Song Wang, Yu Wang, Tyler Derr, and Jundong Li, 'On structural explanation of bias in graph neural networks', in *SIGKDD*, (2022).

[15] Dheeru Dua, Casey Graff, et al., 'Uci machine learning repository', (2017).

[16] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel, 'Fairness through awareness', in *ITCS*, (2012).

[17] Michael Feldman, Sorelle A Friedler, John Moeller, Carlos Scheidegger, and Suresh Venkatasubramanian, 'Certifying and removing disparate impact', in *SIGKDD*, (2015).

[18] Chelsea Finn, Pieter Abbeel, and Sergey Levine, 'Model-agnostic meta-learning for fast adaptation of deep networks', in *ICML*, (2017).

[19] Moritz Hardt, Eric Price, and Nati Srebro, 'Equality of opportunity in supervised learning', in *NeurIPS*, (2016).

[20] Kexin Huang and Marinka Zitnik, 'Graph meta learning via local subgraphs', in *NeurIPS*, (2020).

[21] Anja Lambrecht and Catherine Tucker, 'Algorithmic bias? an empirical study of apparent gender-based discrimination in the display of stem career ads', *Management Science*, (2019).

[22] Moshe Lichman et al. Uci machine learning repository, 2013.

[23] Lu Liu, Tianyi Zhou, Guodong Long, Jing Jiang, and Chengqi Zhang, 'Learning to propagate for graph meta-learning', in *NeurIPS*, (2019).

[24] Christos Louizos, Kevin Swersky, Yujia Li, Max Welling, and Richard Zemel, 'The variational fair autoencoder', *arXiv:1511.00830*, (2015).

[25] Nikhil Mishra, Mostafa Rohaninejad, Xi Chen, and Pieter Abbeel, 'A simple neural attentive meta-learner', in *ICLR*, (2018).

[26] Sérgio Moro, Paulo Cortez, and Paulo Rita, 'A data-driven approach to predict the success of bank telemarketing', *Decision Support Systems*, (2014).

[27] Aaron van den Oord, Yazhe Li, and Oriol Vinyals, 'Representation learning with contrastive predictive coding', in *arXiv:1807.03748*, (2018).

[28] Sachin Ravi and Hugo Larochelle, 'Optimization as a model for few-shot learning', in *ICLR*, (2016).

[29] Soumajyoti Sarkar and Hamidreza Alvari, 'Mitigating bias in online microfinance platforms: A case study on kiva. org', in *ECMLPKDD*, (2020).

[30] Ashudeep Singh and Thorsten Joachims, 'Fairness of exposure in rankings', in *SIGKDD*, (2018).

[31] Dylan Slack, Sorelle A Friedler, and Emile Givental, 'Fairness warnings and fair-maml: learning fairly with minimal data', in *FAccT*, (2020).

[32] Jake Snell, Kevin Swersky, and Richard Zemel, 'Prototypical networks for few-shot learning', in *NeurIPS*, (2017).

[33] Megan Stevenson, 'Assessing risk assessment in action', *Minn. L. Rev.*, (2018).

[34] Flood Sung, Yongxin Yang, Li Zhang, Tao Xiang, Philip HS Torr, and Timothy M Hospedales, 'Learning to compare: relation network for few-shot learning', in *CVPR*, (2018).

[35] Latanya Sweeney, 'Discrimination in online ad delivery', *Communications of the ACM*, (2013).

[36] Zhen Tan, Song Wang, Kaize Ding, Jundong Li, and Huan Liu, 'Transductive linear probing: A novel framework for few-shot node classification', *arXiv:2212.05606*, (2022).

[37] Yonglong Tian, Yue Wang, Dilip Krishnan, Joshua B Tenenbaum, and Phillip Isola, 'Rethinking few-shot image classification: a good embedding is all you need?', in *ECCV*, (2020).

[38] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin, 'Attention is all you need', in *NeurIPS*, (2017).

[39] Oriol Vinyals, Charles Blundell, Timothy Lillicrap, Daan Wierstra, et al., 'Matching networks for one shot learning', in *NeurIPS*, (2016).

[40] Song Wang, Chen Chen, and Jundong Li, 'Graph few-shot learning with task-specific structures', in *NeurIPS*, (2022).

[41] Song Wang, Kaize Ding, Chuxu Zhang, Chen Chen, and Jundong Li, 'Task-adaptive few-shot node classification', in *SIGKDD*, (2022).

[42] Song Wang, Yushun Dong, Xiao Huang, Chen Chen, and Jundong Li, 'Faith: Few-shot graph classification with hierarchical task graphs', in *IJCAI*, (2022).

[43] Song Wang, Xiao Huang, Chen Chen, Liang Wu, and Jundong Li, 'Reform: Error-aware few-shot knowledge graph completion', in *CIKM*, (2021).

[44] Mikhail Yurochkin, Amanda Bower, and Yuekai Sun, 'Training individually fair ml models with sensitive subspace robustness', in *ICLR*, (2020).

[45] Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rodriguez, and Krishna P Gummadi, 'Fairness beyond disparate treatment & disparate impact: Learning classification without disparate mistreatment', in *WWW*, (2017).

[46] Rich Zemel, Yu Wu, Kevin Swersky, Toni Pitassi, and Cynthia Dwork, 'Learning fair representations', in *ICML*, (2013).

[47] Brian Hu Zhang, Blake Lemoine, and Margaret Mitchell, 'Mitigating unwanted biases with adversarial learning', in *AIES*, (2018).

[48] Chen Zhao and Feng Chen, 'Unfairness discovery and prevention for few-shot regression', in *ICKG*, (2020).

[49] Chen Zhao, Feng Chen, Zhuoyi Wang, and Latifur Khan, 'A primal-dual subgradient approach for fair meta learning', in *ICDM*, (2020).

[50] Chen Zhao, Changbin Li, Jincheng Li, and Feng Chen, 'Fair meta-learning for few-shot classification', in *ICKG*, (2020).

[51] Fan Zhou, Chengtai Cao, Kunpeng Zhang, Goce Trajcevski, Ting Zhong, and Ji Geng, 'Meta-gnn: On few-shot node classification in graph meta-learning', in *CIKM*, (2019).