Fair Weak-Supervised Learning: A Multiple-Instance Learning Approach

Yucong Dai[1], Xiangyu Jiang[1], Yaowei Hu[2], Lu Zhang[3], Yongkai Wu[1]
[1]*Electrical and Computer Engineering*, *Clemson University*, Clemson, SC, USA
E-mail: {yucongd, xiangyj, yongkaw}@clemson.edu
[2]*Walmart Inc*, Bentonville, AR, USA
E-mail: yaowei.hu@walmart.com
[3]*Computer Science & Computer Engineering*, *University of Arkansas*, Fayetteville, AR, USA
E-mail: lz006@uark.edu

2024 International Joint Conference on Neural Networks (IJCNN) | 979-8-3503-5931-2/24/$31.00 ©2024 IEEE | DOI: 10.1109/IJCNN60899.2024.10651225

***Abstract*—With the prevalence of machine learning in many high-stakes decision-making processes, e.g., hiring and admission, it is important to take fairness into account when practitioners design and deploy machine learning models, especially in scenarios with imperfectly labeled data. Multiple-Instance Learning (MIL) is a weakly supervised approach where instances are grouped in labeled bags, each containing several instances sharing the same label. However, current fairness-centric methods in machine learning often fall short when applied to MIL due to their reliance on instance-level labels. In this work, we introduce a Fair Multiple-Instance Learning (FMIL) framework to ensure fairness in weakly supervised learning. In particular, our method bridges the gap between bag-level and instance-level labeling by leveraging the bag labels, inferring high-confidence instance labels to improve both accuracy and fairness in MIL classifiers. Comprehensive experiments underscore that our FMIL framework substantially reduces biases in MIL without compromising accuracy.**

***Index Terms*—fairness, multiple-instance learning, weak supervised learning**

I. Introduction

Machine learning algorithms have been widely used in many high-stakes applications, such as hiring and banking. Since conventional machine learning models primarily aim to maximize predictive accuracy, they may incur or exacerbate unintended biases in these high-stakes scenarios. As bias in machine learning becomes a matter of concern, it is imperative to ensure that machine learning models can strike a balance between accuracy and fairness. Recently, many fairness-aware machine learning solutions and models have been proposed to mitigate unfairness [1]–[25], including data modification, model tweaking, and decision flipping, in the supervised learning paradigm. Although these methods have successfully balanced model performance and fairness, it is important to note that many tasks struggle to obtain strong supervision information, such as fully verified ground-truth labels, due to the high costs associated with data labeling and acquisition. Taking American Census Data [26] as an example, it is common that data are collected from a household perspective rather than from individuals. In contrast, there is a need to predict income at the individual level rather than the household level. It is challenging to predict individual income using household-based data due to the lack of individual-level ground truth. It is worth pointing out that a transition from conventionally strongly supervised learning to weakly supervised learning [27] may lead to existing bias detection and mitigation algorithms being ineffective or unable to maintain the proper balance between accuracy and fairness. Hence, it is imperative to develop a new fair machine-learning framework that can effectively mitigate bias in weakly supervised machine learning.

In this paper, we target multiple-instance machine learning, one of the most important tasks in weakly supervised learning, formulate a set of novel fairness metrics leveraging the inexact bag labels, and propose a general framework to mitigate adverse bias. The multiple-instance learning paradigm leverages inexact supervision from bag labels to develop a robust instance-level model in situations where only coarse-grained label information is available. The proposed bias mitigation framework for the MIL task focuses on two widely used metrics, *demographic parity* [28] and *equalized odds* [1]. We follow the in-processing bias mitigation paradigm and design fairness constraints that are effective in multiple-instance learning without instance labels. Inspired by *pseudo-labeling* [29], we exploit as much unlabeled instance data as possible to improve the accuracy and fairness of a classifier. Our experimental results demonstrate that FMIL maintains accuracy while achieving fairness, outperforming common baselines. Our main contributions are as follows.

- We propose a framework that takes fairness into consideration in multiple-instance learning settings. To the best of our knowledge, there has not been any work that aims to improve fairness in multiple-instance learning settings.
- We design novel fairness constraints that can mitigate bias when only coarse-grained labels are available. Integrating these constraints with the existing MIL learning paradigm, we develop bias mitigation solutions for weakly supervised learning.
- We provide empirical evidence showing the ability of the FMIL framework to sustain a high level of accuracy while ensuring fairness.

II. Related works

Fair machine learning has been studied extensively in the literature. A set of fairness notions and metrics have been proposed in the past years. Among them, *demographic parity* [28], [30] and *equalized odds* [1] which measure the dependencies between the decision and the sensitive information have been widely adopted. Based on the notions, three types of approaches [1], [7], [10], [14], [31], [32], *pre-processing*, *in-processing*, and *post-processing*, have been proposed to tackle bias in machine learning. Pre-processing methods focus on adjusting training data to eliminate bias before model training, such as re-sampling or re-weighing instances to ensure fairness in representation. *Massaging* [7] flips the labels of selected examples close to the decision boundary to eliminate discrimination. *Reweighting* [5] assigns weights to individuals to balance the majority and minority groups. *Preferential sampling* [33] resamples subgroups to make the dataset discrimination-free. Most of the *in-processing* approaches, e.g., [34]–[36], involve directly incorporating fairness constraints or objectives during the model training process, which may include modifying the learning algorithm itself. Those approaches are flexible in balancing the trade-off between model performance and fairness requirements. The *post-processing* approaches adjust the output of already trained models to correct for bias, typically by altering decision thresholds for different groups. Hardt et al. [1] develop an optimization solution to adjust any learned predictor to remove discrimination according to *equalized odds*. Kamiran et al. [37] exploit the low-confidence region of a single or an ensemble of probabilistic classifiers for discrimination reduction. Together, these strategies aim to enhance fairness and reduce discrimination in machine learning applications, with a trade-off between fairness and model performance.

The majority of fair machine learning literature focuses on supervised learning, where models are trained exclusively with precisely labeled data, which can be limiting and expensive to obtain. Semi-supervised learning (SSL) [38], [39] is a machine learning approach that bridges the gap between supervised and unsupervised learning by utilizing both labeled and unlabeled data. Translating fairness principles from supervised to semi-supervised learning involves adapting bias mitigation techniques to work with unlabeled data. A few works employ neural networks to optimize the trade-off between fairness and accuracy in a semi-supervised setting where only a subset of samples are labeled. Noroozi et al. [40] utilize *pseudo-labeling* to exploit unlabeled data to improve performance. Zhang et al. [41] leverage *pseudo-labeling* to predict labels for unlabeled data, divide the whole dataset into groups, and finally re-sampled within the groups.

Despite efforts at semi-supervised learning, fair weakly supervised learning has rarely been considered in the literature. Weakly supervised learning [42]–[44] is a branch of machine learning that deals with scenarios where the training data is imprecisely labeled. The goal is to develop models that can effectively leverage this imperfect information to make accurate predictions, enhancing the ability to work with less-than-ideal data while minimizing the labor-intensive process of manual labeling. In the weakly supervised learning setting, only partial supervision information is given while it is not exactly as desired. A typical scenario is multiple-instance learning where only coarse-grained label information is available. Thus, we cannot apply existing fairness-aware learning techniques directly to multiple-instance learning. This leaves a fundamental challenge – how to leverage the coarse-grained labels and alleviate the adverse bias in the MIL tasks. To tackle these challenges, we formulate the fair multiple-instance learning problem and propose an in-processing framework, namely FMIL, to minimize the objective with a fairness regularizer. To the best of our knowledge, FMIL is the first fairness-aware classification algorithm in the weakly supervised learning paradigm.

III. Preliminaries

We first introduce the classic fairness-aware classification that incorporates fairness constraints to eliminate statistical discrepancies between the prediction and the sensitive information. We present the key symbols used throughout the paper in Table I. Then, we formally introduce the multiple-instance learning problem.

Throughout the paper, the set of features is represented by $\mathbf{X}$ while the prediction attribute is denoted by Y. Z denotes sensitive information, such as gender, age, and race. Without loss of generality, we assume the sensitive information S and the decision Y as binary. The positive/favored value is denoted by 1, and the negative/unfavored value is denoted by -1. The binary setting can be readily extended to non-binary cases [45], [46]. The subscript after variables indicates the sample index. For example, $\mathbf{X}_i$ and Y_i represent the features and labels for i-th sample for the strongly supervised learning while $\mathbf{X}_{i,j}$ in the conventional strong supervised machine learning setting. In the weakly supervised learning setting, $X_{i,j}$ and $Y_{i,j}$ represent the instance features and labels for the j-th sample in the i-th bag in the MIL setting. Specially, the i-th bag label is denoted as $Y_{i,*}$. The lowercase letters with scripts represent the realization values. For example, $\mathbf{x}_{i,j}$ denotes the feature values of $\mathbf{X}_{i,j}$.

A. Fairness-aware classification

The learning goal of conventional strong supervised learning is to find a classifier $f : \mathcal{X} \to \mathcal{Y}$ from the training dataset $\mathcal{D}_s = \{\mathbf{X}_i, Z_i, Y_i\}$. In a traditional classification problem, we minimize the average of the classification errors (a.k.a the empirical loss) given by

$$\min_f \mathbb{E}_{\mathcal{D}_s}[\mathbb{1}_{f(\mathbf{X}_i)\neq Y_i}] \tag{1}$$

where $\mathbb{1}_{(\cdot)}$ is an indicator function. The indicator function can be replaced with various surrogate functions, such as the hinge function or the logit function [47].

Several fairness notions or definitions are proposed in the literature, such as *demographic parity* [28] and *equalized odds* [1]. Quantitative fairness constraints or regularization

Symbols	Definitions and Descriptions
i	the number of samples (bags in MIL)
j	the number of instances
$\{\mathbf{X}_i, Z_i, Y_i\}$	a strongly supervised dataset $\mathcal{D}_s$
$\mathbf{X}_i$	the features of i-th sample
Y_i	the label of i-th sample
Z_i	the sensitive attribute of i-th sample
$\{\mathbf{X}_{i,*}, Z_{i,*}, Y_{i,*}\}$	a weakly supervised dataset $\mathcal{D}_w$
$\mathbf{X}_{i,*}, x_{i,j}$	the features of j-th instance in the i-th bag/its realization values
$Y_{i,j}, y_{i,j}$	the label of j-th instance in the i-th bag/its realization values
$Y_{i,*}$	the label of the i-th bag
$Z_{i,j}, z_{i,j}$	the sensitive attribute of j-th instance in the i-th bag/its realization values
$f(\cdot)$	a classifier
$\mathbb{F}_{\mathbb{M}}[\cdot]$	the fairness constraints of certain metric $\mathbb{M}$

TABLE I
TABLE OF SYMBOLS AND NOTATIONS.

terms are designed to mitigate bias regarding those corresponding fairness notions. For example, the *demographic parity* can be quantitatively formulated as:

$$\mathbb{F}[\cdot] = \mathbb{E}_{\mathbf{X}_i|Z_i=1}[\mathbb{1}_{f(\mathbf{X}_i)=1}] - \mathbb{E}_{\mathbf{X}_i|Z_i=-1}[\mathbb{1}_{f(\mathbf{X}_i)=1}]. \quad (2)$$

In-processing bias mitigation algorithms incorporate the quantitative and differentiate fairness constraints or regularization terms into the objective function to ensure fairness in prediction, then we have:

$$\min_f \left\{ \alpha \mathbb{E}_{\mathcal{D}_s}\left[\mathbb{1}_{f(\mathbf{X}_i)\neq Y_i}\right] + (1-\alpha)\mathbb{F}\left[f(\mathbf{X}_i), Z_i, Y_i\right] \right\} \quad (3)$$

where $\mathbb{F}[\cdot]$ indicates the fairness loss, which imposes fairness on the prediction of the model. α indicates the penalized magnitude that controls the trade-off between fairness and classification loss.

B. Multiple-instance learning

We start with a weakly-labeled dataset $\mathcal{D}_w = \left\{\mathbf{X}_{i,*}, Z_{i,*}, Y_{i,*}\right\}_{i=1}^{N}$ where each element consists of a bag of m_i instances such that $\mathbf{X}_{i,*} = \left\{\mathbf{X}_{i,j}\right\}_{j=1}^{m_i}$ and $Z_{i,*} = \left\{Z_{i,j}\right\}_{i=1}^{m_i}$. It is worth noting that there is only a single label $Y_{i,*}$ associated with the i-th bag in the dataset. In other words, the instance labels $Y_{i,j}$ are unavailable during training. Additionally, it is common to assume that a bag is positive, i.e., $Y_{i,*} = +1$, if there exists at least one instance that is positive. We can formulate the assumptions of the MIL problem in the following form:

$$Y_{i,*} = \begin{cases} -1, & \forall j \in \{1, 2, \ldots, m_i\}, Y_{i,j} = -1 \\ +1, & otherwise \end{cases} \quad (4)$$

Further, we can reformulate it using the maximum operator:

$$Y_{i,*} = \max_j \{Y_{i,j}\} \quad (5)$$

Multiple-instance learning aims to learn an instance-based classifier $Y_{i,j} = h(\mathbf{X}_{i,j})$ from the dataset $\mathcal{D}_w$ organized by bags. There is a challenge for applying Eq. (3) to the training dataset $\mathcal{D}_w$ where the instance labels are missing. A straightforward approach to infer the instance label $Y_{i,j}$ from the bag $Y_{i,*}$ is to assign the instance labels $Y_{i,j} = Y_{i,*}, \forall j = \{1, 2, \ldots, m_i\}$. This process is referred to as **A**ssign **A**s **B**ag labels (**AAB**). Blum and Kalai [48] described a reduction from multiple-instance examples to PAC-learning with one-side random classification noise. The basic idea of this method is to consider all instances from a negative bag as negative but randomly choose one sole instance as positive from a positive bag. This method is referred to as **A**ssign **A**t **R**andom (**AAR**).

C. Attention-based multiple-instance learning

The aforementioned approaches that adhere to the premise in Eq. (5) exhibit a clear disadvantage, namely, that the `max` operator is non-trainable and inconvenient to incorporate with deep neural networks. Ilse et al. [49] propose a fully trainable MIL pooling based on the attention mechanism. Especially, the bag labels are considered a weighted aggregation of their instance labels. To model the weighted aggregation, Ilse et al. let $\mathbf{H}_i = \{\mathbf{h}_{i,j} | j \in 1, ..., m_i\}$ denote the instance representation of the i-th bag. Then, they formulate the instance-bag connection by the weighted average operator:

$$\mathbf{R}_i = \sum_{j=1}^{m_i} \mathbf{a}_{i,j}\mathbf{h}_{i,j}, \quad (6)$$

and the weights are derived from the attention mechanism:

$$\mathbf{a}_{i,j} = \frac{\exp\left\{w^\top \tanh\left(V\mathbf{h}_{i,j}^\top\right)\right\}}{\sum_{j=1}^{m_i}\exp\left\{w^\top \tanh\left(V\mathbf{h}_{i,j}^\top\right)\right\}}, \quad (7)$$

where $w \in \mathbb{R}^{L\times 1}$ and $V \in \mathbb{R}^{L\times D}$ are trainable parameters, D is the dimension of the hidden representation $\mathbf{h}_{i,j}$, and L is a hyperparameter that controls the dimensions of hidden space of the attention mechanism. Then the aggregated representation $\mathbf{R}_i$ is utilized to predict the label of the i-th bag. The hyperbolic tangent $\tanh(\cdot)$ ensures both negative and positive values for proper gradient flow.

IV. Fair Multiple-instance Learning

Fair multiple-instance learning aims to find an instance-based classifier that minimizes the empirical loss at the bag level while satisfying certain fairness constraints at the instance level. The objective function for FMIL can be reformulated as follows:

$$\min_h \left\{ \alpha \mathbb{E}_{\mathcal{D}_w}\left[\mathbb{1}_{\max\left(h(\mathbf{X}_{i,j})\right)\neq Y_{i,*}}\right] + (1-\alpha)\mathbb{F}\left[h(\mathbf{X}_{i,j}), Z_{i.j}, Y_{i,j}\right] \right\} \quad (8)$$

Since the instance-based labels $Y_{i,j}$ are unavailable in the MIL setting, it is not trivial to construct a bias regularizer $\mathbb{F}[\cdot]$ to achieve fairness in MIL. In this section, we propose new formulations for existing fairness constraints tailored for MIL and refer to the proposed framework as FMIL.

A. Achieve demographic parity in MIL

Demographic parity requires the decision made by the classifier to be independent of the sensitive attribute, which is quantified with regard to risk difference, i.e., the difference of the positive predictions between the favorable group and non-favorable group. The risk difference fairness constraint for the MIL setting is expressed as:

$$\mathbb{F}_{\mathbb{RD}}[\cdot] = \mathbb{E}_{\mathbf{X}_{i,j}|Z_{i,j}=1}[\mathbb{1}_{f(\mathbf{X}_{i,j})=1}] - \mathbb{E}_{\mathbf{X}_{i,j}|Z_{i,j}=-1}[\mathbb{1}_{f(\mathbf{X}_{i,j})=1}] \tag{9}$$

It follows that

$$\begin{aligned}\mathbb{F}_{\mathbb{RD}}[\cdot] =& \mathbb{E}_{\mathbf{X}_{i,j}}\Bigg[\frac{P\left(Z_{i,j}=1|\mathbf{X}_{i,j}\right)}{P\left(Z_{i,j}=1\right)}\mathbb{1}_{h(\mathbf{X}_{i,j})=1} \\ &+ \frac{P\left(Z_{i,j}=-1|\mathbf{X}_{i,j}\right)}{P\left(Z_{i,j}=-1\right)}\mathbb{1}_{h(\mathbf{X}_{i,j})=-1} - 1\Bigg]\end{aligned} \tag{10}$$

The proposed risk difference constraint in Eq. (10) can be applied to the joint objective function Eq. (8) directly without requiring the instance label $Y_{i,j}$.

B. Achieve equalized odds in MIL

Equalized odds requires that the sensitive attributes and the predicted labels are independent conditions on the true label. The challenge here is that the instance labels $Y_{i,j}$ are not available in the training dataset. Despite **AAB** and **AAR** discussed in the aforementioned preliminaries, the induced errors and noise lead to significant performance decreases. Inspired by the pseudo-label approach [29], we design a confidence-based method to tackle the errors and noise. For negative bags, we can consider that all the instances are negative. Thus the equalized odds for instances from the negative bags is formulated as:

$$\begin{aligned}\mathbb{F}_{\mathbb{EO}}^{neg} = \mathbb{E}_{\mathbf{X}_{i,j}|Y_{i,j}=-1}[&\frac{P(Z_{i,j}=1|\mathbf{X}_{i,j},Y_{i,j}=-1)}{P(Z_{i,j}=1|Y_{i,j}=-1)}\mathbb{1}_{h(\mathbf{X}_{i,j})=1} \\ &+ \frac{1-P(Z_{i,j}=1|\mathbf{X}_{i,j},Y_{i,j}=-1)}{1-P(Z_{i,j}=1|Y_{i,j}=-1)}\mathbb{1}_{h(\mathbf{X}_{i,j})=-1}] - 1\end{aligned} \tag{11}$$

For positive bags, the instance labels are derived from highly confident predictions.

$$Y_{i,j} = \begin{cases} 1, & sigmoid\big(h(\mathbf{X}_{i,j})\big) \geq \lambda \\ -1, & sigmoid\big(h(\mathbf{X}_{i,j})\big) \leq (1-\lambda) \\ 0, & otherwise \end{cases} \tag{12}$$

where λ indicates a threshold for the degree of confidence. Therefore, the equalized odds for instances from the positive bags is defined as:

$$\begin{aligned}\mathbb{F}_{\mathbb{EO}}^{pos} = Y_{i,j}^2\Bigg\{\mathbb{E}_{\mathbf{X}_{i,j}|Y_{i,j}}[&\frac{P(Z_{i,j}=1|\mathbf{X}_{i,j},Y_{i,j})}{P(Z_{i,j}=1|Y_{i,j})}\mathbb{1}_{h(\mathbf{X}_{i,j})=1} \\ &+ \frac{1-P(Z_{i,j}=1|\mathbf{X}_{i,j},Y_{i,j})}{1-P(Z_{i,j}=1|Y_{i,j})}\mathbb{1}_{h(\mathbf{X}_{i,j})=-1}] - 1\Bigg\}\end{aligned} \tag{13}$$

The proposed risk difference constraint in Eq. (11) and Eq. (13) can be applied to the joint objective function Eq. (8) directly without requiring the true instance label $Y_{i,j}$. The integration is illustrated in Algorithm 1.

Algorithm 1 FMIL EO algorithm

Input: $\mathcal{D}_w = \left\{\mathbf{X}_{i,*}, Z_{i,*}, Y_{i,*}\right\}_{i=1}^{N}$, λ, $\alpha, Epoches$
Output: Parameters θ of the instance classifier h
for $t := 1$ **to** $Epoches$ **do**
  for $i = 1$ **to** N **do**
    for $j = 1$ **to** m_i **do**
      if $Y_i = 1$ **then**

$$Y_{i,j} = \begin{cases} 1, & sigmoid\big(h(\mathbf{X}_{i,j})\big) \geq \lambda \\ -1, & sigmoid\big(h(\mathbf{X}_{i,j})\big) \leq (1-\lambda) \\ 0, & otherwise \end{cases}$$

      else
        $Y_{i,j} = Y_i$
      end if
    end for
  end for
  $\nabla_1(\theta) = \nabla_\theta\Big\{\alpha\mathbb{E}_{\mathcal{D}_s}\big[\mathbb{1}_{f(\mathbf{X}_i)\neq Y_i}\big]\Big\}$
  $\nabla_2(\theta) = \nabla_\theta\Big\{(1-\alpha)\mathbb{F}\big[f(\mathbf{X}_i), Z_i, Y_i\big]\Big\}$
  $\theta^t = \theta^{t-1} - \big[\nabla_1(\theta) + \nabla_2(\theta)\big]$
end for

V. Experiment

We evaluate the proposed framework and several baselines in various MIL settings using several real-world datasets. We conduct extensive experimental results showing our method outperforms other methods with better trade-offs between accuracy and fairness.

A. Dataset

We evaluate the baselines and proposed methods on two real-world datasets. The **Adult** [26] dataset contains $48,842$ instances, including 14 features, such as age, sex, household, and education. The goal of the Adult dataset is to predict whether an individual earns more than $50,000$ US dollars a year. We consider `sex` as the sensitive attribute and `income` as the target attribute. The **ACSIncome** [50] datasets is an improved alternative to the Adult dataset. The **ACSIncome** dataset consists of $1,664,500$ samples from 2014-2018 for all 50 U.S. states and Puerto Rico and contains 10 features, including age, class of worker, educational attainment, marital status, occupation, etc. In this dataset, we consider `race` as the sensitive attribute with two values. $white$ and $black$. For both datasets, we construct bags based on their similarities and create the bag labels according to the `max` mechanism.

B. Experimental setting

We implement three baseline methods and compare them with the proposed method. We compare these methods with regard to risk difference and equalized odds separately. The three conventional baselines are trained as follows.

- **FReg**: a classic fair classifier [4] is trained on strongly supervised data where the instance labels are available.
- **FAAB**: a classic fair classifier is trained on the bag labels but the instance labels are derived using the **AAB** method.

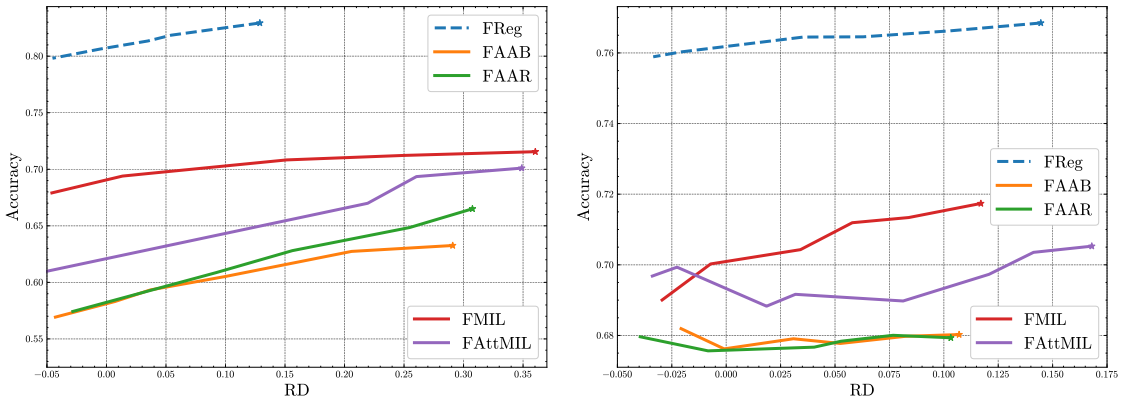

Fig. 1. Trade-off between accuracy and risk difference in **Census Adult** (left) and **ACSIncome** (right).

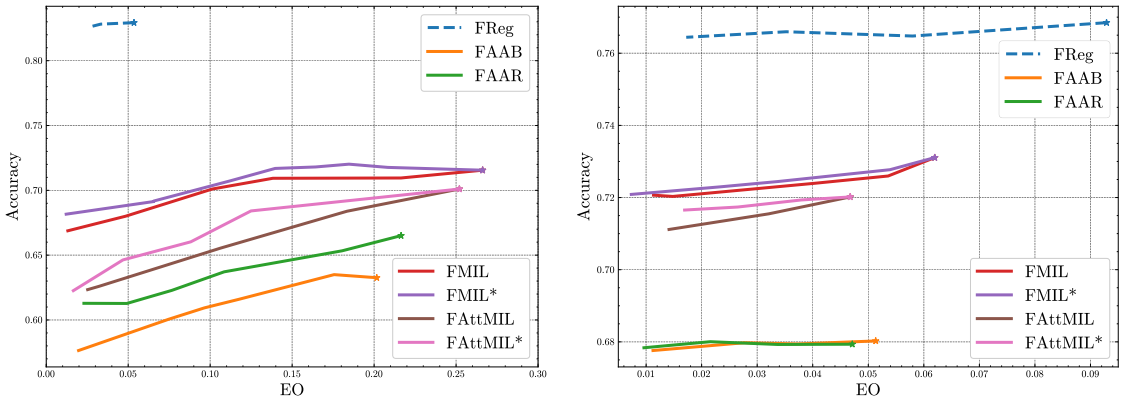

Fig. 2. Trade-off between accuracy and equalized odds in **Census Adult** and **ACSIncome**.

- **FAAR**: a classic fair classifier is trained on the bag labels but the instance labels are derived using the **AAR** method.

For the fair multiple-instance learning framework, we integrate the proposed fairness constraints into the MIL and attention-based MIL methods, namely **FMIL** and **FAttMIL**.

C. Experimental Result

1) Trade-off between accuracy and fairness.: We evaluate the model performance in terms of fairness and accuracy. For a fair comparison, we tune the hyper-parameters α for each approach from $\alpha = 1$ (where the bias is maximal as there is no fairness penalty) until the bias is -0.05 (which is the minimal legal requirement for fairness). We compare their accuracy performance with varying fairness levels. The results for achieving **RD** are shown in Fig. 1. It shows that the proposed methods, **FMIL** and **FAttMIL**, can effectively mitigate adverse bias without true instance-level labels while maintaining better accuracy, compared with two baseline methods **FAAR** and **FAAB** where the instance-level labels are inferred from bag-level labels. The results for achieving **EO** are shown in Fig. 2. It shows that the proposed methods, **FMIL** and **FAttMIL**, outperform baselines. In addition, we implement the confidence-based method to infer the true labels. To promote the performance of the predictors used for confidence calculation, we initially train the predictor for a few epochs without fairness regularizers. After the classifier has demonstrated a certain level of discrimination, the fairness constraint will be applied to the rest of the training process. As shown in Fig. 2, the refined methods (**FMIL*** and **FAttMIL***) can achieve higher accuracy with similar fairness levels. These results also show the flexibility of the proposed framework in dealing with the bias induced in the training process. These results indicate that the utilization of FMIL in mitigating bias by leveraging bag-level data is effective, as evidenced by the consistency of the results across the two fairness metrics risk difference (**RD**) and equalized odds (**EO**).

2) Training efficiency.: In our exploration of the performance efficiency of various methods, we specifically evaluated the training efficiency of each method. As depicted in Fig. 3, the **FAAB** and **FAAR** methods demonstrate an ability to reach fair predictions. However, their performance plateau suggests limitations in achieving higher accuracy for classifying instance labels. In stark contrast, **FMIL** and **FAttMIL**, our proposed methods, consistently outperform the other methods regarding accuracy. Their higher peaks in accuracy and lower **RD** values indicate a more robust and reliable classification capability. Notably, even though they exhibit a faster divergence trend post the 40th epoch, their overall performance trajectory remains superior. This underscores the efficacy of our proposed frameworks in delivering both efficient and

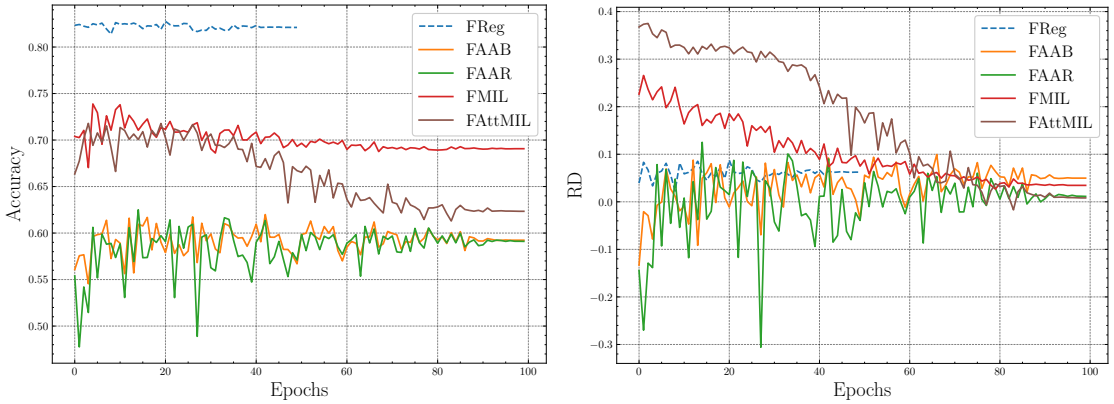

Fig. 3. Accuracy and bias of **FMIL** and baseline methods v.s. the training iterations in **Census Adult** and **ACSIncome**.

effective results.

VI. Conclusion

In this study, we investigate the algorithmic fairness problem in multiple-instance learning, the most common variant of weakly supervised learning. We introduce an innovative approach to combat bias in the multiple-instance learning paradigm. Leveraging coarse-grained bag labels, our framework **FMIL** and its variant, **FAttMIL**, infer instance-level fairness constraints, effectively bridging the gap between traditional fairness methods designed for fully supervised settings and the realities of MIL scenarios. Upon evaluation using demographic parity and equalized odds as fairness metrics, **FMIL** showcased commendable results. Our research reveals that when fairness constraints are integrated into multiple-instance learning algorithms, the proposed framework proves advantageous not only in terms of accuracy but also in fostering fairness.

References

[1] M. Hardt, E. Price, and N. Srebro, "Equality of opportunity in supervised learning," in *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, December 5-10, 2016, Barcelona, Spain*, D. D. Lee, M. Sugiyama, U. von Luxburg, I. Guyon, and R. Garnett, Eds., 2016, pp. 3315–3323.

[2] G. Goh, A. Cotter, M. R. Gupta, and M. P. Friedlander, "Satisfying real-world goals with dataset constraints," in *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, December 5-10, 2016, Barcelona, Spain*, D. D. Lee, M. Sugiyama, U. von Luxburg, I. Guyon, and R. Garnett, Eds., 2016, pp. 2415–2423.

[3] T. Kamishima, S. Akaho, and J. Sakuma, "Fairness-aware learning through regularization approach," in *Data Mining Workshops (ICDMW), 2011 IEEE 11th International Conference on, Vancouver, BC, Canada, December 11, 2011*, M. Spiliopoulou, H. Wang, D. J. Cook, J. Pei, W. Wang, O. R. Zaïane, and X. Wu, Eds. IEEE Computer Society, 2011, pp. 643–650.

[4] Y. Wu, L. Zhang, and X. Wu, "On convexity and bounds of fairness-aware classification," in *The World Wide Web Conference, WWW 2019, San Francisco, CA, USA, May 13-17, 2019*, L. Liu, R. W. White, A. Mantrach, F. Silvestri, J. J. McAuley, R. Baeza-Yates, and L. Zia, Eds. ACM, 2019, pp. 3356–3362.

[5] T. Calders, F. Kamiran, and M. Pechenizkiy, "Building classifiers with independency constraints," in *ICDM Workshops 2009, IEEE International Conference on Data Mining Workshops, Miami, Florida, USA, 6 December 2009*, Y. Saygin, J. X. Yu, H. Kargupta, W. Wang, S. Ranka, P. S. Yu, and X. Wu, Eds. IEEE Computer Society, 2009, pp. 13–18.

[6] D. Pedreschi, S. Ruggieri, and F. Turini, "Discrimination-aware data mining," in *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Las Vegas, Nevada, USA, August 24-27, 2008*, Y. Li, B. Liu, and S. Sarawagi, Eds. ACM, 2008, pp. 560–568.

[7] F. Kamiran and T. Calders, "Classifying without discriminating," in *2009 2nd International Conference on Computer, Control and Communication*. IEEE, Feb. 2009, pp. 1–6. [Online]. Available: http://ieeexplore.ieee.org/document/4909197/

[8] A. Agarwal, A. Beygelzimer, M. Dudík, J. Langford, and H. M. Wallach, "A reductions approach to fair classification," in *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018*, ser. Proceedings of Machine Learning Research, J. G. Dy and A. Krause, Eds. PMLR, 2018, pp. 60–69. [Online]. Available: http://proceedings.mlr.press/v80/agarwal18a.html

[9] M. Kleindessner, M. Donini, C. Russell, and M. B. Zafar, "Efficient fair PCA for fair representation learning," in *International Conference on Artificial Intelligence and Statistics, 25-27 April 2023, Palau de Congressos, Valencia, Spain*, ser. Proceedings of Machine Learning Research, F. J. R. Ruiz, J. G. Dy, and J.-W. van de Meent, Eds. PMLR, 2023, pp. 5250–5270. [Online]. Available: https://proceedings.mlr.press/v206/kleindessner23a.html

[10] M. B. Zafar, I. Valera, M. Gomez-Rodriguez, and K. P. Gummadi, "Fairness constraints: Mechanisms for fair classification," in *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics, AISTATS 2017, 20-22 April 2017, Fort Lauderdale, FL, USA*, ser. Proceedings of Machine Learning Research, A. Singh and X. J. Zhu, Eds. PMLR, 2017, pp. 962–970. [Online]. Available: http://proceedings.mlr.press/v54/zafar17a.html

[11] M. B. Zafar, I. Valera, M. Gomez-Rodriguez, K. P. Gummadi, and A. Weller, "From parity to preference-based notions of fairness in classification," in *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, 4-9 December 2017, Long Beach, CA, USA*, I. Guyon, U. von Luxburg, S. Bengio, H. M. Wallach, R. Fergus, S. V. N. Vishwanathan, and R. Garnett, Eds., 2017, pp. 229–239. [Online]. Available: http://papers.nips.cc/paper/6627-from-parity-to-preference-based-notions-of-fairness-in-classification

[12] M. B. Zafar, I. Valera, M. Gomez-Rodriguez, and K. P. Gummadi, "Fairness beyond disparate treatment & disparate impact: Learning classification without disparate mistreatment," in *Proceedings of the 26th International Conference on World Wide Web, WWW 2017, Perth, Australia, April 3-7, 2017*, R. Barrett, R. Cummings, E. Agichtein, and E. Gabrilovich, Eds. ACM, 2017, pp. 1171–1180.

[13] ——, "Fairness Constraints: A Flexible Approach for Fair Classification," *J. Mach. Learn. Res.*, pp. 75:1–75:42, 2019. [Online]. Available: http://jmlr.org/papers/v20/18-262.html

[14] M. Wan, D. Zha, N. Liu, and N. Zou, "In-Processing Modeling Techniques for Machine Learning Fairness: A Survey," *ACM Transactions on Knowledge Discovery from Data*, p. 3551390, Jul. 2022.

[15] Y. Wu and X. Wu, "Using loglinear model for discrimination discovery and prevention," in *2016 IEEE International Conference on Data*

Science and Advanced Analytics, DSAA 2016, Montreal, QC, Canada, October 17-19, 2016. IEEE, 2016, pp. 110–119.

[16] Y. Wu, L. Zhang, and X. Wu, "Counterfactual fairness: Unidentification, bound and algorithm," in *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI 2019, Macao, China, August 10-16, 2019*, S. Kraus, Ed. ijcai.org, 2019, pp. 1438–1444.

[17] D. Xu, Y. Wu, S. Yuan, L. Zhang, and X. Wu, "Achieving causal fairness through generative adversarial networks," in *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI 2019, Macao, China, August 10-16, 2019*, S. Kraus, Ed. ijcai.org, 2019, pp. 1452–1458.

[18] L. Zhang, Y. Wu, and X. Wu, "Situation testing-based discrimination discovery: A causal inference approach," in *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence, IJCAI 2016, New York, NY, USA, 9-15 July 2016*, S. Kambhampati, Ed. IJCAI/AAAI Press, 2016, pp. 2718–2724. [Online]. Available: http://www.ijcai.org/Abstract/16/386

[19] ——, "A causal framework for discovering and removing direct and indirect discrimination," in *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI 2017, Melbourne, Australia, August 19-25, 2017*, C. Sierra, Ed. ijcai.org, 2017, pp. 3929–3935.

[20] ——, "Achieving non-discrimination in prediction," in *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI 2018, July 13-19, 2018, Stockholm, Sweden*, J. Lang, Ed. ijcai.org, 2018, pp. 3097–3103.

[21] Y. Wu, L. Zhang, and X. Wu, "On discrimination discovery and removal in ranked data using causal graph," in *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, KDD 2018, London, UK, August 19-23, 2018*, Y. Guo and F. Farooq, Eds. ACM, 2018, pp. 2536–2544.

[22] L. Zhang, Y. Wu, and X. Wu, "Achieving non-discrimination in data release," in *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Halifax, NS, Canada, August 13 - 17, 2017*. ACM, 2017, pp. 1335–1344.

[23] Y. Wu, L. Zhang, X. Wu, and H. Tong, "PC-Fairness: A unified framework for measuring causality-based fairness," in *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, 8-14 December 2019, Vancouver, BC, Canada*, H. M. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. B. Fox, and R. Garnett, Eds., 2019, pp. 3399–3409. [Online]. Available: http://papers.nips.cc/paper/8601-pc-fairness-a-unified-framework-for-measuring-causality-based-fairness

[24] L. Zhang, Y. Wu, and X. Wu, "On discrimination discovery using causal networks," in *Social, Cultural, and Behavioral Modeling, 9th International Conference, SBP-BRiMS 2016, Washington, DC, USA, June 28 - July 1, 2016, Proceedings*, ser. Lecture Notes in Computer Science, K. S. Xu, D. Reitter, D. Lee, and N. Osgood, Eds. Springer, 2016, pp. 83–93.

[25] ——, "Causal modeling-based discrimination discovery and removal: Criteria, bounds, and algorithms," *IEEE Transactions on Knowledge and Data Engineering*, no. 11, pp. 2035–2050, 2019.

[26] B. Becker and R. Kohavi, "Adult," UCI Machine Learning Repository, 1996, DOI: https://doi.org/10.24432/C5XW20.

[27] P. Nodet, V. Lemaire, A. Bondu, A. Cornuéjols, and A. Ouorou, "From weakly supervised learning to biquality learning, a brief introduction," *CoRR*, 2020. [Online]. Available: https://arxiv.org/abs/2012.09632

[28] D. Pedreschi, S. Ruggieri, and F. Turini, "A study of top-k measures for discrimination discovery," in *Proceedings of the ACM Symposium on Applied Computing, SAC 2012*, S. Ossowski and P. Lecca, Eds. ACM, 2012, pp. 126–131.

[29] D.-H. Lee *et al.*, "Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks," in *Workshop on challenges in representation learning, ICML*, no. 2, 2013, p. 896.

[30] N. Mehrabi, F. Morstatter, N. Saxena, K. Lerman, and A. Galstyan, "A survey on bias and fairness in machine learning," *ACM Comput. Surv.*, no. 6, pp. 115:1–115:35, 2021.

[31] K. Bhaila, Y. Wu, and X. Wu, "Fair collective classification in networked data," in *IEEE International Conference on Big Data, Big Data 2022, Osaka, Japan, December 17-20, 2022*, S. Tsumoto, Y. Ohsawa, L. Chen, D. V. den Poel, X. Hu, Y. Motomura, T. Takagi, L. Wu, Y. Xie, A. Abe, and V. Raghavan, Eds. IEEE, 2022, pp. 1415–1424.

[32] X. Jiang, Y. Dai, and Y. Wu, "Fair selection through kernel density estimation," in *International Joint Conference on Neural Networks, IJCNN 2023, Gold Coast, Australia, June 18-23, 2023*. IEEE, 2023, pp. 1–8.

[33] F. Kamiran and T. Calders, "Data preprocessing techniques for classification without discrimination," *Knowl. Inf. Syst.*, no. 1, pp. 1–33, 2011.

[34] S. Corbett-Davies, E. Pierson, A. Feller, S. Goel, and A. Huq, "Algorithmic decision making and the cost of fairness," in *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Halifax, NS, Canada, August 13 - 17, 2017*. ACM, 2017, pp. 797–806.

[35] F. Kamiran, T. Calders, and M. Pechenizkiy, "Discrimination aware decision tree learning," in *ICDM 2010, the 10th IEEE International Conference on Data Mining, Sydney, Australia, 14-17 December 2010*, G. I. Webb, B. Liu, C. Zhang, D. Gunopulos, and X. Wu, Eds. IEEE Computer Society, 2010, pp. 869–874.

[36] T. Kamishima, S. Akaho, H. Asoh, and J. Sakuma, "Fairness-aware classifier with prejudice remover regularizer," in *Machine Learning and Knowledge Discovery in Databases - European Conference, ECML PKDD 2012*, ser. Lecture Notes in Computer Science, P. A. Flach, T. D. Bie, and N. Cristianini, Eds. Springer, 2012, pp. 35–50.

[37] F. Kamiran, A. Karim, and X. Zhang, "Decision theory for discrimination-aware classification," in *12th IEEE International Conference on Data Mining, ICDM 2012, Brussels, Belgium, December 10-13, 2012*, M. J. Zaki, A. Siebes, J. X. Yu, B. Goethals, G. I. Webb, and X. Wu, Eds. IEEE Computer Society, 2012, pp. 924–929.

[38] YCAP. Reddy, P. Viswanath, and B. E. Reddy, "Semi-supervised learning: A brief review," no. 1.8, p. 81, 2018.

[39] X. Yang, Z. Song, I. King, and Z. Xu, "A survey on deep semi-supervised learning," *IEEE Trans. Knowl. Data Eng.*, no. 9, pp. 8934–8954, 2023.

[40] V. Noroozi, S. Bahaadini, S. Sheikhi, N. Mojab, and P. S. Yu, "Leveraging semi-supervised learning for fairness using neural networks," in *18th IEEE International Conference On Machine Learning And Applications, ICMLA 2019, Boca Raton, FL, USA, December 16-19, 2019*, M. A. Wani, T. M. Khoshgoftaar, D. Wang, H. Wang, and N. Seliya, Eds. IEEE, 2019, pp. 50–55.

[41] T. Zhang, T. Zhu, J. Li, M. Han, W. Zhou, and P. S. Yu, "Fairness in semi-supervised learning: Unlabeled data help to reduce discrimination," *IEEE Trans. Knowl. Data Eng.*, no. 4, pp. 1763–1774, 2022.

[42] M. Sugiyama, "Machine Learning from Weak Supervision: An Empirical Risk Minimization Approach," p. 47.

[43] LEMAIRE. Vincent, "Weakly supervised learning," p. 77.

[44] Z.-H. Zhou, "A brief introduction to weakly supervised learning," *National Science Review*, no. 1, pp. 44–53, Jan. 2018. [Online]. Available: https://academic.oup.com/nsr/article/5/1/44/4093912

[45] J. Fitzsimons, M. Osborne, and S. Roberts, "Intersectionality: Multiple Group Fairness in Expectation Constraints," *arXiv:1811.09960 [cs, stat]*, Nov. 2018. [Online]. Available: http://arxiv.org/abs/1811.09960

[46] J. R. Foulds, R. Islam, K. N. Keya, and S. Pan, "An intersectional definition of fairness," in *36th IEEE International Conference on Data Engineering, ICDE 2020, Dallas, TX, USA, April 20-24, 2020*. IEEE, 2020, pp. 1918–1921.

[47] P. L. Bartlett, M. I. Jordan, and J. D. McAuliffe, "Convexity, Classification, and Risk Bounds," *Journal of the American Statistical Association*, no. 473, pp. 138–156, Mar. 2006.

[48] A. Blum and A. Kalai, "A note on learning from multiple-instance examples," *Mach. Learn.*, no. 1, pp. 23–29, 1998.

[49] M. Ilse, J. M. Tomczak, and M. Welling, "Attention-based deep multiple instance learning," in *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018*, ser. Proceedings of Machine Learning Research, J. G. Dy and A. Krause, Eds. PMLR, 2018, pp. 2132–2141. [Online]. Available: http://proceedings.mlr.press/v80/ilse18a.html

[50] F. Ding, M. Hardt, J. Miller, and L. Schmidt, "Retiring adult: New datasets for fair machine learning," in *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, 2021, pp. 6478–6490.