Abstract—Bias in training datasets must be managed for various groups in classification tasks to ensure parity or equal treatment. With the recent growth in artificial intelligence models and their expanding role in automated decision-making, it is vital to ensure that these models are not biased. There is an abundance of evidence suggesting that these models could contain or even amplify the bias present in the data on which they are trained, inherent to their objective function and learning algorithms; Existing methods for mitigating bias result in information loss and do not provide a suitable balance between accuracy and fairness or do not ensure limiting the biases in training. To this end, we propose a powerful strategy for training deep learning models called the Distraction module, which can be effective in controlling bias from affecting the classification results. This method can be utilized with different data types (e.g., tabular, images, graphs). We demonstrate the potency of the proposed method by testing it on UCI Adult and Heritage Health datasets (tabular), POKEC-Z, POKEC-N and NBA datasets (graph), and CelebA dataset (vision). Considering state-of-the-art methods proposed in the fairness literature for each dataset, we exhibit that our model is superior to these proposed methods in minimizing bias and maintaining accuracy.

Index Terms—Deep Learning, Fairness, Neural Networks, Adversarial Training, Game Theory

I. INTRODUCTION

Artificial intelligence and machine learning models in real-world applications have grown in past decades and led to automated decision-making in different domains such as hiring pipelines, face recognition, financial services, the healthcare system, and criminal justice. Algorithmic decision-making may cause an algorithmic bias toward the central population subgroup and discrimination and unfairness toward the minority. In recent years, fairness in artificial intelligence has increased ethical concerns and received attention from interdisciplinary research communities [1]. Several definitions of fairness have been put forth as potential solutions to the problem of unwanted bias in machine learning techniques. In most cases, the definitions may be separated into two categories: individual fairness and collective fairness. A system that is individual will treat users that are similar to each other in the same manner, where the similarities between people may be determined by past information [2], [3]. Group fairness metrics are measurements of the statistical equality between different subgroups that are characterized by sensitive characteristics such as race, or gender [4]–[6].

In this paper, our focus is on group fairness, and from this point on, we refer to group fairness as fairness.

We introduced a novel in-process bias mitigation method that does not need adversarial example generation to train a fair classifier and can be trained on available datasets without alteration. We proposed a departure from the typical adversarial paradigm. Rather than using a separate discriminator network or injecting adversarial noise, we introduce a new training approach that embeds a specialized module (Distraction Module) within the main neural network. This embedded module independently manages a subset of weights within the main network and is explicitly trained to optimize the network towards specific criteria or alternative objective functions.

Our primary contributions are:

A novel training method that incorporates a specialized module within the main network to optimize towards specific criteria or alternative objective functions.

Empirical validation of our approach on multiple benchmarks with various data types, demonstrating improvements in network fairness.

New adversarial training procedure which significantly improves the state-of-the-art in both accuracy and fairness metrics.

II. RELATED WORKS

Fairness in machine learning has garnered significant attention, with methods primarily spanning three categories: Pre-process approaches adjust data prior to model training to achieve fair outcomes. This includes changing or reweighing labels [7], [8] and modifying feature distributions to make differentiation between privileged and unprivileged groups [9], [10]. Recently, a GAN was introduced to generate unbiased tabular datasets, focusing on both accuracy and fairness [11].

In-process approaches alter the algorithm during training. Some add regularization terms to the objective function, balancing fairness and accuracy. For example, mutual information between protected attributes and classifier predictions was penalized [12], while others added constraints to satisfy equalized odds proxies [13], [14].

Post-process approaches modify outcomes post-training. Strategies range from flipping certain outcomes [6] to using different thresholds for privileged and unprivileged groups, balancing fairness and accuracy [15], [16].

Different data types, namely tabular, graph, and images, require specialized fairness approaches:

Tabular: Efforts in tabular data focus on mitigating bias. Approaches include regularizing covariance between predictions and sensitive variables [17], standardizing decision bounds [13], and restricting adversaries from inferring sensitive characteristics [18]. Game-theoretic methods have been employed, although scaling them remains challenging [19]. Other noteworthy techniques leverage attention-based approaches, mutual information, and information-theoretic methods [20]–[24].

Graph: Graphs can amplify biases, particularly in networks where nodes with similar sensitive features are more likely to connect [25], [26]. This can lead to severe decision-making biases in Graph Neural Networks (GNNs) [27]. Most fair models were designed for i.i.d data and often don’t cater to
graph data. However, there have been pioneering efforts in learning fair node representations from graphs [25]–[28].

**Vision**: Biases in vision models can manifest in various ways, such as gender biases in action recognition or disparities in face recognition across racial and gender categories [29], [29]. Solutions span from altering GAN utility functions for fair image datasets [30], [31] to adversarial game formulations [32] and methodologies for balanced data generation [33]. Apart from GANs, techniques like U-Nets, deep information [32] and methodologies for balanced data generation [33].

We treat the Distraction module and the encompassing network as described in [37].

### III. Methodology

#### A. Problem Definition

Our focus in this study is binary classification tasks. We posit that the techniques proposed here can be extended to multi-class datasets without constraint.

Let's consider a dataset denoted as \( E = \{ \mathbf{x}^{(i)}, \mathbf{a}^{(i)}, y^{(i)} \} \), where \( \mathbf{x}^{(i)}, \mathbf{a}^{(i)} \), and \( y^{(i)} \) are independently and identically distributed samples drawn from the data distribution \( P(\mathbf{x}, y, \mathbf{a}) \).

Here:

- \( \mathbf{x} \) represents the features of the dataset.
- \( y \in \{0, 1\} \) indicates the label.
- \( \mathbf{a} \), which takes on discrete finite values, is the protected attribute of our data.

#### B. Model Architecture

Proposed method employs two distinct sets of weights for its neural network classifier:

- The primary set, aimed at the classification task, optimizes for accuracy (\( \theta_c \)).
- The second set, associated with the "Distraction module", is designed to enhance fairness within the model (\( \theta_d \)).

A graphical representation of this architecture can be found in Figure 1. This model can be configured using a variety of neural network architectures such as fully connected layers, graph convolution network, convolution neural network, and etc.

#### C. Formulation

The goal of fairness in classification is to ensure that predictions remain consistent regardless of the protected attribute. This concept can be formally captured by:

\[
p(C(x, D(x; \theta_d); \theta_c) = s|\mathbf{a}, y) = p(C(x, D(x; \theta_d); \theta_c) = s|\mathbf{a}', y)
\]

as described in [37].

The methodology incorporates a game-theoretic approach. We treat the Distraction module and the encompassing network as two distinct players in a maximin game. Our objective is to train both sets of weights concurrently, aiming to create a classifier that balances fairness with accuracy. The weights of the Distraction module are entirely isolated from the leading network. The Distraction module tries to make the classifier function results as fair as possible, while the whole network tries to make the classifier function results as accurate as possible. The first player is the Distraction module, and the second is the whole network containing the Distraction module. We train two sets of weights simultaneously to achieve a fair and accurate classifier. This game is established with two utility functions, one for each player. We denote the Distraction module as \( D(x) \) and the whole network as \( C(x, D(x)) \) throughout this paper. To put it differently, \( D \) and \( C \) play the two-player maximin game with a utility function of \( U_1(D, C) \) (Eq. 2) for player \( C \) and \( U_2(D, C) \) (Eq. 3) for player \( D \).

\[
\begin{align*}
\min_{\theta_c} U_1(C, D) = \mathbb{E}_{\mathbf{x} \sim X} \mathbb{E}_{y \sim Y} \left[ -\log p_{\theta_d, \theta_c}(y|x) \right] \\
\max_{\theta_d} U_2(C, D) = -\mathbb{E}_{\mathbf{s} \sim C(x, D(x; \theta_d); \theta_c)} \left[ -\log p_{\theta_d, \theta_c}(s) \right]
\end{align*}
\]

#### D. Trade-off Between Fairness and Accuracy

A common challenge is that a classifier might not be both completely fair and optimal. This is often because the protected attributes, along with their proxies, can influence classifier decisions substantially.

To manage this, we introduce a parameter, \( \eta \), which helps strike a balance between model accuracy and fairness. The role of \( \eta \) is central to the Demographic parity loss (Eq. 4) utilized by the Distraction module. A smaller \( \eta \) leans towards accuracy, while a larger value promotes fairness. By tuning \( \eta \), we can generate a series of Pareto solutions for our multi-objective optimization task.

\[
\eta \sum_{i=1}^{m} -\log p_{\theta_d, \theta_c}(a^{(i)}|C(x^{(i)}, D(x^{(i)})))
\]

#### E. Training Procedure

We employ a mini-batch stochastic gradient descent technique to train our network. Algorithm 1 offers a detailed training procedure used for training networks containing Distraction module.

During each iteration:

- First, the network’s fairness loss is computed, followed by an update to the Distraction module weights.
- Then, the network undergoes another iteration with frozen weights of the Distraction module where the classification loss is computed, and back-propagation is employed for updating the weight of the remaining network.

### IV. Experiments

This section compares our method to other benchmark methods in the literature. The experiment section consists of three parts. First, we experiment with tabular datasets. We compare the classification accuracy and statistical parity of deep learning methods in the benchmark datasets. In the second section, we use graph data for node classification tasks. We evaluate our model on vision datasets in the third and final
Fig. 1. Depiction of Distraction module in MLP model: The red weights are controlled by the Distraction module and are optimized with Eq. 3 objective function, based on Demographic parity (DP) loss. The blue weights have different optimizer and Binary Cross Entropy (BCE) loss function. They are controlled by the main classifier.

Algorithm 1 Minibatch stochastic gradient descent for adversarial training of a network with distraction module

| Input: data \((X, A, Y)\), \(A\) is a set of protected attributes and \(Y\) is the label, Batch Size \(m\), \(C\)-Learning rate \(lr_1\), \(D\)-Learning rate \(lr_2\), \(\eta\) |
|---|
| for number of iterations in training do |
| sample minibatch of size \(m\) samples \(\{x^{(1)}, x^{(2)}, \ldots, x^{(m)}\}\) from data \(P(x)\) |
| updates the Distraction module by ascending the stochastic gradient with learning rate \(lr_2\): |
| \(\nabla_{\theta_d} \eta \sum_{i=1}^{m} - \log p_{\theta_d, \theta_c}(a^{(i)}|C(x^{(i)}, D(x^{(i)})))\) |
| updates the classifier network by descending the stochastic gradient with learning rate \(lr_1\): |
| \(\nabla_{\theta_c} \sum_{i=1}^{m} - \log p_{\theta_d, \theta_c}(y^{(i)}|x^{(i)})\) |
| end for |

In a practical perspective, this method needs two optimizers which can be any standard gradient-based method. We used the Adam optimizer for both of the functions in our experiments.

The Distraction module used on all the datasets consists of only linear layers, and the Distraction module is positioned one layer before the classification layer. This choice was due to experiments conducted on the vision and tabular datasets. The ablation study, and the additional results can be found in the section IV-F. Additionally, we observed that the loss for both the fairness metric and the accuracy is very volatile in training. However, given enough steps, it always converges to a single point which is a Pareto answer for this multi-objective optimization. We decided not to include loss graphs in the paper due to the volatility of the losses.

A. Evaluation Metrics

We employ four evaluation metrics to compare the performance of our model with baseline models. These metrics are as follows:

1) **Average Precision (AP):**
   - **Definition:** The average precision is a measure that computes the average while combining recall and precision at each threshold. It provides an aggregate assessment of the classifier’s performance over all levels of precision-recall.
   - **Objective:** Higher AP values are preferred, indicating better accuracy of the classifiers.

2) **Accuracy:**
   - **Usage:** Consistency with literature suggests using accuracy for tabular and graph datasets.
   - **Objective:** A higher accuracy indicates a better performing model.

3) **Demographic Parity (DP):**

| Method | UCI Adult | Heritage Health |
|---|---|---|
| Distraction (Ours) | 0.411 | 0.503 |
| FCRL (AAAI 2021) | 0.253 | 0.285 |
| Attention | 0.213 | 0.139 |
| CIVB (NeurIPS 2018) | 0.163 | 0.176 |
| MIFR (AIStats 2019) | 0.221 | 0.166 |
| MaxEnt-ARL (CVPR 2019) | 0.133 | 0.0 |
| Adv Forgetting (AAAI 2020) | 0.077 | 0.172 |
TABLE II  
SUMMARY OF PARAMETER SETTING FOR THE DISTRACTION ON TABULAR DATASETS

| Hyperparameters                        | UCI Adult | Health Heritage |
|----------------------------------------|-----------|-----------------|
| FC layers before the Distraction module| 2         | 2               |
| FC layers of the Distraction module    | 1         | 3               |
| FC layers after the Distraction module | 1         | 1               |
| Epoch                                  | 50        | 50              |
| Batch size                             | 100       | 100             |
| Dropout                                | 0         | 0               |
| Network optimizer                      | Adam      | Adam            |
| Distraction module optimizer           | Adam      | Adam            |
| Network learning rate                  | 1e-3      | 1e-3            |
| Distraction module learning rate       | 1e-5      | 1e-5            |
| \(\eta\)                               | 100       | 100             |

- **Definition**: DP is a widely-used fairness metric that captures the difference in the probability of receiving a favorable decision between different protected groups. It is calculated as the absolute difference \(|P(Y = 1 | S = 0) - P(Y = 1 | S = 1)|\). For more than two groups, DP can be computed using \(\Delta_{DP}(a, \hat{y}) = \max_{a_i, a_j} |P(\hat{y} = 1 | a = a_i) - P(\hat{y} = 1 | a = a_j)|\) [24, 38].

- **Objective**: A smaller DP indicates a more fair classification, as it reduces disparity between protected groups.

4) **Difference in Equality of Opportunity (\(\Delta EO\))**:  
- **Definition**: Following [39] and [33], we use \(\Delta EO\) as a fairness metric. It is defined as the absolute difference between the true positive rates for different group expressions \(|TPR(S = 0) - TPR(S = 1)|\).
- **Objective**: Lower values of \(\Delta EO\) are preferred, indicating fairer categorization between group expressions.

In the experiments, we used demographic parity as the fairness criterion during training. This choice indicates that the model is optimized for demographic parity and the demographic parity is the main metric that our model is providing. The better fairness metrics provided are byproduct of optimizing the model with the demographic parity objective.

B. Implementation details

The hyperparameters used in training the models on each tabular, graph, and vision datasets can be found in the tables II and IV respectively. The training was performed on a single NVIDIA GeForce RTX 3090.

C. Tabular

We evaluated our method for bias mitigation to various current state-of-the-art approaches. We concentrate on strategies specifically tuned to achieve the best results in statistical parity metrics on tabular studies.

1) **Datasets**: We conducted experiments using two well-established benchmark datasets in this field:

**UCI Adult Dataset** [40]: This dataset, based on demographic data collected in 1994, includes a training set of 30,000 samples and a test set of 15,000 samples. The task is to predict whether an individual’s salary exceeds $50,000 per annum, with gender serving as the binary protected attribute.

**Heritage Health Dataset**: This dataset involves predicting the Charleson Index, a measure of a patient’s 10-year mortality risk. It comprises samples from approximately 51,000 patients, split into a training set of 41,000 and a test set of 11,000. The protected attribute, age, has nine possible values. We selected the protected and target attributes in accordance with existing literature.

We used training sets to run the Distraction module technique and train the network. Then we assess the performance by running classifiers for subsequent prediction tasks.

2) **Benchmark Methods**: We compared our approach with the following state-of-the-art techniques:

- **CVIB** [21]: This method employs a conditional variational autoencoder for bias mitigation.
- **MIFR** [22]: This technique optimizes fairness by leveraging an information bottleneck factor combined with adversarial learning.
- **FCRL** [24]: This approach uses specific approximations for contrastive information to maximize theoretical goals, which can be employed to strike appropriate trade-offs between statistical demographic parity and accuracy.
- **MaxEnt-ARL** [32]: This method uses adversarial learning to reduce bias in tabular data.
- **Adversarial Forgetting** [23]: Another method that employs adversarial learning for bias mitigation.

3) **Results**: We employed the training sets to execute our Distraction module method and train the network. Subsequent prediction tasks were evaluated using classifiers. The average accuracy, representing the most likely accuracy, and maximum demographic parity, indicating worst-case bias, were computed over five training iterations with random seeds. In contrast to [24], no preprocessing was applied to the data before inputting it into our network. Our reported results represent Pareto solutions for the neural network during training with varying \(\eta\).

Figures 2a and 2b depict trade-offs between statistical demographic parity and accuracy for various bias reduction techniques on the UCI Adult and Heritage Health datasets, respectively. An ideal result would position the curve in the lower right corner of the graph, indicating accurate and fair outcomes concerning protected attributes. Our results demonstrate that the Distraction method significantly outperforms competing methods. This finding is further supported by the area-over-the-curve data for demographic parity and accuracy (Table I), showing that our proposed strategy improves the area over the curve for both datasets by a factor of \(\sim 2\). This implies that our bias reduction framework is the most effective mitigation strategy for tabular data.

D. Graph

We compare our proposed framework with state-of-the-art approaches for fair classification and fair graph embedding
Table III

Summary of Parameter Setting for the Distraction on Graph Datasets

| Hyperparameters                  | POKEC-Z | POKEC-N | NBA   |
|----------------------------------|---------|---------|-------|
| GCN layer before the Distraction module | 2       | 2       | 2     |
| Distraction module FC layers     | 1       | 1       | 1     |
| FC layers after the Distraction module | 1       | 1       | 1     |
| Epoch                            | 5000    | 1000    | 1000  |
| Batch size                       | 1       | 1       | 1     |
| Dropout                          | 0       | 0.5     | 0.5   |
| Network optimizer                | Adam    | Adam    | Adam  |
| Distraction module optimizer     | Adam    | Adam    | Adam  |
| Network learning rate            | 1e-3    | 1e-3    | 1e-2  |
| Distraction module learning rate | 1e-6    | 1e-8    | 1e-5  |
| η                                | 1000    | 100     | 1000  |

Table IV

Summary of Parameter Setting for the Distraction on Vision Dataset

| Hyperparameters                  | CelebA-Attractive | CelebA-Smiling | CelebA-WavyHair |
|----------------------------------|-------------------|----------------|-----------------|
| Distraction module FC layers     | 1                 | 1              | 1               |
| FC layers after the Distraction module | 1                 | 1              | 1               |
| Epoch                            | 30                | 15             | 15              |
| Batch size                       | 128               | 128            | 128             |
| Dropout                          | 0                 | 0              | 0               |
| Network optimizer                | Adam              | Adam           | Adam            |
| Distraction module optimizer     | Adam              | Adam           | Adam            |
| Network learning rate            | 1e-3              | 1e-3           | 1e-3            |
| Distraction module learning rate | 1e-6              | 1e-5           | 1e-5            |
| η                                | 1000              | 100            | 100             |

Fig. 2. Accuracy of different benchmark models and distraction model vs. statistical demographic parity of each model for the (a) UCI Adult dataset and (b) Heritage Health dataset. The ideal area for this graph is the bottom right which indicates high accuracy and low statistical demographic parity. This graph shows that our model performs significantly better than the other benchmark models in both tabular datasets.

Learning, including ALFR [41], ALFR-e, Debias [18], Debias-e, FCGE [28], FairGCN [27], and NT-FAIRGNN [42]. A brief overview of these methods is as follows:

- **ALFR [41]**: A pre-processing approach that removes sensitive information from representations generated by an MLP-based autoencoder using a discriminator. The debiased representations are then used to train a linear classifier.
- **ALFR-e**: An extension of ALFR that incorporates graph structure information by combining user features in ALFR with graph embeddings discovered by deepwalk [43].
- **Debias [18]**: An in-processing fair classification method that directly applies a discriminator to the predicted probability of the classifier.
- **Debias-e**: An extension of Debias that includes deepwalk embeddings into the Debias features.
- **FCGE [28]**: A method for learning fair node embeddings in graphs without node characteristics. Discriminators filter out sensitive data in the embeddings.
- **FairGCN [27]**: A graph convolutional network designed for fairness in graph-based learning. It incorporates fair-
TABLE V  
THE COMPARISONS OF OUR PROPOSED METHOD WITH THE BASELINES ON POKEC-Z

| METHOD       | ACC(%) | AUC(%) | $\Delta_{DP}$(%) | $\Delta_{EO}$(%) |
|--------------|--------|--------|-----------------|-----------------|
| ALFR         | 65.4 ±0.3 | 71.3 ±0.3 | 2.8 ±0.5 | 1.1 ±0.4 |
| ALFR-E       | 68.0 ±0.6 | 74.0 ±0.7 | 5.8 ±0.4 | 2.8 ±0.8 |
| DEBIAS       | 65.2 ±0.7 | 71.4 ±0.6 | 1.9 ±0.6 | 1.9 ±0.4 |
| DEBIAS-E     | 67.5 ±0.7 | 74.2 ±0.7 | 4.7 ±1.0 | 3.0 ±1.4 |
| FCGE         | 65.9 ±0.2 | 71.0 ±0.2 | 3.1 ±0.5 | 1.7 ±0.6 |
| FAIRGCN      | 70.0 ±0.3 | 76.7 ±0.2 | 0.9 ±0.5 | 1.7 ±0.2 |
| FAIRGAT      | 70.1 ±0.1 | 76.5 ±0.2 | 0.5 ±0.3 | 0.8 ±0.3 |
| NT-FAIRGNN   | 70.0 ±0.1 | 76.7 ±0.3 | 1.0 ±0.4 | 1.6 ±0.2 |
| GAT+DISTRACTION (Ours) | **70.97 ±0.16** | **77.58 ±0.13** | 0.93 ±0.44 | **0.97 ±0.40** |

TABLE VI  
THE COMPARISONS OF OUR PROPOSED METHOD WITH THE BASELINES ON POKEC-N

| METHOD       | ACC(%) | AUC(%) | $\Delta_{DP}$(%) | $\Delta_{EO}$(%) |
|--------------|--------|--------|-----------------|-----------------|
| ALFR         | 63.1 ±0.6 | 67.7 ±0.5 | 3.05 ±0.5 | 3.9 ±0.6 |
| ALFR-E       | 66.2 ±0.5 | 71.9 ±0.3 | 4.1 ±0.5 | 4.6 ±1.6 |
| DEBIAS       | 62.6 ±0.9 | 67.9 ±0.7 | 2.4 ±0.7 | 2.6 ±1.0 |
| DEBIAS-E     | 65.6 ±0.8 | 71.7 ±0.7 | 3.6 ±0.2 | 4.4 ±1.2 |
| FCGE         | 64.8 ±0.5 | 69.5 ±0.4 | 4.1 ±0.8 | 5.5 ±0.9 |
| FAIRGCN      | 70.1 ±0.2 | 74.9 ±0.4 | 0.8 ±0.2 | 1.1 ±0.5 |
| FAIRGAT      | 70.0 ±0.2 | 74.9 ±0.4 | 0.6 ±0.3 | **0.8 ±0.2** |
| NT-FAIRGNN   | 70.1 ±0.2 | 74.9 ±0.4 | 0.8 ±0.2 | 1.1 ±0.3 |
| GAT+DISTRACTION (Ours) | **70.07 ±0.05** | **75.8 ±0.38** | 0.62 ±0.14 | 3.0 ±1.0 |

TABLE VII  
THE COMPARISONS OF OUR PROPOSED METHOD WITH THE BASELINES ON NBA

| METHOD       | ACC(%) | AUC(%) | $\Delta_{DP}$(%) | $\Delta_{EO}$(%) |
|--------------|--------|--------|-----------------|-----------------|
| ALFR         | 64.3 ±1.3 | 71.5 ±0.3 | 2.3 ±0.9 | 3.2 ±1.5 |
| ALFR-E       | 66.0 ±0.4 | 72.9 ±1.0 | 4.7 ±1.8 | 4.7 ±1.7 |
| DEBIAS       | 63.1 ±1.1 | 71.3 ±0.7 | 2.5 ±1.5 | 3.1 ±1.9 |
| DEBIAS-E     | 65.6 ±2.4 | 72.9 ±1.2 | 5.3 ±0.9 | 3.1 ±1.3 |
| FCGE         | 66.0 ±1.5 | 73.6 ±1.5 | 2.9 ±1.0 | 3.0 ±1.2 |
| FAIRGCN      | 71.1 ±1.0 | 77.0 ±0.3 | 1.0 ±0.5 | 1.2 ±0.4 |
| FAIRGAT      | 71.5 ±0.8 | 77.5 ±0.7 | 0.7 ±0.5 | **0.7 ±0.3** |
| NT-FAIRGNN   | 71.1 ±1.0 | 77.0 ±0.3 | 1.0 ±0.5 | 1.2 ±0.4 |
| GAT+DISTRACTION (Ours) | **77.09 ±0.45** | **77.99 ±0.58** | 0.34 ±0.21 | 12.78 ±2.9 |

—

ness constraints during training to reduce disparities between protected groups.

- **NT-FAIRGNN [42]**: A graph neural network that aims to achieve fairness by balancing the trade-off between accuracy and fairness. It uses a two-player minimax game between the predictor and the adversary, where the adversary aims to maximize the unfairness.

1) **Datasets**: We conducted experiments using the following datasets obtained from the study by [27]:

- **Pokec [44]**: A widely-used social network dataset from Slovakia, akin to Facebook and Twitter, containing anonymized data from the entire social network of year 2012. The dataset includes user information such as gender, age, interests, hobbies, and profession. Sampled subsets, Pokec-z and Pokec-n, are created based on user provinces. The classification task involves predicting users’ working environments.

- **NBA**: A Kaggle dataset of approximately 400 NBA basketball players, featuring 2016-2017 season statistics, nationality, age, and salary. Graph connections were established using relationships between NBA players on Twitter, collected via the official API. Players are categorized as American or international, a sensitive attribute. The classification task is to predict whether a player’s wage is above the median.

2) **Results**: Each experiment was conducted five times, and Tables V, VI, and VII report the mean and standard deviation of the runs for Pokec-z, Pokec-n, and NBA datasets, respectively. These results represent the selected Pareto solutions for comparison with the benchmarks. The tables reveal that, in comparison to GAT, generic fair classification techniques and graph embedding learning approaches exhibit inferior classification performance, even when utilizing graph information. In contrast, our Distraction method performs comparably to...
baseline GNNs. FairGCN is close to the baseline, but our Distraction technique outperforms it. When sensitive information is scarce (e.g., NBA dataset), baselines exhibit clear bias, with graph-based baselines performing worse. However, our proposed model yields near-zero statistical demographic parity, indicating effective discrimination mitigation.

E. Vision

We compare our proposed method on vision tasks with several existing approaches, including: (1) Empirical Risk Minimization (ERM) which achieves the training task without any regularization, (2) Gap Regularization which directly regularizes the model, (3) Adversarial Debiasing [18], and (4) FairMixup [19].

1) Dataset and Setup: To demonstrate the efficacy of our method, we used the CelebA dataset of face attributes [45], consisting of over 200,000 images of celebrities. Each image in this dataset has been annotated with 40 binary attributes, including gender, by human annotators. We chose three attributes – attractive, smiling, and wavy hair – for binary classification tasks, using gender as the protected attribute. These attributes were selected because each of them has a sensitive group that receives a disproportionately high number of positive samples. For each task, we employed a ResNet-18 architecture [46], augmented with two additional layers for outcome prediction.

2) Results: The trade-off between Average Precision (AP), Demographic Parity (DP), and Equality of Opportunity (EO) for attributes “Attractive”, “Smiling”, and “Wavy Hair” is illustrated in the figures 3, 4, and 5 respectively. Our pro-
Fig. 5. “Wavy Hair” Attribute of CelebA Dataset as the Target Attribute. (a) reflects the trade-off between Average Precision and Demographic Parity Difference. (b) shows the trade-off between Average Precision and Equalized Odds Difference. The Distraction module is showing competitive results to the baseline.

Fig. 6. CelebA Dataset – t-SNE visualization $z$ without the Distraction module (a) and with the Distraction module (b) labeled with gender classes. The invariant encoding $z$ (a) shows clustering by gender. In contrast (b) shows no recognizable clustering. These plots are generated using “Attractive” attribute.

Our proposed Distraction module provides a more balanced trade-off between accuracy and fairness. Instead of prioritizing one over the other, our method strikes a better balance, ensuring that the trained model is both accurate and fair. Moreover, our Distraction module consistently provides better equality of opportunity across various accuracy levels compared to benchmark models. Through empirical validation on multiple benchmarks, we’ve shown that the Distraction module consistently outperforms other methods in achieving equality of opportunity across various accuracy levels. This indicates that our method can provide fair treatment to different protected groups while still maintaining high predictive accuracy. To further illustrate the effectiveness of the Distraction module, we present t-SNE visualizations of the output embeddings from the ResNet-18 model both with and without the Distraction module. t-SNE (t-distributed Stochastic Neighbor Embedding) is a dimensionality reduction technique that is particularly well-suited for visualizing high-dimensional data in a low-dimensional space.

In Figure 6a, we show the t-SNE plot of the embeddings, $z$, produced by the ResNet-18 model without the Distraction module. In this plot, it is evident that the embeddings are clustered by gender, suggesting that the model has learned to rely on gender information for its classifications. This clustering by gender is problematic as it indicates that the model is biased and may exhibit unfair behavior when making predictions.

Conversely, in Figure 6b, we present the t-SNE plot of the embeddings, $\tilde{z}$, generated by the ResNet-18 model with the Distraction module. In this visualization, the gender-based clustering observed in the previous plot is no longer apparent. Instead, the embeddings are distributed more evenly in the low-dimensional space, indicating that the Distraction module has successfully reduced the influence of gender information.
on the model’s embeddings.

The comparison of these two t-SNE plots demonstrates the capability of the Distraction module to mitigate bias in the model’s embeddings. By preventing the model from relying on protected attributes such as gender for its classifications, the Distraction module promotes fairness and reduces the risk of discriminatory behavior in the model’s predictions.

F. Ablation Study

In this subsection, we perform an ablation study to investigate the effects of different functions for the fairness layer in the Distraction module. The fairness layer can be any differentiable function with controllable parameters denoted as \( b_d \). We experimented with three configurations for the Distraction module: one linear layer, two linear layers, and three linear layers on tabular datasets. The results of the ablation study are summarized in Table VIII.

For the CelebA dataset, we explored three types of fairness layers: linear layers, Residual Blocks (ResBlocks), and Convolutional Neural Network (CNN) layers. The mean scores of each category of CelebA attributes for each type of fairness layer are provided in Table IX. The "Inconsistency between experts" (Incons.), "Gender-dependent" (G-dep), and "Gender-independent" (G-indep) columns in Table IX represent different fairness metrics related to the CelebA dataset.

The justification for the performance differences between the ResBlock and the fully connected models in our ablation study lies in the proportion of the model occupied by the Distraction module and the specific contributions of these modules to different parts of the network. In particular, there are two primary factors that explain the observed performance differences: the roles of the modules in the network and the flow of data through these modules.

Role in the Network: The ResBlock and the fully connected modules serve different purposes within the network. The ResBlock contributes to the embedding space of the image, which includes feature extraction and representation learning. This enables the model to capture the essential characteristics of the image while minimizing the effect of the protected attributes (e.g., gender) on the classification task. In contrast, the fully connected module is mainly involved in the classification part of the network, where it contributes to the decision-making process based on the features extracted from the previous layers. This distinction in roles explains why the ResBlock provides more fair results, as it directly affects the representation learning and reduces the influence of the protected attributes on the embeddings.

Flow of Data: The flow of data through the ResBlock is different from the flow through the fully connected and CNN modules. ResBlocks have skip connections that allow the input to bypass some layers and directly flow to the subsequent layers. These skip connections help in preserving the original information and preventing the loss of critical features during the network’s forward pass. As a result, the ResBlock is more effective in capturing the inherent relationships in the data while mitigating the bias from the protected attributes \( \theta \). In contrast, CNNs involve multiple convolution and pooling operations, which can cause the loss of some information relevant to fairness. The fully connected module, with its dense layers, lacks the skip connections present in the ResBlock, which can lead to less effective bias mitigation.

In conclusion, our ablation study demonstrates that the choice of fairness layer in the Distraction module can significantly impact the fairness and accuracy of the model. It is essential to strike a balance between fairness and accuracy and to select the appropriate fairness layer for the specific dataset and application at hand.

V. CONCLUSION

In this paper, we introduced a novel bias mitigation strategy, the Distraction module, for training deep learning models on various data types, including tabular, image, and graph data. Our proposed approach addresses the problem of unwanted bias in machine learning models, which can lead to algorithmic discrimination and unfairness towards certain population subgroups, particularly minority groups. Unlike existing methods that may result in information loss or fail to balance accuracy and fairness, the Distraction module effectively controls bias while maintaining model performance.

The Distraction module operates within the main neural network, independently managing a subset of weights and optimizing towards specific criteria or alternative objective functions. This in-process bias mitigation method does not require adversarial example generation and can be trained on available datasets without any alterations.

Our empirical evaluation on multiple benchmarks, including the UCI Adult, Heritage Health, POKEC-Z, POKEC-N, NBA, and CelebA datasets, demonstrates the effectiveness of our approach. Compared to state-of-the-art methods in the fairness literature for each dataset, our model exhibits superior performance in minimizing bias while preserving accuracy. We also introduced a new adversarial training procedure that further enhances both accuracy and fairness metrics.

In conclusion, the Distraction module provides a powerful and flexible strategy for mitigating bias in deep learning models. Its ability to operate within the main neural network and optimize specific criteria enables more precise control over bias without sacrificing model performance. As a result, the Distraction module offers a promising approach for addressing the ethical concerns associated with algorithmic bias and promoting fairness in artificial intelligence applications.
VI. ACKNOWLEDGEMENT

REFERENCES

[1] D. Pessach and E. Shmueli, “Algorithmic fairness,” arXiv preprint arXiv:2001.09784, 2020.

[2] C. Dwork, M. Hardt, T. Pitassi, O. Reingold, and R. Zemel, “Fairness through awareness,” in Proceedings of the 3rd innovations in theoretical computer science conference, 2012, pp. 214–226.

[3] M. Yurochkin, A. Bower, and Y. Sun, “Training individually fair ml models with sensitive subspace robustness,” arXiv preprint arXiv:1907.00020, 2019.

[4] R. Zemel, Y. Wu, K. Swersky, T. Pitassi, and C. Dwork, “Learning fair representations,” in International conference on machine learning, PMLR, 2013, pp. 325–333.

[5] C. Louizos, K. Swersky, Y. Li, M. Welling, and R. Zemel, “The variational fair autoencoder,” arXiv preprint arXiv:1511.00830, 2015.

[6] M. Hardt, E. Price, and N. Srebro, “Equality of opportunity in supervised learning,” Advances in neural information processing systems, vol. 29, pp. 3315–3323, 2016.

[7] F. Kamiran and T. Calder, “Data preprocessing techniques for classification without discrimination,” Knowledge and Information Systems, vol. 33, no. 1, pp. 1–33, 2012.

[8] B. T. Luong, S. Ruggieri, and F. Turini, “k-nn as an implementation of situation testing for discrimination discovery and prevention,” in Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining, 2011, pp. 502–510.

[9] M. Feldman, S. A. Friedler, J. Moeller, C. Scheidegger, and S. Venkatasubramanian, “Certifying and removing disparate impact,” in proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining, 2015, pp. 259–268.

[10] A. Tayebi, N. Yousefi, M. Yazdani-Jahromi, E. Kolanthai, C. J. Neal, S. Seal, and O. O. Garibay, “Unbiaseddti: Mitigating real-world bias of drug-target interaction prediction by using deep ensemble-balanced learning,” Molecules, vol. 27, no. 9, p. 2980, 2022.

[11] A. Rajabi and O. O. Garibay, “Tablefairgan: Fair tabular data generation with generative adversarial networks,” arXiv preprint arXiv:2109.00666, 2021.

[12] T. Kamishima, S. Akaho, H. Asoh, and J. Sakuma, “Fairness-aware classifier with prejudice remover regularizer,” in Joint European Conference on Machine Learning and Knowledge Discovery in Databases. Springer, 2012, pp. 35–50.

[13] 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, 2017, pp. 1171–1180.

[14] M. B. Zafar, I. Valera, M. G. Rodrigue, and K. P. Gummadi, “Fairness constraints: Mechanisms for fair classification,” in Artificial Intelligence and Statistics. PMLR, 2017, pp. 962–970.

[15] A. K. Menon and R. C. Williamson, “The cost of fairness in binary classification,” in Conference on Fairness, Accountability and Transparency. PMLR, 2018, pp. 107–118.

[16] S. Corbett-Davies, E. Pierson, A. Sella, S. Goll, 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, 2017, pp. 797–806.

[17] A. Cotter, M. Gupta, H. Jiang, N. Srebro, K. Sridharan, S. Wang, B. Woodworth, and S. You, “Training well-generalizing classifiers for fairness metrics and other data-dependent constraints,” in International Conference on Machine Learning. PMLR, 2019, pp. 1397–1405.

[18] B. H. Zhang, L.emoine, and M. Mitchell, “Mitigating unwanted biases with adversarial learning,” in Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society, 2018, pp. 335–340.

[19] C.-Y. Chung and Y. Mroueh, “Fair mixup: Fairness via interpolation,” arXiv preprint arXiv:2103.06503, 2021.

[20] N. Mehrabi, U. Gupta, F. Mostatter, G. V. Steeg, and A. Galstyan, “Attributing fair decisions with attention interventions,” arXiv preprint arXiv:2109.03952, 2021.

[21] D. Moyer, S. Gao, R. Brekelmans, G. V. Steeg, and A. Galstyan, “Invariant representations without adversarial training,” Advances in Neural Information Processing Systems, volume 31, 9084–9093, 2018.

[22] J. Song, P. Kalluri, A. Grover, S. Zhao, and S. Ermon, “Learning controllable fair representations,” in The 22nd International Conference on Artificial Intelligence and Statistics. PMLR, 2019, pp. 2164–2173.

[23] A. Jaiswal, D. Moyer, G. Ver Steeg, W. AbdAlmageed, and P. Natarajan, “Invariant representations through adversarial forgetting,” in Proceedings of the AAAI Conference on Artificial Intelligence, vol. 34, no. 04, 2020, pp. 4372–4379.

[24] U. Gupta, A. Feber, B. Dilkina, and G. Ver Steeg, “Controllable guarantees for fair outcomes via contrastive information estimation,” in Proceedings of the AAAI Conference on Artificial Intelligence, vol. 35, no. 09, pp. 7610–7619.

[25] Y. Dong, O. Lizardo, and N. V. Chawla, “Do the young live in a ‘smaller world’ than the old? age-specific degrees of separation in a large-scale mobile communication network,” arXiv preprint arXiv:1606.07556, 2016.

[26] T. Rahman, B. Surma, M. Backes, and Y. Zhang, “Fairwalk: Towards fair graph embedding.” 2019.

[27] E. Dai and S. Wang, “Say no to the discrimination: Learning fair graph neural networks with limited sensitive attribute information,” in Proceedings of the 14th ACM International Conference on Web Search and Data Mining, 2021, pp. 680–688.

[28] A. Bose and W. Hamilton, “Compositional fairness constraints for graph embeddings,” in International Conference on Machine Learning. PMLR, 2019, pp. 715–724.

[29] J. Zhao, T. Wang, M. Yatskar, V. Ordonez, and K.-W. Chang, “Men also like shopping: Reducing gender bias amplification using corpus-level constraints,” arXiv preprint arXiv:1709.09457, 2017.

[30] P. Sattigeri, S. C. Hoffman, V. Chenthamarakshan, and K. R. Varshney, “Fairness gan: Generating datasets with fairness properties using a generative adversarial network,” IBM Journal of Research and Development, vol. 63, no. 4/5, pp. 3–1, 2019.

[31] S. Hwang, S. Park, D. Kim, and H. Byun, “Fairfacegan: Fairness-aware facial image-to-image translation,” arXiv preprint arXiv:2012.00282, 2020.

[32] P. C. Roy and V. N. Boddeti, “Mitigating information leakage in image the wild: Reducing racial bias by information maximization adaptation,” arXiv preprint arXiv:2209.08648, 2022.

[33] V. V. Ramaswamy, S. S. Kim, and O. Russakovsky, “Fair attribute classification through latent space de-biasing,” in Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2019, pp. 2586–2594.

[34] A. Rajabi, M. Yazdani-Jahromi, O. O. Garibay, and G. Sukthankar, “Through a fair looking-glass: mitigating bias in image datasets,” arXiv preprint arXiv:2209.08649, 2022.

[35] M. Wang, W. Deng, J. Hu, X. Tao, and Y. Huang, “Racial faces in the wild: Reducing racial bias by information maximization adaptation network,” in Proceedings of the ieee/cvf international conference on computer vision, 2019, pp. 692–702.

[36] M. Wang and W. Deng, “Mitigate bias in face recognition using skewness-aware reinforcement learning,” arXiv preprint arXiv:1911.10692, 2019.

[37] G. Louppe, M. Kagan, and K. Cranmer, “Learning to pivot with adversarial networks,” Proceedings of Advances in Neural Information Processing Systems 30 (NIPS 2017), 2016.

[38] N. Mehrabi, F. Mostatter, N. Saxena, K. Lerman, and A. Galstyan, “A survey on bias and fairness in machine learning,” ACM Computing Surveys (CSUR), vol. 54, no. 6, pp. 1–35, 2021.

[39] V. S. Lokhane, A. K. Akash, S. N. Ravi, and V. Singh, “Fairlam: Augmented lagrangian method for training fair models with little regret,”
in European Conference on Computer Vision. Springer, 2020, pp. 365–381.

[40] D. Dua and C. Graff, “UCI machine learning repository,” 2017. [Online]. Available: [http://archive.ics.uci.edu/ml]

[41] H. Edwards and A. Storkey, “Censoring representations with an adversary,” arXiv preprint arXiv:1511.05897, 2015.

[42] E. Dai and S. Wang, “Learning fair graph neural networks with limited and private sensitive attribute information,” IEEE Transactions on Knowledge and Data Engineering, vol. 35, no. 7, pp. 7103–7117, 2023.

[43] B. Perozzi, R. Al-Rfou, and S. Skiena, “Deepwalk: Online learning of social representations,” in Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining, 2014, pp. 701–710.

[44] L. Takac and M. Zabovsky, “Data analysis in public social networks,” in International scientific conference and international workshop present day trends of innovations, vol. 1, no. 6. Present Day Trends of Innovations Lamza Poland, 2012.

[45] Z. Liu, P. Luo, X. Wang, and X. Tang, “Deep learning face attributes in the wild,” in Proceedings of International Conference on Computer Vision (ICCV), December 2015.

[46] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 770–778.