DISTRACTION IS ALL YOU NEED FOR FAIRNESS

Anonymous authors
Paper under double-blind review

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, ensuring that these models are not biased is vital. 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; Many researchers direct their attention to this issue in different directions, namely, changing data to be statistically independent, adversarial training for restricting the capabilities of a particular competitor who aims to maximize parity, etc. These methods 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 theoretically proven effective in controlling bias from affecting the classification results. This method can be utilized with different data types (e.g., Tabular, images, graphs, etc.). 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). Using state-of-the-art methods proposed in the fairness literature for each dataset, we exhibit our model is superior to these proposed methods in minimizing bias and maintaining accuracy.

1 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 (Pessach & Shmueli, 2020). 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 either past information (Dwork et al., 2012; Yurochkin et al., 2019). Group fairness metrics are measurements of the statistical equality between different subgroups that are characterized by sensitive characteristics such as race, or gender (Zemel et al., 2013; Louizos et al., 2015; Hardt et al., 2016). In this paper, we focus on group fairness, and from this point on, we refer to group fairness as fairness.

Fairness approaches are commonly categorized into three groups: (1) pre-process approaches: these methods consist of changing the data before training the models to improve model outcomes in terms of fairness. Preliminary studies comprise changing the labels or reweighing them, such as (Kamiran & Calders, 2012; Luong et al., 2011) or adjusting the features to minimize differences in the distribution of privileged and unprivileged groups so that it becomes more challenging for the classifier to differentiate the two groups (Feldman et al., 2015; Tayebi et al., 2022). More recently, Rajabi & Garibay (2021) proposed a generative adversarial network to produce unbiased tabular datasets by altering the value function of the generator network, to account for accuracy and fairness. (2) In-process approaches: these methods modify the algorithm during the training to achieve more unbiased classifiers. These methods suggest adding regularization terms to the objective function to

secure fairness. For instance, Kamishima et al. (2012) utilized a regularization term that penalized the mutual information between protected attributes and the classifier's prediction and used a tuning parameter to adjust the trade-off between fairness and accuracy. Zafar et al. (2017a;b) proposed adding a constraint to the model that needs to satisfy a proxy for equalized odds. Moreover, (3) post-process approaches: these techniques are applied after training and modify the unfair outcomes. For instance, Hardt et al. (2016) suggested a method to flip some of the outcomes to improve the fairness of the classifier. Menon & Williamson (2018); Corbett-Davies et al. (2017) in parallel proposed utilizing different thresholds for the privileged and unprivileged groups to optimize the trade-off between accuracy and fairness.

1.1 TABULAR

There are numerous studies that focus on fairness in training on tabular datasets. Cotter et al. (2019) enforce independence by regularizing the covariance between predictions and sensitive variables which reduces the amount of variation in the relationship between the two. In order to reduce the amount of variation that exists across the different groups, Zafar et al. (2017a) standardized the decision bounds of a convex margin-based classifier. Zhang et al. (2018) reduced the effects of bias by restricting an adversary's capacity to infer sensitive characteristics based on predictions. However, since these constraints are reliant on the data, even if they are met when the model is being trained, the model may exhibit a different behavior while it is being evaluated. Agarwal et al. (2018) investigate the generalization error of fair classifiers that were produced via two-player games. Cotter et al. (2019) inherit the two-player situation while simultaneously training each player on two distinct datasets in order to increase the generalizability of their findings. Despite the analytic answers and theoretical assurances, it may be challenging to scale game-theoretic techniques for more complicated model classes (Chuang & Mroueh, 2021).
Mehrabi et al. (2021a) showed the effect of proxy attributes leading to indirect unfairness using their attention-based approach. Using that, they utilized a post-processing approach for bias mitigation that reduces the weight of the attributes more responsible for unfairness. Avoiding increasing levels of complexity introduced by adversarial training, Moyer et al. (2018) use mutual information instead of adversarial training to achieve invariant representations of data concerning specified factors. Song et al. (2019) proposed an information-theoretic method which, by exploiting information-theoretic and adversarial methods, could achieve controllable fair representations of data with respect to the notion of demographic parity. By utilizing a forget-gate in a neural network architecture similar to forget gates in LSTMs, Jaiswal et al. (2020) propose adversarial forgetting to fairness. Finally, Gupta et al. (2021) utilizes certain estimates for contrasting information in order to optimize theoretical objectives that may be used to determine suitable trade-offs between demographic parity and accuracy in the statistical population.

1.2 GRAPH

The message-passing structure of GNNs and the topology of graphs both have the potential to amplify the bias. In general, in graphs such as social networks, nodes with sensitive features similar to one another are more likely to link to one another than nodes with sensitive attributes dissimilar from one another (Dong et al., 2016; Rahman et al., 2019). On social networks, for instance, persons of younger generations have a higher tendency to form friendships with others of a similar age (Dong et al., 2016). This results in the aggregation of neighbors' features in GNN having similar representations for nodes of similar sensitive information while having different representations for nodes of different sensitive features, which leads to severe bias in decision making, in the sense that the predictions are highly correlated with the sensitive attributes of the nodes. GNNs have a greater bias due to the adoption of graph structure than models which employ node characteristics (Dai & Wang, 2021). Because of this bias, the widespread use of GNNs in areas such as the evaluation of job candidates (Mehrabi et al., 2021b) and the prediction of drug-target interaction (Yazdani-Jahromi et al., 2022; Yousefi et al., 2022) would be significantly hindered. As a result, it is essential to research equitable GNNs. The absence of sensitive information presents significant problems to the work that has already been done on fair models (Beutel et al., 2017; Creager et al., 2019; Locatello et al., 2019; Louizos et al., 2015). Despite the significant amount of work that has been put into developing fair models through the revision of features Kamiran & Calders (2009; 2012); Zhang et al. (2017), disentanglement (Creager et al., 2019; Louizos et al., 2015), adversarial debiasing (Beutel et al., 2017; Edwards & Storkey, 2015), and fairness constraints (Zafar et al., 2017a;b), these models are

almost exclusively designed for independently and identically distributed (i.i.d) data, meaning that they are unable to be directly applied to graph data due to the fact that they do not simultaneously take into consideration the bias that comes from node attributes and graph. In recent years, Bose & Hamilton (2019); Rahman et al. (2019) have been published with the aim of learning fair node representations from graphs. These approaches only deal with simple networks that do not have any properties on any of the nodes, and they place their emphasis on fair node representations rather than fair node classifications. Finally, Dai & Wang (2021) used graph topologies and a restricted amount of protected attributes and designed FairGNN to reduce the bias of GNNs while retaining high node classification accuracy.

1.3 VISION

The challenges caused by bias in computer vision might appear in various ways. It has been found, for instance, that in action recognition models, when the data include gender bias, the bias is exacerbated by the models trained on such datasets (Zhao et al., 2017). Face detection and recognition models may be less precise for some racial and gender categories (Buolamwini & Gebru, 2018). Methods for mitigating bias in vision datasets are suggested in (Wang et al., 2022) and (Yang et al., 2020). Several researchers have used GANs on image datasets for bias reduction. Sattigeri et al. (2019) altered the utility function of GAN in order to generate equitable picture datasets. FairFace-GAN (Hwang et al., 2020) provides facial image-to-image translation to avoid unintended transfer of protected characteristics. Roy & Boddeti (2019) developed a method to mitigate information leakage on image datasets by formulating the problem as an adversarial game to maximize data utility and minimizing the amount of information contained in the embedding, measured by entropy. Ramaswamy et al. (2021) present a methodology to generate balanced training data for each protected property by perturbing the latent vector of a GAN. Other experiments using GANs to generate accurate data are (Choi et al., 2020; Sharmanska et al., 2020). Beyond GANs, many strategies have addressed the challenge of AI fairness. (Rajabi et al., 2022) proposed an U-Net for creating unbiased image data. Deep information maximization adaption networks were employed to eliminate racial bias in face vision datasets (Wang et al., 2019a), while reinforcement learning was utilized for training a race-balanced network (Wang & Deng, 2019). Wang et al. (2021) offer a generative few-shot cross-domain adaptation method for performing fair cross-domain adaptation and enhancing minority category performance. The research in (Xu et al., 2021) recommends adding a penalty term to the softmax loss function to reduce bias and enhance face recognition fairness performance. Quadrianto et al. (2019) describe a technique for discovering fair data representations with the same semantic information as the original data. There have also been effective applications of adversarial learning for this purpose (Wang et al., 2019b; Zhang et al., 2018). (Chuang & Mroueh, 2021) proposed fair mixup, which uses data augmentation to mitigate bias in data. In this work, we introduced a novel in-process bias mitigation method that does not need adversarial example generation to generate results and can be trained on available datasets without alteration. Our **contributions** in this work are: (1)We proposed a new model designed to mitigate data bias effects on neural networks. (2) We provide an adversarial procedure with theoretical analysis for training networks with attention modules. (3) The introduced model using the new adversarial training procedure significantly improves the state-of-the-art accuracy and fairness metrics over the curve accuracy and statistical parity in all types of datasets. (4) The proposed method using the self-attention mechanism increases explainability and assists in identifying proxy attributes.

2 METHODOLOGY

We concentrate on 2-class classification problems in this paper. This assumption is not limiting because the proposed method can be generalized to datasets with more than two classes. Suppose we have a dataset $D = \{x^{(i)}, a^{(i)}, y^{(i)}\}$; $x^{(i)}$, $y^{(i)}$, and $a^{(i)}$ are i.i.d. samples from a distribution of the data $P(x, y, a)$. a is discrete with finite values, which are the protected attributes of the dataset. $y \in \{0, 1\}$ is the label, and x denotes other features in the dataset. This section introduces a new approach for training a fair classifier based on deep neural networks. The proposed deep learning model includes two sets of weights for a classifier neural network. This model can consist of different neural network layers (FF, GCN, CNN, ...). The first set of weights is related to the classification task of the DL model, and the second is related to the Distraction module, which is the fairness part of the model. The first set of weights is trained to maximize the classifier's accuracy, and the second

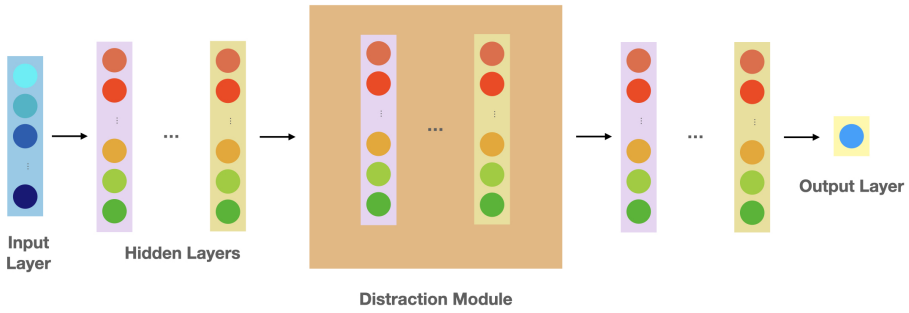


Figure 1: Depiction of Distraction module in MLP model: weights of the layers in the orange box are controlled by the Distraction module and are optimized with 3 objective function. The black arrows are weights that are controlled by the main classifier.

set is trained to minimize demographic parity. The model's architecture is depicted in Figure 1. Fairness in classification means that the predicted class of the data should be invariant to the protected attribute. This can be formalized as 1 Louppe et al. (2016).

$$p(C(x, D(x;\theta_d);\theta_c) = s|a, y) = p(C(x, D(x;\theta_d);\theta_c) = s|a', y) \tag{1}$$

The proposed method is formulated as a maximin game with a global optimal that meets specific criteria. The weights of the Distraction module are entirely isolated from the leading network. The Distraction module is trying to make the classifier function results as fair as possible, while the whole network is trying to make the classifier function results as accurate as possible. The first player is the Distraction module, and the second is the 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 $U_1(D, C)$ 2 for player C and $U_2(D, C)$ 3 for player D.

$$\min_{\theta_c} U_1(C, D) = \mathbb{E}_{x\sim X}\mathbb{E}_{y\sim Y}\Big[-\log p_{\theta_d,\theta_c}(y|x))\Big] \tag{2}$$

$$\max_{\theta d} U_2(C, D) = -\mathbb{E}_{s\sim C(x,D(x;\theta_d);\theta_c)}\Big[\mathbb{E}_{a\sim A}\Big[-\log p_{\theta_d,\theta_c}(a|s)\Big]\Big] \tag{3}$$

Proposition 2.1. *If there exist an optimal solution for minimax problem with utility functions 2 and 3, the optimal $C(x, D(x))$ is an optimal classifier and invariant to the protected attributes.*

Proof. For θ_c fixed. the D network is optimal at:

$$\hat{\theta}_d = \arg\max_{\theta_d} U_2(C, D) \tag{4}$$

Considering

$$p_{\theta_c,\theta_d}(a|s) = \frac{p_{\theta_c,\theta_d}(s|a)p_{\theta_c,\theta_d}(s)}{p(a)} \tag{5}$$

We have:

$$\max_{\theta d} U_2(C, D) = -\mathbb{E}_{s\sim C(x,D(x;\theta_d);\theta_c)}\mathbb{E}_{a\sim A}\Big[-\log\frac{p_{\theta_c,\theta_d}(s|a)p_{\theta_c,\theta_d}(s)}{p(a)}\Big]$$
$$= -\mathbb{E}_{s\sim C(x,D(x;\theta_d);\theta_c)}\mathbb{E}_{a\sim A}\Big[-\log p_{\theta_c,\theta_d}(s|a)p_{\theta_c,\theta_d}(s)\Big] + \mathbb{E}_{a\sim A}[\log p(a)]$$

the $\mathbb{E}_{a\sim A}[\log p(a)]$ is constant; therefore, it can be omitted from the objective function without affecting the θ_d consequently:

$$\max_{\theta d} U_2(C,D) = -\mathbb{E}_{s\sim C(x,D(x;\theta_d);\theta_c)}\mathbb{E}_{a\sim A}\Big[-\log p_{\theta_c,\theta_d}(s|a)p_{\theta_c,\theta_d}(s)\Big]$$

$$= -\mathbb{E}_{s\sim C(x,D(x;\theta_d);\theta_c)}[-\log p_{\theta_c,\theta_d}(s)] - \mathbb{E}_{s\sim C(x,D(x;\theta_d);\theta_c)}\mathbb{E}_{a\sim A}\big[-\log p_{\theta_c,\theta_d}(s|a)\big] \quad (6)$$

$$= -H(C(x,D(x;\theta_d);\hat{\theta}_c)) - H(C(x,D(x;\theta_d);\hat{\theta}_c)|A) \quad (7)$$

For $\hat{\theta}c$ function $U_1(C,D)$ is optimal which means it's value reduces to $H(Y|X)$. With $\hat{\theta}c$ fixed, $H(C(x,D(x;\theta_d);\theta_c))$ is reduced to $H(Y|X)$ for sufficiently small changes in θ_d. The function of $U_2(C,D)$ is maximized when $H(C(x,D(x;\theta_d);\hat{\theta}_c)|A)$ is reduced to $H(C(x,D(x;\hat{\theta}_d);\hat{\theta}_c)$ or less which means the $C(x,D(x;\theta_d);\hat{\theta}_c)$ and A are independent.
Hence, the $C(x,D(x))$ is an optimal classifier, and this function is invariant to protected attributes; This concludes the proof.

□

Proposition 2.1 suggests that if the weights of the main network is allowed to reach it's optimal $C(x,D(x))$ and $D(x)$ is updated to improve 3 with sufficiently small steps, consequently $C(x,D(x))$ converges to a accurate classifier invariant to the protected attribute.
In practical cases, the classifier cannot be fair and optimal simultaneously; This is due to the fact that the protected and proxy attributes contribute significantly to the decision made by the classifier. We introduce the parameter η to create a trade-off between accuracy and fairness. η is the coefficient of demographic parity loss 8, which is being used by the Distraction module. When the η is near 0, the classifier tends to be more accurate than fair. Higher η means the classifier prioritizes fairness over accuracy, which indicates that manipulating η results in the generation of Pareto solutions for this multi-objective optimization problem.

$$\eta \sum_{i=1}^{m} -\log p_{\theta_d,\theta_c}(a^{(i)}|C(x^{(i)},D(x^{(i)}))) \quad (8)$$

Naturally, as mentioned earlier, we can develop a mini-batch stochastic gradient descent procedure to train the network. This procedure can be found in the appendix A.1.

3 EXPERIMENTS

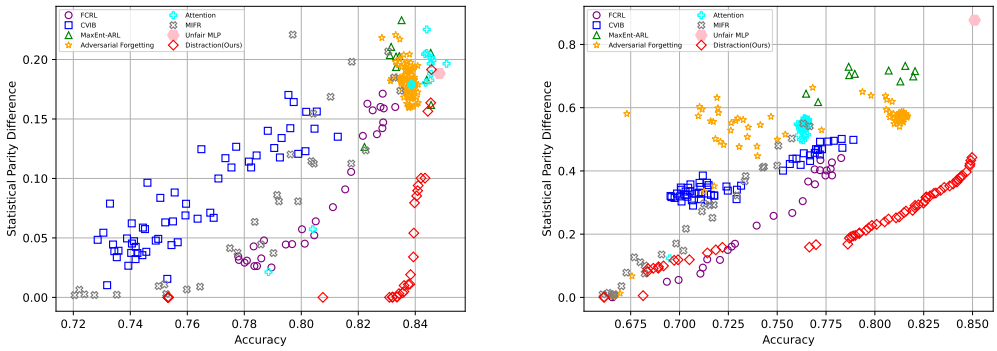
This section compares our method to other benchmark methods in the literature. The experiment section consists of three parts. First, we experiment with tabular datasets. We compare the classification accuracy and statistical parity with the benchmark datasets. In the second section, we use graph data for node classification tasks. We evaluate our model on vision datasets in the third and final section.
The appendix A.3 contains information on the baselines and experimental settings for each dataset and evaluation metrics used for comparison. The Distraction module used on all the datasets reported here 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, ablation study, and the additional results can be found in the appendix A.5.

3.1 TABULAR

We evaluated our method for bias mitigation to a range of current state-of-the-art approaches. We concentrate on strategies specifically tuned to achieve the best results in statistical parity metrics on tabular studies.
We used two well-known benchmark datasets in this field for our experiments which are as follows:
UCI Adult Dataset Dua & Graff (2017) This dataset is based on demographic data gathered in 1994, including a train set of 30000 and a test set of 15,000 samples. The goal is to forecast if the salary is more than $50,000 yearly, and the binary protected attribute is the gender of samples gathered in the dataset.
Heritage Health Predicting the Charleson Index, a measure of a patient's 10-year mortality, is

Table 1: Area over the curve of statistical demographic parity and accuracy (Higher is better). The proposed model (Distraction) significantly outperforms other benchmark models in this quantitative metric.

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
ADV FORGETTING (AAAI 2020)	0.077	0.172



(a) UCI Adult dataset

(b) Heritage Health dataset

Figure 2: Graph of the accuracy of different benchmark models and distraction model vs. statistical demographic parity of each model for (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 datasets.

the aim of this dataset. The Heritage Health dataset contains samples from roughly 51,000 patients which 41000 of them are in the training set and 11000 of them are in the test set, and the goal is to predict the Charleson Index. The protected attribute, which has nine potential values, is age.
We use training sets to run the Distraction module technique and train the network, which includes the Distraction Module. Then we assess the performance by running classifiers for subsequent prediction tasks. We provide average accuracy as a measure of most probable accuracy and maximum demographic parity as a measure of worst-case scenario bias, calculated across five iterations of the training process using random seeds. Unlike Gupta et al. (2021), we did not use any preprocessing on the data before feeding it to our network. Reported results for our model are Pareto solutions for the neural network during training with different ηs. Results are reported for methods with a multi-layer perceptron classifier with two hidden layers. We compare our results with the following state-of-the-art methods as the benchmark: CVIB Moyer et al. (2018) achieves this goal using a conditional variational autoencoder, MIFR Song et al. (2019) optimizes the fairness goal with a mix of information bottleneck factor and adversarial learning. FCRL Gupta et al. (2021) uses specific approximations for contrastive information to maximize theoretical goals that may be utilized to make appropriate trade-offs over statistical, demographic parity, and precision. Further, we investigated baselines including MaxEnt-ARL Roy & Boddeti (2019) and Adversarial Forgetting Jaiswal et al. (2020) that employ adversarial learning.
To provide a quantitative measure, we used the Gupta et al. (2021) method for the normalized area over the parity-accuracy curve. It is the standardized volume of the feasible statistical demographic parity vs. accuracy area, and it should be maximized via an effective fairness classification approach. Table 1 is the table of aforementioned quantitative metric. Our method (Distraction) is significantly

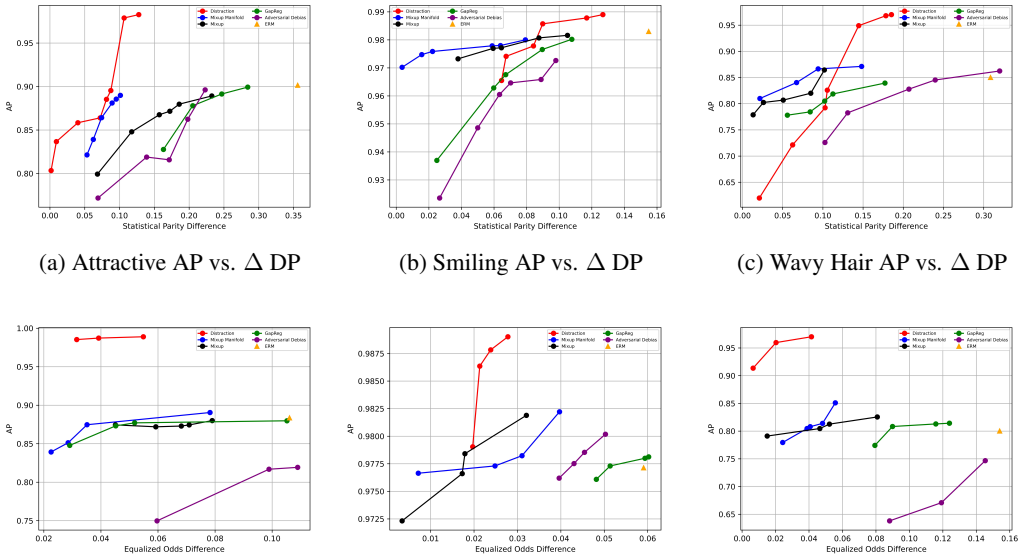
Table 2: The comparisons of our proposed method with the baselines on graph datasets.

METHODS	POKEC-Z		POKEC-N		NBA	
	ACCURACY	Δ_{DP}	ACCURACY	Δ_{DP}	ACCURACY	Δ_{DP}
GAT (BASELINE)	0.704	0.091	0.703	0.094	0.719	0.102
DISTRACTION (OURS)	**0.702**	**0.001**	**0.703**	0.015	**0.717**	**0.002**
FAIRGAT	0.701	0.005	0.700	**0.006**	0.715	0.007
ALFR	0.654	0.028	0.631	0.0305	0.643	0.023
ALFR-E	0.680	0.058	0.662	0.041	0.660	0.047
DEBIAS	0.652	0.019	0.626	0.024	0.631	0.025
DEBIAS-E	0.675	0.047	0.656	0.036	0.656	0.053
FCGE	0.659	0.031	0.648	0.041	0.660	0.029

better than other methods statistically.
Figures 2a and 2b show trade-offs of the statistical demographic parity vs. accuracy associated with various bias reduction strategies in the UCI Adult dataset and Heritage Health dataset, respectively. The ideal area of the graph for the result of a method is to measure how much the curve is located in the lower right corner of the graph, which means accurate and fair results concerning protected attributes. Results indicate that the Distraction method outperforms other methods significantly. Additionally, our bias reduction framework, which is built on adversarial training, is the best mitigation strategy for tabular data. Moreover, all of the examined methods except for Attention are built to reduce data bias and lack interpretability. Our method can be interpretable with the help of the self-attention mechanism while reaching state-of-the-art in both fairness and accuracy.

3.2 GRAPH

We compare our suggested framework with the cutting-edge approaches for fair classification, and fair graph embedding learning, including ALFR Edwards & Storkey (2015), which is a pre-processing technique. The sensitive information in the representations created by an MLP-based autoencoder is eliminated using a discriminator. Then the debiased representations are used to train the linear classifier. ALFR-e, which is a method to make use of the graph structure information, ALFR-e joins the user features in the ALFR with the graph embeddings discovered by deepwalk (Perozzi et al., 2014). Debias Zhang et al. (2018), which is a fair categorization technique used throughout processing. It immediately applies a discriminator to the predicted likelihood of the classifier. Debias-e, which is similar to the ALFR-e, this method also includes deepwalk embeddings into the Debias characteristics. FCGE (Bose & Hamilton, 2019), which is suggested as a method for learning fair node embeddings in graphs without node characteristics. Discriminators screen out the delicate data in the embeddings. We used the Dai & Wang (2021) study's obtained datasets for our investigation which are as follows:
Pokec (Takac & Zabovsky, 2012) is among the most well-known social network in Slovakia and resembles Facebook and Twitter greatly. This dataset includes anonymous information from the whole social network for the year 2012. User profiles on Pokec include information on gender, age, interests, hobbies, profession, and more. There are millions of users in the original Pokec dataset. Sampled Pokec-z and Pokec-n datasets are based on the provinces that users are from. Users from two crucial regions of the corresponsive provinces make up Pokec-z and Pokec-n. The region is handled as a sensitive property. The categorization task involves predicting the users' working environment.
NBA is a Kaggle dataset 1 with about 400 NBA basketball players served as the basis for this extension. Players' 2016–2017 season success statistics are presented, along with additional details like nationality, age, and income. They gathered the relationships between NBA basketball players on Twitter using its official crawling API to create the graph connecting the NBA players. They separated the nationality into two groups, American players and international players, which is a sensitive characteristic. The classification job is to predict whether a player's wage is above the median. Dai & Wang (2021) removed any nodes with no connections to other nodes across all datasets. As validation sets and test sets, they randomly choose 25% and 50% of nodes in Pokec-z, Pokec-n, and NBA that have both sensitive characteristics and labels. The validation sets and test



(a) Attractive AP vs. Δ DP (b) Smiling AP vs. Δ DP (c) Wavy Hair AP vs. Δ DP

(d) Attractive AP vs. Δ EO (e) Smiling AP vs. Δ EO (f) Wavy Hair AP vs. Δ EO

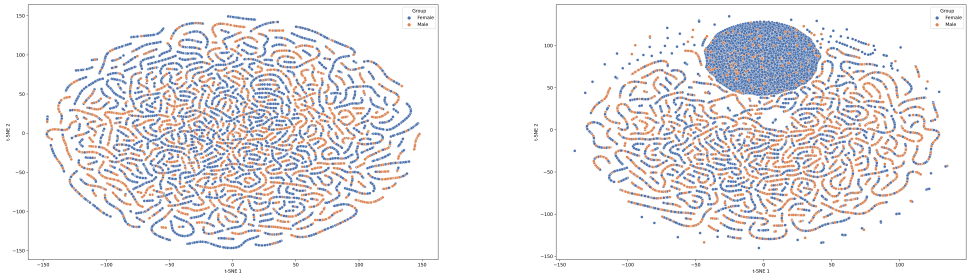
Figure 3: *Dataset for Celebs*. Each task's first and second rows reflect the trade-off between AP and DP/EO. Across all tasks, the Distraction module is significantly better than the baseline in EO and performs better in higher AP on DP.

sets don't overlap with VL and VS, as should be noted. They also provided the minority group to majority group ratio and the number of edges connecting one group to another. Sensitive attributes are skewed, and most relationships are between people who share the same sensitive attribute.

Each experiment is done five times, and table 2 reports the mean of the runs. We may deduce from the table that, in comparison to GAT, generic fair classification techniques and the graph embeddings learning approach exhibit subpar classification performance even when using graph information, while Distraction performs highly similar to the baseline GNNs. FairGCN is quite close to the baseline, but the Distraction technique is also outperforming it. When sensitive information is limited, baselines exhibit obvious bias, and graph-based baselines are much worse. On the other hand, the statistical demographic parity that our suggested model provides is near 0, indicating that discrimination has essentially been eradicated.

3.3 Vision

We compare our method on vision task with (1) empirical risk minimization (ERM), which accomplish training task without any regularization, (2) gap regularization, which directly regularizes the model, (3) adversarial debiasing (Zhang et al., 2018), and (4) Fairmixup (Chuang & Mroueh, 2021). We used CelebA face attributes dataset (Liu et al., 2015) to demonstrate the potency of our method. Over 200,000 celebrity photos of faces can be found in CelebA, where each image is assigned 40 human-labeled binary characteristics, including gender. We choose attractive, smiling, and wavy hair among the attributes and utilize them in three binary classification tasks, considering gender as the protected attribute. We chose these three characteristics because each of these activities has a sensitive group that receives a disproportionately high number of positive samples. We train a ResNet-18 He et al. (2016) for each task in addition to two hidden layers for the outcome prediction. Figure 3's top row depicts how AP and DP are traded off for each assignment. Once again, Distraction consistently performs far better than the baselines. In figure 3, in the second row, are the tradeoffs between AP and EO. Although without mixup augmentation, adversarial debiasing, and gap regularization have similar trajectories and bigger DPs, the Distraction module outperforms them significantly. Additionally, we see that the Distraction curve has the shortest slope and hence the smallest DP. Our method is performing on par if not better than the Fairmixup, which is a pre-



(a) With the Distraction module $\tilde{z}$

(b) Without the Distraction module z

Figure 4: CelebA Dataset – t-SNE visualization of z and $\tilde{z}$ labeled with gender classes. The invariant encoding $\tilde{z}$ shows no clustering by gender. These plots are generated using attractive attribute.

processing method. The output of each layer with and without the Distraction module and ResNet-18 filters can be found in the appendix A.6.
Furthermore, we demonstrate the power of the Distraction module by visualizing the t-SNE plot. The t-SNE visualization of z (output of the ResNet-18 before the classification layer without the Distraction module) and $\tilde{z}$ (output of the ResNet-18 before the classification layer with the Distraction module) is shown in Figure 4, demonstrating that z clusters by gender, but $\tilde{z}$ does not.

4 EXPLAINABILITY

Our method not only significantly outperforms all the bias mitigation strategies, including the only interpretable approach to the best of our knowledge (Attention Mehrabi et al. (2021a)) but also the method can be used to identify proxy attributes which are exacerbating the bias in the classifier besides the protected attribute. To explain our method, we can use the cross-attention module in the Distraction module. We use the cross-attention with the pre-trained classifier for the protected attribute and the output of the classifier, with the extraction of the Distraction module's weights matrix and multiplying this matrix with the pre-trained protected attribute classifier weights to get the heat map for the UCI adult dataset. The model architecture and the heatmap can be found in the appendix A.4. This heat map shows the amount of contribution of each attribute to the final layer of the classifier. If the absolute value is greater than 0, it means that the attribute is contributing toward the final result. We used the mean of each column and used a threshold of (-0.1, 0.1) to determine the proxy features in the dataset. For the UCI Adult dataset, "age," "race," "education," "capital loss," and "native country" is detected as proxy attribute. The nature of detected attributes for the adult dataset has the potential to introduce bias into the classifier model.

5 CONCLUSION

We introduce a new learning procedure for learning a fair and accurate classifier through the Distraction module. We mathematically prove that the proposed method is effective for achieving an accurate and fair classifier. We demonstrate that the proposed method is highly effective in different data types. We leveraged the cross-attention mechanism to achieve interpretability and identify proxy attributes and their effect on the model outcomes. This method not only shows promising results on benchmark datasets on a tabular, graph, and vision in the fairness domain but also provides insight into protected and proxy attributes. Finally, fairness in AI and Human-Centered AI design is essential. This work contributes toward a better understanding of neural network architectures designed for accuracy, fairness, and explainability.

REFERENCES

Alekh Agarwal, Alina Beygelzimer, Miroslav Dudík, John Langford, and Hanna Wallach. A reductions approach to fair classification. In *International Conference on Machine Learning*, pp.

60–69. PMLR, 2018.

Alex Beutel, Jilin Chen, Zhe Zhao, and Ed H Chi. Data decisions and theoretical implications when adversarially learning fair representations. *arXiv preprint arXiv:1707.00075*, 2017.

Avishek Bose and William Hamilton. Compositional fairness constraints for graph embeddings. In *International Conference on Machine Learning*, pp. 715–724. PMLR, 2019.

Joy Buolamwini and Timnit Gebru. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on fairness, accountability and transparency*, pp. 77–91. PMLR, 2018.

Kristy Choi, Aditya Grover, Trisha Singh, Rui Shu, and Stefano Ermon. Fair generative modeling via weak supervision. In *International Conference on Machine Learning*, pp. 1887–1898. PMLR, 2020.

Ching-Yao Chuang and Youssef Mroueh. Fair mixup: Fairness via interpolation. *arXiv preprint arXiv:2103.06503*, 2021.

Sam Corbett-Davies, Emma Pierson, Avi Feller, Sharad Goel, and Aziz Huq. Algorithmic decision making and the cost of fairness. In *Proceedings of the 23rd acm sigkdd international conference on knowledge discovery and data mining*, pp. 797–806, 2017.

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 *International Conference on Machine Learning*, pp. 1397–1405. PMLR, 2019.

Elliot Creager, David Madras, Jörn-Henrik Jacobsen, Marissa Weis, Kevin Swersky, Toniann Pitassi, and Richard Zemel. Flexibly fair representation learning by disentanglement. In *International conference on machine learning*, pp. 1436–1445. PMLR, 2019.

Enyan Dai and Suhang 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*, pp. 680–688, 2021.

Yuxiao Dong, Omar Lizardo, and Nitesh 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.

Dheeru Dua and Casey Graff. UCI machine learning repository, 2017. URL `http://archive.ics.uci.edu/ml`.

Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. Fairness through awareness. In *Proceedings of the 3rd innovations in theoretical computer science conference*, pp. 214–226, 2012.

Harrison Edwards and Amos Storkey. Censoring representations with an adversary. *arXiv preprint arXiv:1511.05897*, 2015.

Michael Feldman, Sorelle A Friedler, John Moeller, Carlos Scheidegger, and Suresh Venkatasubramanian. Certifying and removing disparate impact. In *proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining*, pp. 259–268, 2015.

Umang Gupta, Aaron Ferber, Bistra Dilkina, and Greg Ver Steeg. Controllable guarantees for fair outcomes via contrastive information estimation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pp. 7610–7619, 2021.

Moritz Hardt, Eric Price, and Nati Srebro. Equality of opportunity in supervised learning. *Advances in neural information processing systems*, 29:3315–3323, 2016.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Sunhee Hwang, Sungho Park, Dohyung Kim, Mirae Do, and Hyeran Byun. Fairfacegan: Fairness-aware facial image-to-image translation. *arXiv preprint arXiv:2012.00282*, 2020.

Ayush Jaiswal, Daniel Moyer, Greg Ver Steeg, Wael AbdAlmageed, and Premkumar Natarajan. Invariant representations through adversarial forgetting. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pp. 4272–4279, 2020.

Faisal Kamiran and Toon Calders. Classifying without discriminating. In *2009 2nd international conference on computer, control and communication*, pp. 1–6. IEEE, 2009.

Faisal Kamiran and Toon Calders. Data preprocessing techniques for classification without discrimination. *Knowledge and Information Systems*, 33(1):1–33, 2012.

Toshihiro Kamishima, Shotaro Akaho, Hideki Asoh, and Jun Sakuma. Fairness-aware classifier with prejudice remover regularizer. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pp. 35–50. Springer, 2012.

Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of International Conference on Computer Vision (ICCV)*, December 2015.

Francesco Locatello, Gabriele Abbati, Thomas Rainforth, Stefan Bauer, Bernhard Schölkopf, and Olivier Bachem. On the fairness of disentangled representations. *Advances in Neural Information Processing Systems*, 32, 2019.

Vishnu Suresh Lokhande, Aditya Kumar Akash, Sathya N Ravi, and Vikas Singh. Fairalm: Augmented lagrangian method for training fair models with little regret. In *European Conference on Computer Vision*, pp. 365–381. Springer, 2020.

Christos Louizos, Kevin Swersky, Yujia Li, Max Welling, and Richard Zemel. The variational fair autoencoder. *arXiv preprint arXiv:1511.00830*, 2015.

Gilles Louppe, Michael Kagan, and Kyle Cranmer. Learning to pivot with adversarial networks. *Proceedings of Advances in Neural Information Processing Systems 30 (NIPS 2017)*, 2016.

Binh Thanh Luong, Salvatore Ruggieri, and Franco 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*, pp. 502–510, 2011.

Ninareh Mehrabi, Umang Gupta, Fred Morstatter, Greg Ver Steeg, and Aram Galstyan. Attributing fair decisions with attention interventions. *arXiv preprint arXiv:2109.03952*, 2021a.

Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. A survey on bias and fairness in machine learning. *ACM Computing Surveys (CSUR)*, 54(6):1–35, 2021b.

Aditya Krishna Menon and Robert C Williamson. The cost of fairness in binary classification. In *Conference on Fairness, Accountability and Transparency*, pp. 107–118. PMLR, 2018.

Daniel Moyer, Shuyang Gao, Rob Brekelmans, Greg Ver Steeg, and Aram Galstyan. Invariant representations without adversarial training. *Advances in Neural Information Processing Systems, volume 31, 9084–9093*, 2018.

Bryan Perozzi, Rami Al-Rfou, and Steven Skiena. Deepwalk: Online learning of social representations. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 701–710, 2014.

Dana Pessach and Erez Shmueli. Algorithmic fairness. *arXiv preprint arXiv:2001.09784*, 2020.

Novi Quadrianto, Viktoriia Sharmanska, and Oliver Thomas. Discovering fair representations in the data domain. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 8227–8236, 2019.

Tahleen Rahman, Bartlomiej Surma, Michael Backes, and Yang Zhang. Fairwalk: Towards fair graph embedding. 2019.

Amirarsalan Rajabi and Ozlem Ozmen Garibay. Tabfairgan: Fair tabular data generation with generative adversarial networks. *arXiv preprint arXiv:2109.00666*, 2021.

Amirarsalan Rajabi, Mehdi Yazdani-Jahromi, Ozlem Ozmen Garibay, and Gita Sukthankar. Through a fair looking-glass: mitigating bias in image datasets. *arXiv preprint arXiv:2209.08648*, 2022.

Vikram V Ramaswamy, Sunnie SY Kim, and Olga Russakovsky. Fair attribute classification through latent space de-biasing. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 9301–9310, 2021.

Proteek Chandan Roy and Vishnu Naresh Boddeti. Mitigating information leakage in image representations: A maximum entropy approach. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2586–2594, 2019.

Prasanna Sattigeri, Samuel C Hoffman, Vijil Chenthamarakshan, and Kush R Varshney. Fairness gan: Generating datasets with fairness properties using a generative adversarial network. *IBM Journal of Research and Development*, 63(4/5):3–1, 2019.

Viktoriia Sharmanska, Lisa Anne Hendricks, Trevor Darrell, and Novi Quadrianto. Contrastive examples for addressing the tyranny of the majority. *arXiv preprint arXiv:2004.06524*, 2020.

Jiaming Song, Pratyusha Kalluri, Aditya Grover, Shengjia Zhao, and Stefano Ermon. Learning controllable fair representations. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pp. 2164–2173. PMLR, 2019.

Lubos Takac and Michal Zabovsky. Data analysis in public social networks. In *International scientific conference and international workshop present day trends of innovations*, volume 1. Present Day Trends of Innovations Lamza Poland, 2012.

Aida Tayebi, Niloofar Yousefi, Mehdi Yazdani-Jahromi, Elayaraja Kolanthai, Craig J Neal, Sudipta Seal, and Ozlem Ozmen Garibay. Unbiaseddti: Mitigating real-world bias of drug-target interaction prediction by using deep ensemble-balanced learning. *Molecules*, 27(9):2980, 2022.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention Is All You Need. *arXiv*, 2017.

Angelina Wang, Alexander Liu, Ryan Zhang, Anat Kleiman, Leslie Kim, Dora Zhao, Iroha Shirai, Arvind Narayanan, and Olga Russakovsky. Revise: A tool for measuring and mitigating bias in visual datasets. *International Journal of Computer Vision*, pp. 1–21, 2022.

Mei Wang and Weihong Deng. Mitigate bias in face recognition using skewness-aware reinforcement learning. *arXiv preprint arXiv:1911.10692*, 2019.

Mei Wang, Weihong Deng, Jiani Hu, Xunqiang Tao, and Yaohai 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*, pp. 692–702, 2019a.

Tianlu Wang, Jieyu Zhao, Mark Yatskar, Kai-Wei Chang, and Vicente Ordonez. Balanced datasets are not enough: Estimating and mitigating gender bias in deep image representations. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 5310–5319, 2019b.

Tongxin Wang, Zhengming Ding, Wei Shao, Haixu Tang, and Kun Huang. Towards fair cross-domain adaptation via generative learning. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 454–463, 2021.

Xingkun Xu, Yuge Huang, Pengcheng Shen, Shaoxin Li, Jilin Li, Feiyue Huang, Yong Li, and Zhen Cui. Consistent instance false positive improves fairness in face recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 578–586, 2021.

Kaiyu Yang, Klint Qinami, Li Fei-Fei, Jia Deng, and Olga Russakovsky. Towards fairer datasets: Filtering and balancing the distribution of the people subtree in the imagenet hierarchy. In *Proceedings of the 2020 conference on fairness, accountability, and transparency*, pp. 547–558, 2020.

Mehdi Yazdani-Jahromi, Niloofar Yousefi, Aida Tayebi, Elayaraja Kolanthai, Craig J Neal, Sudipta Seal, and Ozlem Ozmen Garibay. Attentionsitedti: an interpretable graph-based model for drug-target interaction prediction using nlp sentence-level relation classification. *Briefings in Bioinformatics*, 23(4):bbac272, 2022.

Niloofar Yousefi, Mehdi Yazdani-Jahromi, Aida Tayebi, Elayaraja Kolanthai, Craig J Neal, Tanumoy Banerjee, Agnivo Gosai, Ganesh Balasubramanian, Sudipta Seal, and Ozlem Ozmen Garibay. Bindingsiteaugmenteddta: Enabling a next-generation pipeline for interpretable prediction models in drug-repurposing. *bioRxiv*, 2022.

Mikhail Yurochkin, Amanda Bower, and Yuekai Sun. Training individually fair ml models with sensitive subspace robustness. *arXiv preprint arXiv:1907.00020*, 2019.

Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rodriguez, and Krishna P Gummadi. Fairness beyond disparate treatment & disparate impact: Learning classification without disparate mistreatment. In *Proceedings of the 26th international conference on world wide web*, pp. 1171–1180, 2017a.

Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rogriguez, and Krishna P Gummadi. Fairness constraints: Mechanisms for fair classification. In *Artificial Intelligence and Statistics*, pp. 962–970. PMLR, 2017b.

Rich Zemel, Yu Wu, Kevin Swersky, Toni Pitassi, and Cynthia Dwork. Learning fair representations. In *International conference on machine learning*, pp. 325–333. PMLR, 2013.

Brian Hu Zhang, Blake Lemoine, and Margaret Mitchell. Mitigating unwanted biases with adversarial learning. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, pp. 335–340, 2018.

Lu Zhang, Yongkai Wu, and Xintao Wu. Achieving non-discrimination in data release. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 1335–1344, 2017.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. Men also like shopping: Reducing gender bias amplification using corpus-level constraints. *arXiv preprint arXiv:1707.09457*, 2017.

A Appendix

A.1 Training Algorithm

Algorithm 1 is designed to be used with deep learning frameworks and optimization algorithms. The Distraction module's weights are entirely isolated from the rest of the network. In each step, first, the fairness loss of the networks gets calculated, and the weights of the Distraction module get updated. Then we have another pass through the network, and the classification loss of the network gets calculated, and the weights of the rest of the network get updated using back-propagation.

A.2 Evaluation Metrics

Four indicators are utilized to compare the outcomes of our model to the baseline models. We assess the average precision to determine the classifiers' accuracy. This measure computes the average while combining recall and accuracy at each place. One would like a greater average precision (AP). We use accuracy in tabular and graph datasets to be consistent with the literature. Numerous criteria have been presented in the literature to assess fairness (Mehrabi et al., 2021b). Demographic parity is one of the measures that is most often utilized (DP). The difference between getting a favorable decision for various protected groups is captured by this measure ($|P(Y = 1|S = 0)P(Y = 1|S = 1)|$). For more than two groups the demographic parity can be calculated using $\Delta_{DP}(a, \hat{y}) = \max_{a_i,a_j} |P(\hat{y} = 1|a = a_i) - P(\hat{y} = 1|a = a_j))|$ (Gupta et al., 2021). A smaller DP indicates a more fair categorization and is preferred. Following (Lokhande et al., 2020) and (Ramaswamy et al., 2021), we utilize the difference in equality of opportunity (ΔEO), which is the absolute difference between the true positive rates for both gender expressions ($|TPR(S = 0) - TPR(S = 1)|$), as our final fairness metric. We want the ΔEO to be smaller.

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

Input: data (X, A, Y) A is set of protected attribute and Y is the label, Batch Size m, C Learning rate lr_1, D Learning rate lr_2, η

for number of iterations in training **do**

sample minibatch of size m samples $\{x^{(1)}, x^{(2)}, ..., x^{(m)}\}$ from data $P(x)$

updates the Distraction module by ascending the stochastic gradient with learning rate lr_1:

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

updates the classifier network by descending the stochastic gradient with learning rate lr_2:

$$\nabla_{\theta_c} \sum_{i=1}^{m} - \log p_{\theta_d, \theta_c}(y^{(i)} | x^{(i)})$$

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.

Table 3: 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
η	100	100

A.3 Implementation details

The hyperparameters used in training the models on each dataset can be found in the tables 3, 4 and 5. The training was carried out on a computer equipped with an NVIDIA GeForce RTX 3090.

Table 4: 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	500	1000	1000
Batch size	1	1	1
Dropout	0.5	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-5	1e-5	1e-5
η	100	100	1000

Table 5: 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	15	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-5	1e-5	1e-5
η	100	100	100

A.4 Explainability

In this part we used the Distraction module which is inspired by Vaswani et al. (2017) work introducing transformers.
This method consists of two separate networks. First network is pre-trained for predicting protected attribute of the datasets with respect to other attributes which excludes the main target attribute. The second network includes two sets of weights the first set of weights are related to the classifier and the second set is related to distraction module. the first set of weights are trained in order to maximize the accuracy of the classifier and the second set is trained to minimize the demographic parity. The architecture of the model is depicted in Figure 13.

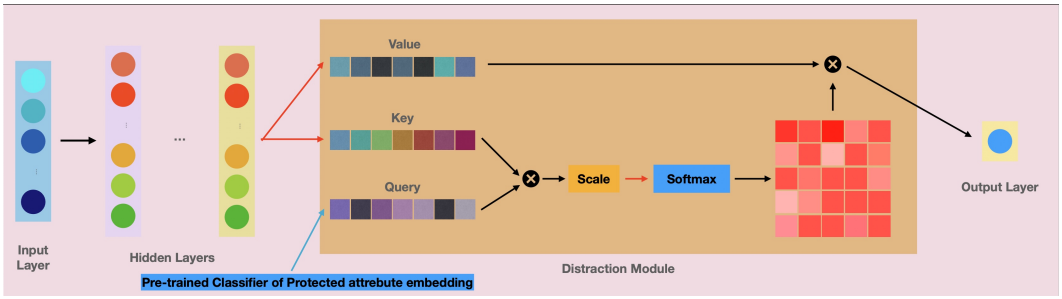


Figure 5: Depiction of Distraction Module in MLP model: the First network is the network for classifying protected attributes to get the most related embedding from proxy features this pre-trained network send the computed embedding to the Distraction module (blue arrow) as the query and the main network is providing the key and the value. the red arrows in this figure are the weights controlled by the Distraction Module and are optimized with 3 objective function. the black arrows are weights which are controlled by the main classifier. the first network is pre-trained and the weights of this network does not change in the main training process.

A.5 Ablation Study

Due to the theoretical analysis of the proposed model, the fairness layer can be any differentiable function with controllable parameters θ_d. In this subsection, we conduct an ablation study to analyze the effect of different functions in the network. For the Distraction module, we used one linear layer, two linear layers, and three linear layers. The results can be seen in table 6. For the CelebA dataset we experimented with linear, ResBlock and CNN leyers and mean of the each category of CelebA attributes can be found table in 7.

A.6 CelebA Layer Visualization

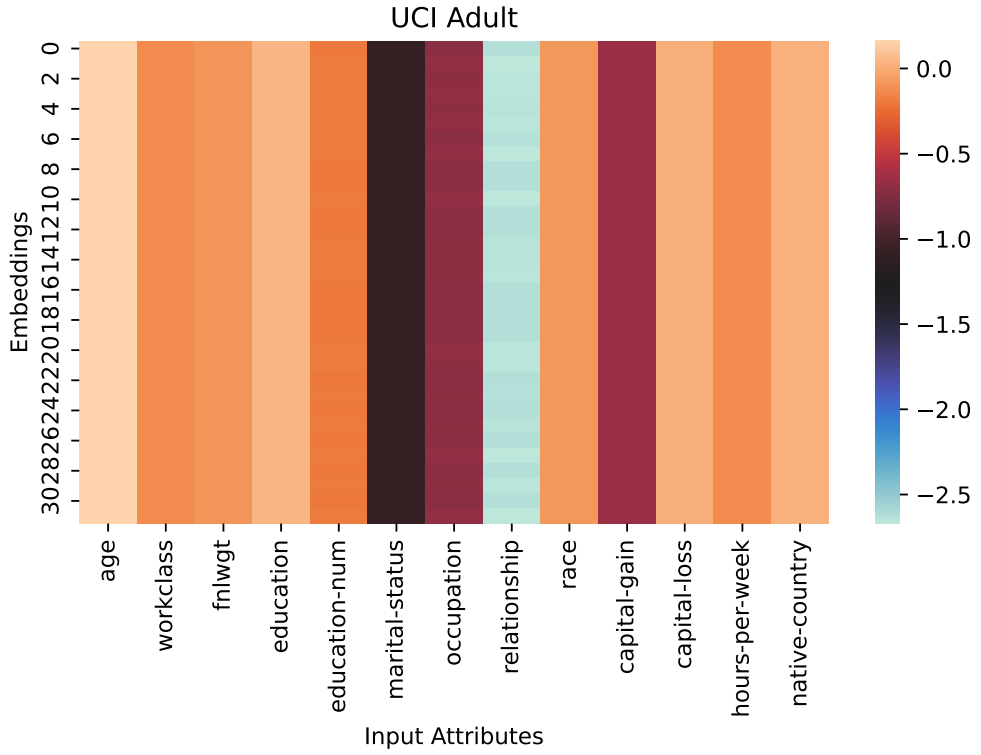


Figure 6: Heat map of effect for each of attributes towards the final classification results in UCI Adult dataset.The color light orange in this heat-map indicates the attributes that the distraction module decides to gave near zero coefficient to that attribute in classifying the data which means they are the proxy attributes.

Table 6: Area over the curve of statistical demographic parity and accuracy for model ablation

METHOD	UCI ADULT	HERITAGE HEALTH
CROSS-ATTENTION	0.411	0.503
ONE LINEAR LAYER	**0.411**	0.492
TWO LINEAR LAYERS	0.404	0.513
THREE LINEAR LAYERS	0.349	**0.531**

Table 7: Accumulative comparison between different Distraction layers

CNNBlock	AP			ΔDP			ΔEO		
	Incons.	**G-dep**	**G-indep**	**Incons.**	**G-dep**	**G-indep**	**Incons.**	**G-dep**	**G-indep**
One Linear Layer	0.646	0.755	0.841	0.072	0.115	0.085	0.084	0.069	0.089
CNN Res Block	0.568	0.699	0.768	0.04	0.035	0.026	0.126	0.067	0.062
CNN Layer	0.617	0.731	0.822	0.058	0.092	0.069	0.0.99	0.067	0.073

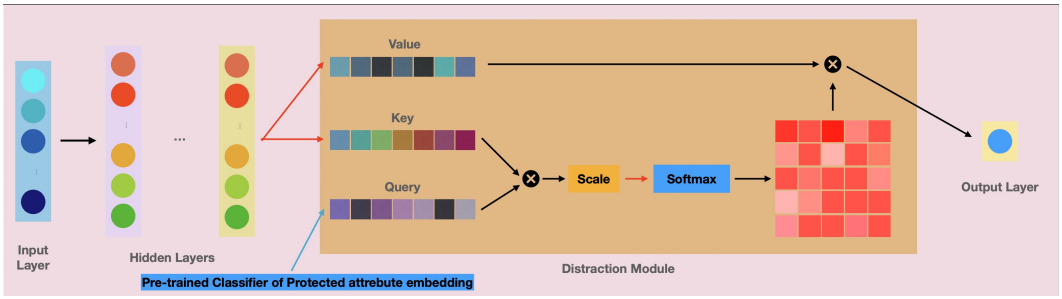


Figure 7: Depiction of Distraction Module in MLP model: the First network is the network for classifying protected attributes to get the most related embedding from proxy features this pre-trained network send the computed embedding to the Distraction module (blue arrow) as the query and the main network is providing the key and the value. the red arrows in this figure are the weights controlled by the Distraction Module and are optimized with 3 objective function. the black arrows are weights which are controlled by the main classifier. the first network is pre-trained and the weights of this network does not change in the main training process.

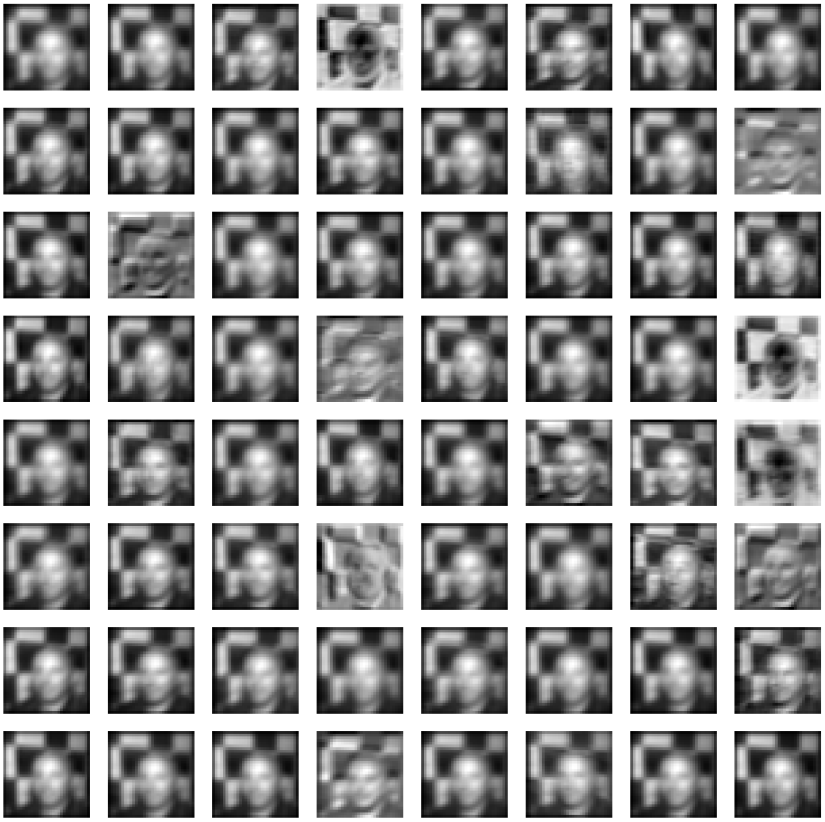


Figure 8: Visualization of the filtered image without the Distraction module

Figure 9: Visualization of the filtered image with the Distraction module as a CNN Block.

Figure 10: Visualization of the filtered image with the Distraction module as a Linear layer.

Figure 11: Visualization of the filters with the Distraction module as a CNN Block.

Figure 12: Visualization of the filters with the Distraction module as a linear layer

Figure 13: Visualization of the filters without the Distraction module