BIAS MITIGATION FRAMEWORK FOR INTERSECTIONAL SUBGROUPS IN NEURAL NETWORKS

Anonymous authors
Paper under double-blind review

ABSTRACT

We propose a fairness-aware learning framework that mitigates intersectional subgroup bias associated with protected attributes. Prior research has primarily focused on mitigating one kind of bias by incorporating complex fairness-driven constraints into optimization objectives or designing additional layers that focus on specific protected attributes. We introduce a simple and generic bias mitigation framework that prevents models from learning relationships between protected attributes and output variable by reducing mutual information. We demonstrate that our approach is effective in reducing bias with little or no drop in accuracy. We also show that our approach mitigates intersectional bias even when other attributes in the dataset are correlated with protected attributes. Finally, we validate our approach by studying feature interactions between protected and non-protected attributes. We demonstrate that these interactions are significantly reduced when applying our bias mitigation.

1 INTRODUCTION

The unprecedented adoption of Machine Learning (ML) in critical sectors such as finance, healthcare and education has made fairness-related bias detection and mitigation a crucial part of ML systems. It is important that ML models do not discriminate against individuals based on protected attributes such as race, gender or skin color when making predictions. Fairness research has typically focused on detecting and mitigating bias for a single protected attribute. While this alleviates the bias with respect to that specific attribute, the fairness gap for *intersectional subgroups* such as `Female` and `Black` might still be high. This phenomenon is called fairness gerrymandering (Ghosh et al., 2021; Kearns et al., 2018; Buolamwini & Gebru, 2018). Several works (Kearns et al., 2018; Yang et al., 2020; Kang et al., 2021; Morina et al., 2019) proposed techniques to address fairness gerrymandering, which require predetermined fairness violation metrics. Nevertheless, a recent survey (Du et al., 2021) highlights the scarcity of research in detecting and mitigating intersectional bias. One of the key challenges of this research is the limited access to protected attributes and their *unknown correlations* with other attributes in the dataset. These limitations are common due to privacy or data restrictions. Another key challenge of fairness research is both reducing the bias across groups and *intersectional subgroups* without a significant drop in accuracy.

In this paper we propose a generic and simple fairness-aware learning framework that addresses the intersectional subgroup bias without requiring specific fairness metrics to be predetermined. It learns latent representations without relying on protected attributes and their interactions with other attributes in the dataset. Previous work (Cho et al., 2020; Song et al., 2018) shows that minimizing mutual information between a protected attribute and the output variable plays a significant role in reducing the bias associated with that protected attribute. We propose a generic framework that reduces the mutual information not only between a single protected attribute and the output variable but also between any subset of protected attributes and the output variable. Experimentally, we show that this approach debiases the model significantly with little or no drop in the accuracy. Furthermore, our framework mitigates the fairness gap both for individual protected attributes as well as for *intersectional subgroups* defined by multiple protected attributes, such as $\texttt{Female} \cap \texttt{Black}$ defined by `Gender` and `Race`. With the help of well-known metrics, Equalized Odds (Hardt et al., 2016) and Demographic Parity (Dwork et al., 2012; Kusner et al., 2017), we show that the equality and parity gaps Beutel et al. (2017) can be reduced even when the dataset is unbalanced with respect to different subgroups of protected attributes. We elaborate more on those metrics in section 5.2.

Table 1: Accuracy and True Positive Rate (TPR) for two subpopulations, computed for the original model, our fairness-aware model and two baseline models.

	Accuracy	TPR	
		Male ∩ White	Female ∩ Black
Original model	0.87	0.96	0.77
Removed protected attributes	0.86	0.96	0.83
Upsampled minority groups	0.86	0.88	0.82
Our fairness-aware framework	0.86	0.94	0.90

Furthermore, we demonstrate the effectiveness of our approach when fairness bias is leaked through other features in the dataset.

Motivating example: Table 1 illustrates accuracy and True Positive Rate (TPR) gaps between two subpopulations, Male ∩ White and Female ∩ Black, for the Law School Admissions Council (LSAC) dataset Wightman & Council (1998). TPR is an important metric in this specific example since it measures the advantaged outcome. We compare our fairness-aware framework with three baseline models: 1) the original model without mitigation, 2) a baseline model which is trained on the dataset after removing protected attributes and 3) a baseline model which is trained using the dataset where we upsample minority subgroups and balance sample distribution across subgroups. We observe a significant TPR gap between two populations in the original model. Two of the baseline models reduce the fairness gap, however, they are less effective compared to our approach. Our approach learns latent representations that do not rely on protected attributes and their relationship with the output variable, and thus reduces the fairness gap significantly.

The main contributions of this paper are as follows:

- We introduce a novel bias mitigation framework that aims to reduce the mutual information between intersectional subgroups and the output variable.
- We show empirically that our approach reduces the equality and demographic parity gaps significantly and outperforms state-of-the-art approaches.
- We study the sensitivity of our debiased model to the presence of protected attributes and the effectiveness of our approach when non-protected attributes are correlated with protected ones.
- We demonstrate that feature interactions between protected and non-protected attributes reduce significantly when the models are trained using our bias mitigation framework.

2 Related Work

Fairness literature offers numerous definitions of fairness (Narayanan, 2018), its measurement and mitigation. We base our fairness definition and measurement on the work of (Hardt et al., 2016) and on three well-known metrics: Demographic Parity (Feldman et al., 2015; Dwork et al., 2012; Kusner et al., 2017), Equalized Odds, and Equality of Opportunity (Hardt et al., 2016). Demographic Parity compares the average prediction score across different subgroups. Equality of Opportunity, in addition to that, takes the label distribution into account and measures the TPR gap among different groups. Equalized Odds (Hardt et al., 2016) measures both the TPR and False Positive Rate (FPR) gaps among different groups. Specific metrics have been developed for intersectional subgroups such as the min-max ratio (Ghosh et al., 2021) and differential fairness metric (Foulds et al., 2020). In this paper we focus on measuring the commonly used Demographic Parity and Equalized Odds for intersectional subgroups, which facilitate comparisons with previous work.

Bias mitigation techniques reduce the disparities among the groups and intersectional subgroups measured by the aforementioned metrics. Three types of mitigation techniques have been proposed to combat fairness bias (Du et al., 2021):

Postprocessing techniques aim to reduce fairness bias during model inference. Those approaches enforce model predictions to follow the same distribution observed during training (Zhao et al., 2017), transforming model predictions to follow a specific fairness measure such as Equality of Opportunity (Hardt et al., 2016) or ϵ-differential fairness metric (Morina et al., 2019). These techniques, however, require access to protected attributes during inference, which is not always available due to data scarcity or privacy reasons.

Dataset preprocessing techniques such as balancing the distribution of data labels, downsampling and sample re-weighing (Kamiran & Calders, 2011) alleviate modelling bias to a certain extent. Nonetheless, these approaches do not always work when the number of samples in a subpopulation is small. However, (Wang et al., 2019) show that data preprocessing and balancing datasets often have limited effect, compared with training inherently unbiased models. Apart from data balancing, one can also delete protected attributes from the training set or mask them with neutral terms as part of preprocessing. However, this is not sufficient, since protected attributes are often correlated with other attributes in the data.

Train-time techniques aim to combat potential fairness bias during model training. This can be accomplished using constraints based on adversarial loss (Beutel et al., 2017; Zhang et al., 2018; Wang et al., 2019), feature importance (Liu & Avci, 2019; Du et al., 2019; Ross et al., 2017), fairness measurement (Agarwal et al., 2018), decision boundary (Zafar et al., 2019) or statistical dependence (Kamishima et al., 2012). Adversarial loss requires defining additional heads or constraints for a specific protected attribute. It maximizes the primary objective of a specific task while minimizing the model's ability to predict specific protected attributes. Constraints based on feature importance, on the other hand, heavily rely on the feature contribution score (Liu & Avci, 2019; Du et al., 2019), which is not always reliable (Hooker et al., 2019) and requires human annotated labels of important features. Yang et al. (2020) propose Bayes-optimal classification framework for interesectional group fairness which is also tied to fairness metric constraints. Approaches that involve a certain fairness metric in the training objective require an upfront selection of this metric, which is not always a straightforward choice. Kang et al. (2021) propose a framework based on mutual information minimization for intersectional subgroups. This framework, however, requires two estimators and two additional predictors which makes the mitigation process complex.

In contrast, we propose a simple and generic training-time fairness-aware framework that doesn't rely on specific fairness metrics, or architectural modifications such as adversarial heads. It accounts for intersectional fairness of any subsets of input features and is straightforward to implement.

3 PRELIMINARIES

In this section we formalize the problem, introduce preliminary notations and concepts that are used later in the paper. The goal is to learn a fair Neural Network (NN) model that mitigates the fairness gap for intersectional subgroups formed by multiple protected attributes. We seek to reach this goal with minimal impact on accuracy. During training, we assume that there is a small number of examples for which protected attributes (and their correlations with non-protected attributes) are available. However, during inference, no information about the protected attributes is necessary.

3.1 NOTATION

We consider a typical ML model, $f : R^M \rightarrow R^C$, that is trained on a dataset $D = \{(x_1, y_1), ..., (x_N, y_N)\}$, where each sample x_i consists of a set of M features $x_i = \{x_i^1, x_i^2 ..., x_i^M\}$ where $x_i^j \in R^d$ represents the j^{th} feature in the i^{th} sample. $y_i \in [1, ..., C]$ is the label corresponding to sample x_i. Let $x'^j \in R^d$ denote a shared baseline across all samples for input feature j. It indicates the absence of signal or feature value in the input. Traditionally, the zero value is used to indicate the absence of signal but for certain features zero value might represent a meaning. For example, 0 and 1 might indicate male and female for the gender feature in some datasets.

Protected Attribute Notation. Let $A_k \subseteq \{1, 2, ..., M\}$ be a subset of features that are known to be protected. For example, A_k may correspond to $\{race\}$ or $\{gender, race\}$. Let also $\mathcal{A}$ denote a set of the subsets of A_k. For example, $\mathcal{A}$ may correspond to $\mathcal{A} = \{\{gender\}, \{race\}, \{gender, race\}\}$.

Finally, let $\mathcal{S}(x_i, x', A_k)$ denote a substitution function that replaces the features that are not in A_k with values from baseline x':

$$\mathcal{S}(x_i, x', A_k) = \begin{cases} x_i^j, & j \in A_k \\ x'^j, & \text{otherwise} \end{cases} \quad \forall j \in [M] \tag{1}$$

We denote by x^{A_k} the subset of features in x corresponding to protected features A_k. Analogously, $x^{\setminus A_k}$ denotes a subset of features that excludes protected attributes A_k.

3.2 Information Theory

We briefly revisit information theory in order to analyse the relationship between a subset of protected attributes A_k and a dependant output variable y. This analysis explains the reasoning behind our bias mitigation framework. The entropy $H(X)$ measures the average uncertainty (Cover & Thomas, 2006) of a random variable X. It takes its highest value for a uniformly distributed random variable X and is equal to zero when X is constant. Mutual Information (MI) uses entropy to measure shared information between two random variables. In our case, the mutual information between an input feature x^j and the output variable y can be measured as $MI(x^j; y) = H(x^j) - H(x^j|y)$.

During bias mitigation, we aim to reduce the MI between the protected attributes x^{A_k} and y. Reducing $MI(x^{A_k}; y)$ implies increasing the uncertainty $H(x^{A_k}|y)$. We increase that uncertainty by associating protected attributes with a uniformly distributed random output variable $Unif(1, C)$, given that

$$H\big(x^{A_k}|Unif(1, C)\big) \geq H\big(x^{A_k}|y\big) \tag{2}$$

This way the model learns to de-correlate the protected attributes and the model output. Since computing MI requires discretization of the continuous features and it doesn't directly depend on model parameters, we propose a proxy measure of MI that can be incorporated into the optimization objective. The proxy measure aims to reduce the distance between $\mathcal{S}(x_i, x', A_k)$ and $Unif(1, C)$ which can be represented as a regularization term in the optimization objective.

Table 2 shows how the mutual information between protected attributes `Gender` and `Race` reduces as we augment the data with samples that associate intersectional subgroups of protected attributes with random guesses. In the next section we describe how we incorporate this regularization term into the model's optimization objective.

Table 2: MI between protected attributes `Gender` and `Race` and output variable `Passed Bar` before and after MI constraint-based data augmentation for the LSAC and Adult datasets.

	Passed Bar (Before Mitigation)		Passed Bar (After Mitigation)	
	LSAC	Adult	LSAC	Adult
Gender	0.00101	0.043532	0.00029	0.011177
Race	0.02265	0.010469	0.004876	0.0010130

4 Fairness-Aware Learning Algorithm

The learning algorithm we propose is similar in spirit to the ones which incorporate predetermined constraints into optimization objective as regularization terms. It is more generic in its nature and does not use EO, DP or adversarial heads as a proxy for the minimization of the mutual information (Cho et al., 2020; Song et al., 2018) between the protected attributes and the output variable. We give a formal definition of the objective function starting from the definition of the the proxy constraint for mutual information.

Definition 4.1 *Given a subset of protected attributes $A_k \in \mathcal{A}$, an input example x_i, a uniformly random chosen label y_{rand} and a baseline x', the proxy constraint for the mutual information is defined as follow:*

$$L^{\mathcal{A}}(x_i, x', \mathcal{A}, y_{rand}) = -\sum_{A_k \in \mathcal{A}} \sum_{c \in C} \mathbb{1}(y_{rand} = c) \cdot \log(f_c(\mathcal{S}(x_i, x', A_k))) \tag{3}$$

where $c \in C$ are possible prediction classes and $y_{rand} = Unif(1, C)$ is a label drawn uniformly random from C. The first summation over the subsets of protected attributes A_k helps us to address bias mitigation for multiple subsets of protected attributes. The joint objective of a multiclass classification problem is the following.

$$L_{combined} = \sum_{(x_i, y_i) \in D} L(x_i, y_i) + \alpha \cdot \sum_{(x_i, x') \in D'} L^{A}(x_i, x', \mathcal{A}, Unif(1, C)) \tag{4}$$

In our setup we use a classification loss but other loss definitions can be used instead. $L(x_i, y_i)$ represents the loss for the original model. The hyperparameter α is used to balance the amount of regularization that we incorporate into the loss. $D' = \{(x_1, x'), ..., (x_N, x')\}$ represents the dataset with a set of baseline values $x' = \{x'^j\}_{j=1}^{j=M}$ for each feature j. $Unif(1, C)$ chooses a label uniformly at random from $[1, ..., C]$. Algorithm 1 illustrates an example of how our proposed loss can be computed during training using Gradient Descent (GD) algorithm. Our method is not limited to GD and can be trained with other optimizations algorithms as well.

Algorithm 1: Fairness-Aware Learning Algorithm

Input: Training Datasets $D = \{(x_i, y_i)\}_{i=1}^{i=N}$, Validation dataset $D^{\text{valid}} = \{(x_i, y_i)\}_{i=1}^{i=F}$, Baseline $x' = \{x'^j\}_{j=1}^{j=M}$, A set of subsets of protected attributes $\mathcal{A}$, hyperparameter α, learning rate η, max_epochs.

Output: W_{best} for the best accuracy of the model f on D^{valid} while minimizing reliance on the feature sets in $\mathcal{A}$.

Initialize the model parameters W_0, set epoch=0;
while $epoch < max\_epochs$ **do**
  $L_{\text{initial}} = \frac{1}{N} \cdot \sum_{i=1}^{i=N} \sum_{c=1}^{C} -\mathbb{1}(y_i = c) \cdot log(f_c(x_i))$;
  $y_{\text{rand}} = uniform\_rand(C)$; `// Uniformly at random chooses class in` $[1, C]$ `for each example`
  $L^{\mathcal{A}} = \frac{1}{N} \sum_{i=1}^{i=N} \sum_{k=1}^{|\mathcal{A}|} \sum_{c=1}^{C} -\mathbb{1}(y_i = c) \cdot log(f_c(\mathcal{S}(x_i, x', A_k)))$;
  $L_{\text{combined}} = L_{\text{initial}} + \alpha \cdot L^{\mathcal{A}}$
  epoch = epoch + 1 ;
  $W_{\text{epoch}} = Optimizer(L_{\text{combined}}, \eta)$; `// Optimizer can be Adam, for example`
  Update W_{best} based on the highest accuracy measures on D^{valid} so far.

5 EXPERIMENTS

In this section we present experimental results of our fairness-aware learning framework and a number of state-of-the-art baseline approaches. We also describe experimental setup, evaluation metrics and discuss empirical results.

5.1 EXPERIMENTAL SETUP

The UCI Adult dataset (Dua & Graff, 2017) and the Law School Admissions Council (LSAC) dataset (Wightman & Council, 1998) are two highly unbalanced datasets used in our experiments. Appendix A.1 provides details about those two datasets and an additional COMPAS 7 dataset used in the appendix. For these two datasets we built a 2-layer RELU-BatchNorm-Linear NN models similar to the one demonstrated in the literature (Beutel et al., 2017). The first linear layer contains

128, the second 64 and the last output layer only a single neuron. The models are trained using approximately 100 epochs with Adam optimizer, 0.001 learning rate and binary cross entropy logit loss as the baseline model's loss, $L(x_i, y_i)$. As a bias mitigation constraint, $L^{\mathcal{A}}$, we used squared distance between the output and the provided label instead of cross entropy loss since it is a convenient approach for binary classification. We chose baseline values x' carefully to indicate missingness of corresponding attributes in the dataset.

We compared our method with the well known GerryFair (Kearns et al., 2018) and a mutual information reduction-based approach (Cho et al., 2020) adopted for intersectional fairness. GerryFair performs a zero-sum optimization between a fairness auditor and a classifier subject to the auditor's constraints. On the other hand, (Cho et al., 2020; Song et al., 2018; Louppe et al., 2017) show that mutual information reduction-based approaches can be formulated as generative adversarial optimization problems. Here the classifier plays the role of the generator and the discriminator aims to reduce the mutual information between the protected attributes and the output of the classifier. Inspired by (Song et al., 2018; Louppe et al., 2017) we implemented an adversarial network with two 2-layer RELU-BatchNorm-Linear NNs. One of those networks serves as a generator and the other one as a discriminator. The last layer of discriminator is equal to six; one representing binary-valued gender and other 5 representing different values of race.

5.2 Evaluation metrics

As fairness measurement metrics we adopted Equalized Odds (EO) Hardt et al. (2016); Beutel et al. (2017) and Demographic Parity (DP) Dwork et al. (2012); Kusner et al. (2017) metrics to measure model's intersectional subgroup biases. Inspired by Beutel et al. (2017) and Ghosh et al. (2021) we measure EO as the differences between minimum and maximum TPR and FPR scores across all subgroups formed by a given subset of protected attributes. Let G denote the set of different combinatorial options formed by the subgroups of protected attributes in the set $A_k \in \mathcal{A}$. For example $G = \{(Male, White), (Male, Black), (Male, Asian), (Female, White), (Female, Black), (Female, Asian), ...\}$. The values in G are then denoted by $G_i \in G$. EO based on TPR and FPR are then defined as $EO_G^{TPR} = |max(TPR(G_i)) - min(TPR(G_j))|$ and $EO_G^{FPR} = |max(FPR(G_i)) - min(FPR(G_j))|, G_i \in G, G_j \in G$ accordingly. Similarly, demographic parity is measured as $DP_G = |max(DP(G_i)) - min(DP(G_j))|, G_i \in G, G_j \in G$.

5.3 Experimental Results

In order to better understand the accuracy vs fairness gap relationship, we run multiple experiments by varying the weight of the fairness component from zero to a large number both for our approach and two baseline approaches. We compare our approach against GerryFair (Kearns et al., 2018) and a mutual information-based approach (Cho et al., 2020) in terms of Accuracy vs TPR, FPR and DP gaps. As described in section 5.1 mutual information based fairness mitigation is performed based on adversarial training. Our experimental results depicted on figure 1 reveal that our method results in higher accuracy when reducing TPR gap from 0.5 to 0.2 for LSAC and from 0.46 to 0.23 for Adult datasets. All three methods exhibit similar accuracy when the TPR gap is further reduced from approximately 0.2 to 0.0. We examine similar patterns for Accuracy vs FPR and DP gap measurements as well. We observe that the adversarially trained approach slightly underperforms two other approaches in terms of accuracy. We hypothesize that generator-discriminator based approaches are more effective for binary protected attributes as also shown in (Louppe et al., 2017). When dealing with multiple non-binary protected attributes, generator-discriminator based networks become less effective and might require further fine tuning. We conclude that our method can be beneficial especially when the goal is to reduce fairness gaps substantially without hurting the accuracy too much. Appendix A.2 provides additional results for COMPAS (Barenstein, 2019) dataset.

Furthermore, we compare our method with the original and two additional baseline models. One of the baselines represents a model trained without protected attributes. In this case we removed `Gender` and `Race` attributes from the dataset. The second baseline model is trained on a dataset where examples for all underrepresented intersectional subgroups of `Gender` and `Race` are upsampled. Figure 2 summarizes Accuracy, TPR, FPR and DP gaps across all baselines and our approach for LSAC and Adult datasets. We observe that our method is comparable or better in reducing the fairness gap for almost the same accuracy trade-off.

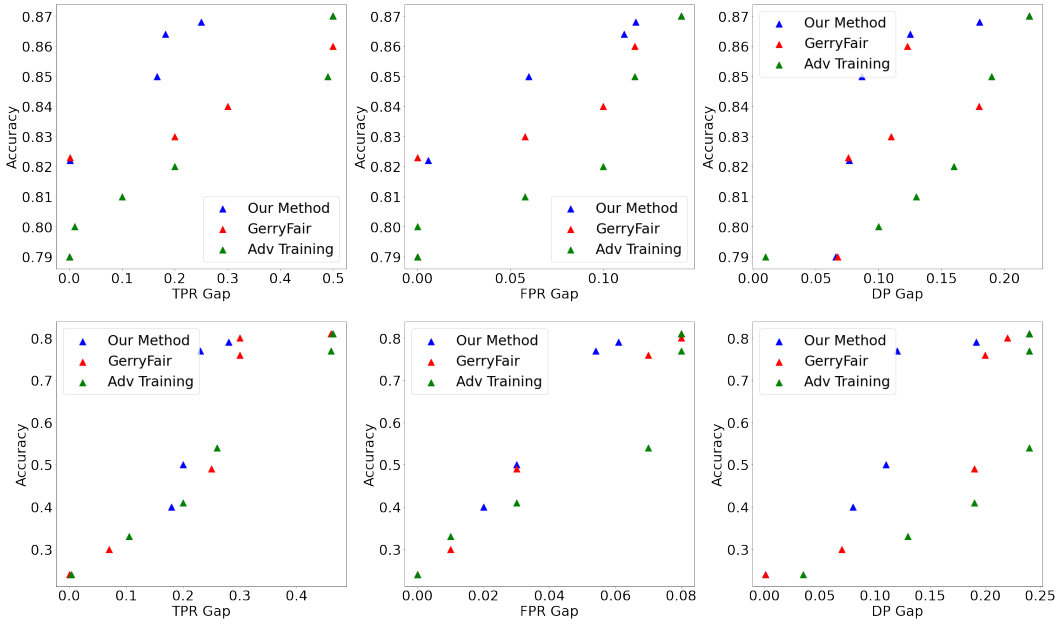

Figure 1: Our method, GerryFair and adversarially trained approaches compared on accuracy vs fairness metrics (TPR, FPR and DP gaps) for LSAC (top) and Adult (bottom) datasets. Best viewed in color.

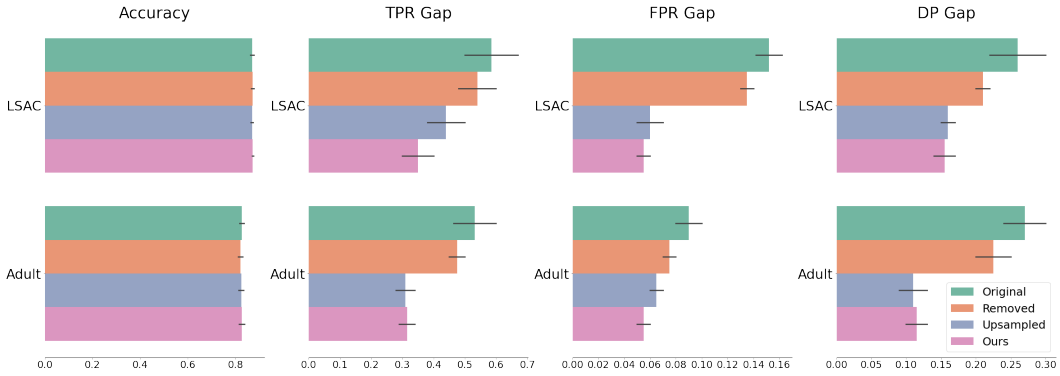

Figure 2: Our method ($\alpha = 0.5$) vs original model, removed `Gender` and `Race` attributes and upsampled for underrepresented subgroups. Best viewed in color.

In order to validate the effectiveness of our approach we perform two additional studies. We aim to understand how inference time masking of protected attributes and the presence of attributes strongly correlated with protected attributes, change the accuracy, TPR, FPR and DP gaps.

Masking protected attributes: We mask protected attributes in the test dataset and compare model accuracy and fairness metrics before and after masking. This validates the hypothesis that fairness gap changes will remain changed before and after protected attributes are masked in the test dataset, if our method is effective in mitigating bias. Table 3 showcases the high sensitivity of the original model's accuracy, TPR, FPR and DP gaps when protected attributes are masked. In contrast to the original model, our model is almost insensitive to the masking of protected attributes showing its robustness to the presence of protected attributes. Appendix A.2 demonstrates the results of the same experiment for Adult dataset.

Studying the effects of strongly correlated features with protected attributes: In order to understand the effectiveness of our approach in the presence of features strongly correlated with pro-

Table 3: Test Accuracy, TPR, FPR and DP Passed Bar Gaps for the Original and Our Models ($\alpha = 0.5$) with and without masking of protected attributes applied on LSAC dataset.

	Accuracy	TPR	FPR	DP passed
Original model	0.87	0.55	0.14	0.23
Original model w/masked $A^{(Gender,Race)}$	0.85	0.53	0.05	0.15
Our model	0.86	0.28	0.06	0.11
Our model w/masked $A^{(Gender,Race)}$	0.86	0.28	0.06	0.10

Table 4: Accuracy and TPR for two subpopulations, computed for the original model, a baseline model trained without `Gender` and `Race`, and our approach. The dataset contains an additional attribute `Race1` for which mitigation is intentionally not performed.

	Accuracy	TPR	
		`Male`$\cap$`White`	`Female`$\cap$`Black`
Original model	0.87	0.99	0.76
Removed protected attributes	0.86	0.96	0.77
Our model ($\alpha = 0.5$)	0.86	0.99	0.90

tected attributes we use two race related features `Race1` and `Race`. We apply bias mitigation only to fields `Gender` and `Race`. No mitigation for field `Race1` is carried out. Table 4 shows that our approach is still effective in mitigating the bias in underrepresented groups such as `Female`$\cap$`Black`. It is not as effective as in the absence of `Race1` field 1, however, it is more effective than removing `Gender` and `Race` from the dataset.

Feature interaction effects: In addition to accuracy and fairness evaluation metrics, we also analyse pairwise feature interaction effects of protected attributes for the biased and unbiased versions of the model. We compute pairwise feature interactions based on the formulations in Tsang et al. (2020). We validate the hypothesis that pairwise feature interaction of protected attributes with non-protected ones drops significantly in the unbiased model. Furthermore, our experiments reveal that the decline of feature interaction scores for protected attributes leads to the emergence of stronger interaction patterns between other attributes. Figure 3 visualizes aggregated pairwise feature interaction heatmaps for the original, biased model, at the top and unbiased model, based on our approach, at the bottom of the diagram. The results suggest that feature interaction effects for `Male` and `Female` look very similar. We also observe that the protected attribute `gender` has a relatively strong interaction pattern with the `age` attribute and `race` with the `fulltime` in the biased model. The unbiased model, however, exhibits no distinct and strong feature interaction patterns for `gender` and `race` with `age` and `fulltime` respectively. On the other hand, we discern stronger emerging interaction patterns between `parttime` and `zgpa`, `fulltime` and `zgpa`. This helps us better understand how feature interaction patterns are impacted when the models are trained with bias mitigation constraints. These findings can serve as sanity checks and facilitate better understanding of bias mitigation techniques.

Our experimental studies provide the following empirical evidence:

- Our approach achieves higher accuracy when reducing the same amount of TPR, FPR and DP gaps compared to the GerryFair Kearns et al. (2018) and mutual information Cho et al. (2020) based baseline approaches.
- In numerous instances, our approach is more effective, when reducing TPR, FPR and DP gaps, compared to removing protected attributes from the dataset or upsampling examples for underrepresented subgroups.
- Masking protected attributes neither changes the accuracy nor the measured gap for our approach.
- Our approach is still effective even when protected attributes are highly correlated with non-protected ones.

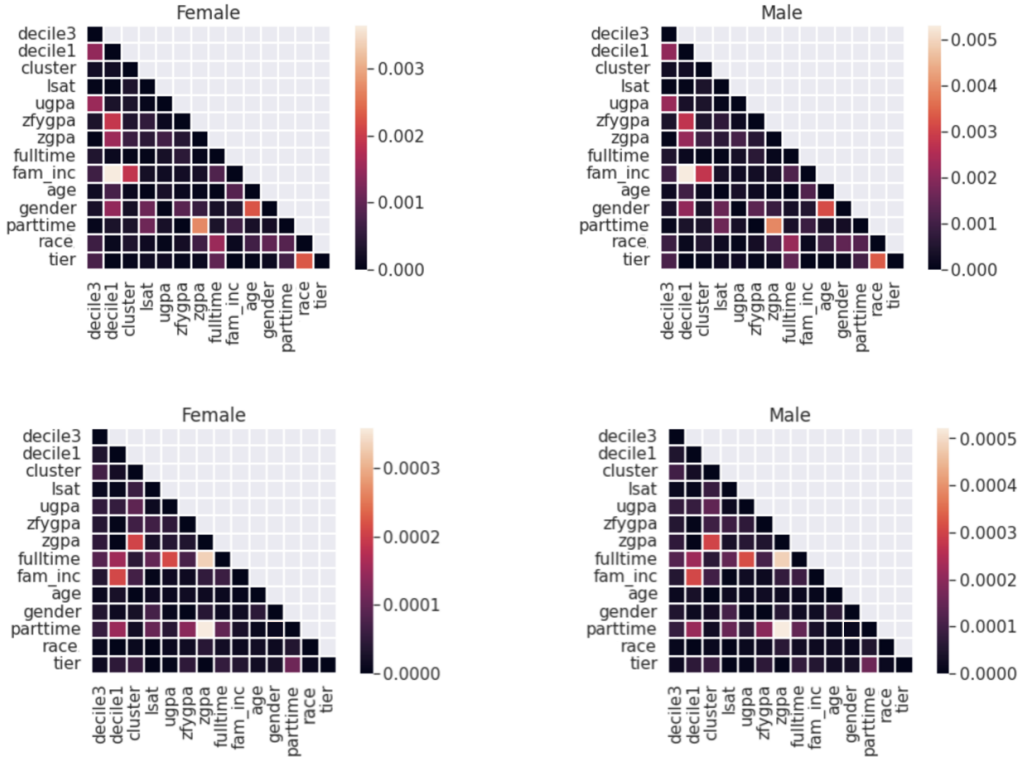

Figure 3: Aggregated pairwise feature interaction heatmaps before (top) and after (bottom) bias mitigation for both `Male` and `Female` based on LSAC test dataset. Best viewed in color.

- A model trained with our approach exhibits lower feature interaction effects between protected and non-protected attributes in the dataset.

6 Conclusion and Future Work

In this paper, we propose a framework to mitigate modeling bias in intersectional subgroups independent of the type of the protected attributes. The framework incorporates a generic proxy constraint into the optimization objective which increases the uncertainty between protected attributes and the output variable, thus reduces mutual information. We study accuracy vs TPR, FPR and DP gap trade-offs both with respect to state of the art approaches as well as data pre-processing techniques such as upsampling and removing protected attributes. We show, empirically, that our approach surpasses GerryFair and adversarially trained approaches when reducing the fairness gaps up to a certain extent. In addition to that we also demonstrate that our approach outperforms data pre-processing based baseline approaches. Furthermore, we show empirically that our approach is still effective when other features are correlated with protected attributes. Lastly, our experiments reveal that bias removal reduces the interaction effects between protected attributes and other attributes in the dataset.

In the future, we plan to analyse the effects of different types of bias mitigation constraints on a specific ML task of interest. This type of research can help better understand what kind of constraints might work better for a specific problem space and why. In addition to that it is also valuable to investigate techniques for estimating the optimal hyperparameter value for the weight of bias mitigation term, instead of a grid search approach.

REFERENCES

Alekh Agarwal, Alina Beygelzimer, Miroslav Dudík, John Langford, and Hanna Wallach. A reductions approach to fair classification. In *International Conference on Machine Learning (ICML)*, 2018.

Matias Barenstein. Propublica's compas data revisited, 2019. URL https://arxiv.org/abs/1906.04711.

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.

Joy Buolamwini and Timnit Gebru. Gender shades: Intersectional accuracy disparities in commercial gender classification. In Sorelle A. Friedler and Christo Wilson (eds.), *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*, volume 81 of *Proceedings of Machine Learning Research*, pp. 77–91. PMLR, 23–24 Feb 2018. URL https://proceedings.mlr.press/v81/buolamwini18a.html.

Jaewoong Cho, Gyeongjo Hwang, and Changho Suh. A fair classifier using mutual information. In *2020 IEEE International Symposium on Information Theory (ISIT)*, pp. 2521–2526, 2020. doi: 10.1109/ISIT44484.2020.9174293.

Thomas M. Cover and Joy A. Thomas. *Elements of Information Theory (Wiley Series in Telecommunications and Signal Processing)*. Wiley-Interscience, USA, 2006. ISBN 0471241954.

Mengnan Du, Ninghao Liu, Fan Yang, and Xia Hu. Learning credible deep neural networks with rationale regularization. *IEEE International Conference on Data Mining (ICDM)*, pp. 150–159, 2019.

Mengnan Du, Fan Yang, Na Zou, and Xia Hu. Fairness in deep learning: A computational perspective. *IEEE Intelligent Systems*, 36(4):25–34, 2021. doi: 10.1109/MIS.2020.3000681.

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 *Innovations in Theoretical Computer Science Conference*, ITCS '12, pp. 214–226, New York, NY, USA, 2012. Association for Computing Machinery. ISBN 9781450311151. doi: 10.1145/2090236.2090255. URL https://doi.org/10.1145/2090236.2090255.

Michael Feldman, Sorelle A. Friedler, John Moeller, Carlos Scheidegger, and Suresh Venkatasubramanian. Certifying and removing disparate impact. In *ACM SIGKDD international conference on knowledge discovery and data mining*, pp. 259–268. Association for Computing Machinery, 2015. ISBN 9781450336642. doi: 10.1145/2783258.2783311. URL https://doi.org/10.1145/2783258.2783311.

James R. Foulds, Rashidul Islam, Kamrun Naher Keya, and Shimei Pan. An intersectional definition of fairness. In *2020 IEEE 36th International Conference on Data Engineering (ICDE)*, pp. 1918–1921, 2020. doi: 10.1109/ICDE48307.2020.00203.

Avijit Ghosh, Lea Genuit, and Mary Reagan. Characterizing intersectional group fairness with worst-case comparisons, 2021.

Moritz Hardt, Eric Price, and Nathan Srebro. Equality of opportunity in supervised learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2016.

Sara Hooker, D. Erhan, Pieter-Jan Kindermans, and Been Kim. A benchmark for interpretability methods in deep neural networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

Faisal Kamiran and Toon Calders. Data preprocessing techniques for classification without discrimination. *Knowledge and Information Systems*, 33:1–33, 2011.

Toshihiro Kamishima, Shotaro Akaho, Hideki Asoh, and Jun Sakuma. Fairness-aware classifier with prejudice remover regularizer. In Peter A. Flach, Tijl De Bie, and Nello Cristianini (eds.), *Machine Learning and Knowledge Discovery in Databases*, pp. 35–50, Berlin, Heidelberg, 2012. Springer Berlin Heidelberg. ISBN 978-3-642-33486-3.

Jian Kang, Tiankai Xie, Xintao Wu, Ross Maciejewski, and Hanghang Tong. Multifair: Multi-group fairness in machine learning. *ArXiv*, abs/2105.11069, 2021.

Michael Kearns, Seth Neel, Aaron Roth, and Zhiwei Steven Wu. Preventing fairness gerrymandering: Auditing and learning for subgroup fairness. In Jennifer Dy and Andreas Krause (eds.), *International Conference on Machine Learning (ICML)*, volume 80 of *Proceedings of Machine Learning Research*, pp. 2564–2572. PMLR, 10–15 Jul 2018. URL https://proceedings.mlr.press/v80/kearns18a.html.

Matt J Kusner, Joshua Loftus, Chris Russell, and Ricardo Silva. Counterfactual fairness. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (eds.), *Advances in Neural Information Processing Systems (NeurIPS)*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper/2017/file/a486cd07e4ac3d270571622f4f316ec5-Paper.pdf.

Frederick Liu and Besim Avci. Incorporating priors with feature attribution on text classification. In *Annual Meeting of the Association for Computational Linguistics*, 2019.

Gilles Louppe, Michael Kagan, and Kyle Cranmer. Learning to pivot with adversarial networks. In *NIPS*, 2017.

Giulio Morina, Viktor Mykhailovych Oliinyk, Julian Waton, Ines Marusic, and Konstantinos Georgatzis. Auditing and achieving intersectional fairness in classification problems. *ArXiv*, abs/1911.01468, 2019.

Arvind Narayanan. Translation tutorial: 21 fairness definitions and their politics, 2018.

Andrew Slavin Ross, Michael C. Hughes, and Finale Doshi-Velez. Right for the right reasons: Training differentiable models by constraining their explanations. In *International Joint Conference on Artificial Intelligence (IJCAI)*, pp. 2662–2670, 2017. doi: 10.24963/ijcai.2017/371. URL https://doi.org/10.24963/ijcai.2017/371.

Jiaming Song, , Aditya Grover, Shengjia Zhao, and Stefano Ermon. Learning controllable fair representations. *arXiv preprint arXiv:1812.04218*, 2018.

Michael Tsang, Sirisha Rambhatla, and Yan Liu. How does this interaction affect me? interpretable attribution for feature interactions. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/443dec3062d0286986e21dc0631734c9-Abstract.html.

T. Wang, J. Zhao, M. Yatskar, K. Chang, and V. Ordonez. Balanced datasets are not enough: Estimating and mitigating gender bias in deep image representations. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 5309–5318, Los Alamitos, CA, USA, nov 2019. IEEE Computer Society. doi: 10.1109/ICCV.2019.00541. URL https://doi.ieeecomputersociety.org/10.1109/ICCV.2019.00541.

L.F. Wightman and Law School Admission Council. *LSAC National Longitudinal Bar Passage Study*. LSAC research report series. Law School Admission Council, 1998. URL https://books.google.com/books?id=O9A7AQAAIAAJ.

Forest Yang, Moustapha Cisse, and Sanmi Koyejo. Fairness with overlapping groups. *Advances in Neural Information Processing Systems*, 2020-December, 2020. ISSN 1049-5258. Publisher Copyright: © 2020 Neural information processing systems foundation. All rights reserved.; 34th Conference on Neural Information Processing Systems, NeurIPS 2020 ; Conference date: 06-12-2020 Through 12-12-2020.

Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez-Rodriguez, and Krishna P. Gummadi. Fairness constraints: A flexible approach for fair classification. *Journal of Machine Learning Research*, 20(75):1–42, 2019. URL http://jmlr.org/papers/v20/18-262.html.

B. Zhang, Blake Lemoine, and Margaret Mitchell. Mitigating unwanted biases with adversarial learning. In *AAAI/ACM Conference on AI, Ethics, and Society*, 2018.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. Men also like shopping: Reducing gender bias amplification using corpus-level constraints. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 2979–2989, Copenhagen, Denmark, September 2017. Association for Computational Linguistics. doi: 10.18653/v1/D17-1323. URL https://aclanthology.org/D17-1323.

A Appendix

A.1 Dataset

In this section we describe LSAC and UCI Adult datasets in detail. Both datasets have a highly unbalanced distribution over their protected attributes such as race and gender. Approximately 84% of all samples in the LSAC dataset have `White` as race, while only 1.8% have `Other` as race. Likewise, there are more examples for `Male` than for `Female` gender. The labels have skewed distributions as well; with approximately 94% samples labelled as `Passed` while only 6% are labeled as `Not Passed`. The full breakdown of LSAC dataset is presented in Table 5. We observe similar data distribution patterns for the Adult dataset. Approximately 86% of all samples in the Adult dataset are associated with `White` race, with only 0.7% with $other$ race. There are more samples for `Male` than for `Female` and more sample for `<=50K` salary range than for `>50K`. The full breakdown of Adult dataset is presented in the table 6. Similar to LSAC and Adult dataset we observe uneven distribution of samples across `Gender` and `Race` intersectional subgroups 7. We observe that, specifically, the number of examples with recidivism label for `Male` and `Not Caucasian` subgroup surpass the number of examples with no recidivism label for the same subgroup. This is not the case for any other subgroup formed by `Gender` and `Race` attributes.

Table 5: LSAC dataset breakdown by gender, race and dataset label (Bar Pass and Bar Not Pass)

	Black	Hispanic	Asian	White	Other
Male (Not Pass / Pass)	99 / 352	64 / 443	33 / 363	311 / 9622	17 / 197
Female (Not Pass / Pass)	167 / 580	51 / 368	27 / 367	249 / 6957	20 / 140

Table 6: Adult dataset breakdown by gender, race and income category

	Black	Asian-Pac-Isl	Amer-Ind-Esk	White	Other
Male(<=50k/>50k)	1736 / 408	563 / 304	230 / 39	18268 / 8752	191 / 36
Female(<=50k/>50k)	1958 / 126	371 / 65	152 / 14	10428 / 1455	117 / 9

Table 7: COMPAS dataset breakdown by gender, race and recidivism label

	Caucasian	Not Caucasian
Male(No recid. / Did recid.)	968 / 652	1630 / 1744
Female(No recid. / Did recid.)	310 / 170	450 / 230

A.2 Additional Experiments

Similar to LSAC dataset, we perform additional experiments on the Adult dataset when protected attributes `Gender` and `Race` are masked in the test dataset. We observe that after masking those attributes, Accuracy, TPR, FPR and DP Passed gaps do not change much. This validates the hypothesis that our method learns latent representations that do not rely on protected attributes.

Table 8: Test Accuracy, TPR, FPR and DP Passed Bar gaps for the original and our models ($\alpha = 0.5$) with and without masking of protected attributes applied on Adult dataset.

	Accuracy	TPR	FPR	DP Passed
Original model	0.81	0.46	0.08	0.24
Original model w/masked $A^{(Gender,Race)}$	0.80	0.40	0.06	0.19
Our model	0.80	0.23	0.06	0.19
Our model w/masked $A^{(Gender,Race)}$	0.80	0.22	0.06	0.19

Similar to LSAC and Adult datasets, we perform additional experiments with COMPAS dataset in order to further support our findings and empirical evidence. Table 4 compares our method against GerryFair and adversarialy trained methods when mitigating the intersectional subgroup bias for `Gender` and `Race` protected attributes. We observe that our approach outperforms GerryFair and adversarially trained methods in terms of reducing Accuracy vs TPR, FPR and DP gap tradeoffs.

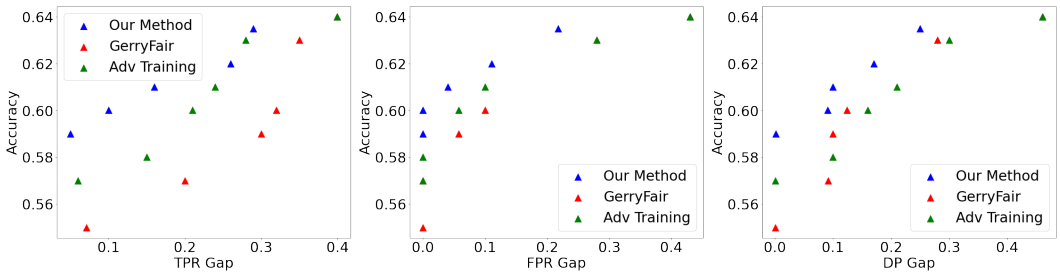

Figure 4: Our method, GerryFair and adversarially trained approaches compared on accuracy vs fairness metrics (TPR, FPR and DP gaps) for COMPAS dataset. Best viewed in color.