Synthetic Data Generation for Intersectional Fairness by Leveraging Hierarchical Group Structure

Anonymous ACL submission

Abstract

In this paper, we introduce a data augmentation approach specifically tailored to enhance intersectional fairness in classification tasks. Our method capitalizes on the hierarchical structure inherent to intersectionality, by viewing groups as intersections of their parent categories. This perspective allows us to augment data for smaller groups by learning a transformation function that combines data from these parent groups. Our empirical analysis, conducted on four diverse datasets including both text and images, reveals that classifiers trained with this data augmentation approach achieve superior intersectional fairness and are more robust to "leveling down" when compared to methods optimizing traditional group fairness metrics.

1 Introduction

The primary objective of fair machine learning is to create models that are free from discriminatory behavior towards subgroups within the population. These subgroups are often defined based on sensitive demographic attributes such as gender (e.g., Male/Female), race (e.g., African-American/European-American), or age (e.g., young/old). To address the above challenge, various strategies have been devised, including preprocessing datasets (Kamiran and Calders, 2012; Feldman et al., 2015), modifying the training process (Cotter et al., 2019; Lohaus et al., 2020), and calibrating outputs of trained models (Iosifidis et al., 2019; Chzhen et al., 2019). Predominantly, these methods have focused on settings where sensitive groups are identified by a *single* demographic attribute. However, recent studies (Yang et al., 2020; Buolamwini and Gebru, 2018; Kirk et al., 2021) demonstrate that ensuring fairness for an individual attribute does not guarantee *intersectional fairness*, which arises when considering *multiple* attributes concurrently (for example, comparing Male European-Americans or Female African-Americans). For instance, Buolamwini and Gebru (2018) found that several face recognition systems exhibit significantly higher error rates for darker-skinned females than for lighter-skinned males. These observations are inline with the hypothesis of Crenshaw (1989) that multiple sensitive attributes "intersect" to create unique effects.

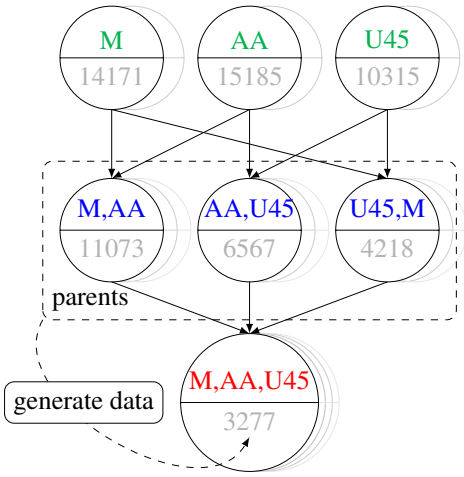


Figure 1: Snippet of the hierarchical structure found in intersectional fairness for Twitter Hate Speech Dataset (Huang et al., 2020) with 3 sensitive attributes. Here, 'M' stands for Male, 'AA' African American, and 'U45' age under 45 years. The group labeled 'M,AA,U45', represents African American men who are less than 45 years old, and has parent groups 'M,AA', 'M,U45', and 'AA,U45'. For each group, the number of examples is reported. The deeper we go in this hierarchical structure, the smaller the number of examples. Our approach consists in generating additional data for smaller groups by combining data from parent groups.

In response to emerging challenges, there has been a notable shift towards intersectionality in fair machine learning research (Filippi et al., 2023; Foulds et al., 2020). Among them, recent studies (Maheshwari et al., 2023; Zietlow et al., 2022; Mittelstadt et al., 2023) have highlighted that several methods improve intersectional fairness by ac-

tually harming the subgroups. In other words, they tend to decrease performance over individual subgroups to achieve better overall fairness, an effect referred to as "leveling down".

In this work, we hypothesize that leveling down can be countered by generating additional data for smaller groups so as to improve their representation. To this end, we propose a data augmentation mechanism that utilizes the hierarchical structure inherent to intersectionality. More precisely, we augment subgroups by modifying and combining data from parent groups (which generally have more examples). Figure 1 illustrates this hierarchical structure for the Twitter Hate Speech Dataset, showing how the group 'African American, Male, under 45' is composed of 'Male, African American', 'Male, under 45', and 'African American, under 45' groups. It also highlights the data scarcity challenge, showing that the number of samples often decreases sharply as we consider more intersections. For example, the 'African American, Male, under 45' group has 3,277 instances, whereas the 'Male' group has 14,171 instances.

In order to produce valuable examples despite limited data availability, we propose a simple parameterization of the generative model and train it using a loss based on Maximum Mean Discrepancy (MMD) (Gretton et al., 2012). This loss quantifies the difference between the original examples from a group and the examples generated by combining examples from its parent groups. Then, we train a classifier on the combination of original and generated examples, using equal sampling (Kamiran and Calders, 2009; González-Zelaya et al., 2021). The first step increases the diversity of examples the classifier is trained on, thereby improving generalization, while the latter ensures that equal importance is given to all subgroups instead of focusing more on larger groups. We empirically evaluate the quality and diversity of the generated examples and their impact on fairness and accuracy. Our results on various datasets show that our proposed approach consistently improves fairness, without harming the groups and at a small cost in accuracy.

2 Related Work

In this section, we provide a brief overview of approaches which specifically optimize intersectional fairness. For a more detailed overview, please refer to Appendix A. Foulds et al. (2020) introduced an in-processing technique that incorporates an intersectional fairness regularizer into the loss function, balancing fairness and accuracy. Conversely, Morina et al. (2019) suggests a post-processing mechanism that adjusts the threshold of the classifier and randomizes predictions for each subgroup independently. InfoFair (Kang et al., 2022) adopts a distinct approach by minimizing mutual information between predictions and sensitive attributes. Recently, research has begun to explore the phenomenon of "leveling down" in fairness. Maheshwari et al. (2023); Mittelstadt et al. (2023) argue that the strictly egalitarian perspective of current fairness measures contributes to this phenomenon. Meanwhile, Zietlow et al. (2022) demonstrates leveling down in computer vision contexts and introduces an adaptive augmented sampling strategy using generative adversarial networks (Goodfellow et al., 2014) and SMOTE (Chawla et al., 2002). Our work aligns with these developments; however, we propose a modality-independent technique that effectively leverages the intrinsic hierarchical structure of intersectionality.

3 Problem Statement

Let p denote the number of distinct *sensitive axes* of interest, which we denote as $\mathcal{A}_1, \ldots, \mathcal{A}_p$. Each of these sensitive axes is a set of discrete-valued *sensitive attributes*.

Consider a feature space $\mathcal{X}$, a finite discrete label space $\mathcal{Y}$. Let $\mathcal{D}$ be an unknown distribution over $\mathcal{X} \times \mathcal{Y} \times \mathcal{A}_1 \times \cdots \times \mathcal{A}_p$ which can be written as:

$$\mathcal{D} = P(X, Y, A_1, \cdots, A_p) \tag{1}$$

We define a *sensitive group* $\mathbf{g}$ as any p-dimensional vector in the Cartesian product set $\mathcal{G} = \mathcal{A}_1 \times \cdots \times \mathcal{A}_p$ of the sensitive axes. For instance, a sensitive group $\mathbf{g} \in \mathcal{G}$ can be represented as $(a_1, \ldots, a_p)$ with corresponding distribution as:

$$\mathcal{D}_{\mathbf{g}} = P(X, Y, A_1 = a_1, \cdots, A_p = a_p)$$

We also introduce a more general group than $\mathbf{g}$ called $\mathbf{g}^{\setminus i}$, referred to as the *parent group* in which the i-th sensitive axis is left underspecified. It can be represented as $(a_1, \cdots, a_{i-1}, a_{i+1}, \cdots, a_p)$ where $i \in \{1, \ldots, p\}$. The distribution over such a group can be written as:

$$\mathcal{D}_{\mathbf{g}^{\setminus i}} = \textstyle\sum_{a_i \in \mathcal{A}_i} P(X, Y, A_1 = a_1, \cdots, A_i = a_i, \cdots, A_p = a_p) \tag{2}$$

In our example above, if group $\mathbf{g}$ is {male, European American, under 45}, then the corresponding

parent groups are: {male, European American}, {male, under 45}, {European American, under 45}.

Finally, in this work, we focus on classification problems and assume K distinct labels. We will denote the distribution of a group conditioned on same label k by $\mathcal{D}_{\mathbf{g}|Y=k}$.

Problem Statement: As standard in machine learning, $\mathcal{D}$ is generally unknown and instead we have access to a finite dataset $\mathcal{T} = \{(x_j, y_j, \mathbf{g}_j)\}_{j=1}^{n}$ consisting of n i.i.d examples sampled from $\mathcal{D}$. This sample can be rewritten as $\mathcal{T} = \bigcup_{\mathbf{g}\in\mathcal{G}} \mathcal{T}_{\mathbf{g}}$ where $\mathcal{T}_{\mathbf{g}}$ represents the subset of examples from group $\mathbf{g}$. Examples belonging to parent group $\mathbf{g}^{\setminus i}$ are denoted by:

$$\mathcal{T}_{\mathbf{g}^{\setminus i}} = \bigcup_{a_i\in\mathcal{A}_i} \mathcal{T}_{a_1,\cdots,a_i,\cdots a_p} \tag{3}$$

The goal of fair machine learning is then to learn an accurate model $h \in \mathcal{H}$, such that $h : \mathcal{X} \to \mathcal{Y}$ is fair with respect to a given group fairness definition like Equal Opportunity (Hardt et al., 2016), Equal Odds (Hardt et al., 2016), Accuracy Parity (Zafar et al., 2017), etc.

4 Approach

In this work, we introduce a novel approach for generating data that leverages the underlying structure of intersectional groups. We begin by highlighting the structural properties of interest, and then present our data generation mechanism. Note that in this work, we treat data as vectors, which allows us to encompass a wide range of modalities including images and text. To convert data into vector representations, we may use pre-trained encoders.

4.1 Structure of the Data

Using the notations introduced in the previous section, we make the following simple but crucial observation about the structure of the data:

$$\mathcal{T}_{\mathbf{g}} = \bigcap_{i=1}^{p} \mathcal{T}_{\mathbf{g}^{\setminus i}} \quad \text{and} \quad \mathcal{T}_{\mathbf{g}} \subset \mathcal{T}_{\mathbf{g}^{\setminus i}} \forall i \in \{1,\ldots,p\}.$$

In other words, the intersection of immediate parent groups constitutes the target group **g**, with each parent group containing more examples than the target group itself. For example, all instances of the group Female African American are also part of both the Female and African American groups. Moreover, the common instances between the Female and African American groups collectively define the Female African American group.

4.2 Data Generation

Our goal is to learn a generative function $gen_{\theta,k}$ such that, given a dataset $\mathcal{T}$, a group $\mathbf{g}$, and task label k, the generated distribution $Z_{gen} \sim gen_{\theta,k}(\mathcal{T}, \mathbf{g})$ is similar to the true distribution $\mathcal{D}_{\mathbf{g}|Y=k}$. Based on the above observations, we propose to generate examples for group $\mathbf{g}$ by combining and transforming the examples from the corresponding parent groups. This can be achieved by appropriate parameterizations of $gen_{\theta,k}$ which we describe next.

Parameterization of the Generative Function: In this work, we explore the use of two simple choices for the generative function $gen_{\theta,k}(\mathcal{T}, \mathbf{g})$ that generates an example $Z_{gen} = (X_{gen}, k, \mathbf{g})$ for a given group $\mathbf{g}$ and label k. The first parameterization is:

$$X_{gen} = \sum_{i=1}^{p} \lambda_i X_{\mathbf{g}^{\setminus i}} \tag{4}$$

with $Z_{\mathbf{g}^{\setminus i}} = (X_{\mathbf{g}^{\setminus i}}, k, \mathbf{g}^{\setminus i}) \sim \mathcal{D}_{\mathbf{g}^{\setminus i}|Y=k}$. In the above equation, $\lambda = (\lambda_1, \ldots, \lambda_p) \in \mathbb{R}^p$ are the parameters to optimize based on the loss we define below. In other words, we generate data for group $\mathbf{g}$ by forming weighted combinations of examples from its parent groups.

The second parameterization we consider is:

$$X_{gen} = \sum_{i=1}^{p} W \cdot X_{\mathbf{g}^{\setminus i}}^{T}, \tag{5}$$

where $\mathbf{W} \in \mathbb{R}^{d\times d}$ is a diagonal matrix with d parameters where d is the dimension of the encoded inputs. Here, we use a uniform combination of examples from parent groups, but learn weights for the different features of the representation.

Given the limited data available for many groups, we opt to share parameters across them instead of learning specific parameters for each group. This approach, combined with the relatively simple parameterizations of the generative function, serves to reduce the risk of overfitting (recall that in practice we have very limited data for many groups). However, we still learn a separate model for each label, i.e., $gen_{\theta,k}(\mathcal{T}, \mathbf{g})\ \forall k \in K$, to avoid the added complexity of jointly learning $\mathcal{X} \times \mathcal{Y}$.

Training the Generative Models: To train the generative model $gen_{\theta,k}$, we minimize a loss based on Maximum Mean Discrepancy (MMD). MMD is a non-parametric kernel-based divergence used to assess the similarity between distributions by using samples drawn from those distributions (Gretton

et al., 2012). Formally, the MMD between two samples $S = (z_1, \ldots, z_m)$ and $S' = (z'_1, \ldots, z'_m)$ can be written as

$$MMD^2(S, S') = \frac{1}{m(m-1)} \big[\sum_i \sum_{j \neq i} k(z_i, z_j) + \sum_i \sum_{j \neq i} k(z'_i, z'_j) \big] + \frac{1}{m^2} \sum_i \sum_j k(z_i, z'_j)$$

where k is a reproducing kernel. In this work, we use the radial basis function kernel $k : (z, z') \mapsto \exp(\|z - z'\|^2 / 2\sigma^2)$ where σ is a free parameter. For completeness, more details about MMD are given in Appendix B

Our loss function is the MMD between the generated samples and the samples from group $\mathbf{g}$, to which we add the MMD between the generated samples and those from its parent groups.[1] Formally, this can be written as:

$$L_{\mathbf{g},k}(\theta) = MMD(S_{gen}, S_{\mathbf{g},k}) + \sum_{i=1}^{p} MMD(S_{gen}, S_{\mathbf{g}^{\setminus i},k}), \quad (6)$$

where S_{gen} is a batch of examples generated from $gen_{\theta,k}$, $S_{\mathbf{g},k}$ and $S_{\mathbf{g}^{\setminus i},k}$ are batches of examples respectively drawn from $\mathcal{D}_{\mathbf{g}|Y=k}$ and $\mathcal{D}_{\mathbf{g}^{\setminus i}|Y=k}$. Since $\mathcal{D}_{\mathbf{g}|Y=k}$ and $\mathcal{D}_{\mathbf{g}^{\setminus i}|Y=k}$ are unknown, we approximate them with the empirical distribution by sampling with replacement from $\mathcal{T}_{\mathbf{g}|Y=k}$ and $\mathcal{T}_{\mathbf{g}^{\setminus i}|Y=k}$. Appendix C details the precise training process to learn the generative models.

Training Classifiers on Augmented Data: After training the generative models $gen_{\theta,k}$, we use them to create additional training data. Specifically, for a group $\mathbf{g}$, we sample examples from its corresponding parent groups and pass these samples through the generative models as previously described. In this way, we can generate additional data for smaller groups that we use to augment the original training dataset, so as to enhance their representation in downstream tasks. As we will see in the next section, this helps to improve the fairness of the classifier.

Alternative formulations: An alternative approach to learn $gen_{\theta,k}$ involves using a generative adversarial network (GAN) (Goodfellow et al., 2014). In this setup, the adversary aims to differentiate between two distributions, while the encoder strives to mislead the adversary. However, training GANs presents notable challenges (Thanh-Tung and Tran, 2020; Bau et al., 2019), including the risk of mode collapse, the complexity of nested optimization, and substantial computational demands. By contrast, MMD is more straightforward to implement and train, with significantly less computational burden. We also note that, while this work primarily employs MMD, our methodology can be adapted to work with other divergences between distributions, such as Sinkhorn Divergences and the Fisher-Rao Distance. We keep the exploration of other choices of divergences for future work.

5 Experiments

Our experiments are designed to (i) assess the quality of the data generated by our approach, and (ii) examine the influence of this data on fairness with a focus on avoiding leveling down as well as maximizing the classification performance for the worst-off group. Before presenting results, we start by outlining the datasets, baselines, and fairness metrics we employ. The code base is available here[2].

Datasets: To demonstrate the broad applicability of our proposed approach, we used four diverse datasets varying in size, demographic diversity, and modality, encompassing both text and images. These datasets are: (i) *Twitter Hate Speech* (Huang et al., 2020) comprising of tweets annotated with 4 demographic attributes; (ii) *CelebA* (Liu et al., 2015) composed of human face images annotated with various attributes; (iii) *Numeracy* (Abbasi et al., 2021) compiles free text responses denoting the numerical comprehension capabilities of individuals; and (iv) *Anxiety* (Abbasi et al., 2021): indicative of a patient's anxiety levels. Experimental setup, splits, and preprocessing are identical to those of Maheshwari et al. (2023). Detailed descriptions are available in the Appendix D.1.

Methods: We benchmark against 6 baselines, encompassing both generative approaches and methods optimizing for intersectional fairness: (i) **`Unconstrained`** solely optimizes model accuracy, ignoring any fairness measure; (ii) **`Adversarial`** adds an adversary (Li et al., 2018) to **`Unconstrained`**, implementing standard adversarial learning approach; (iii) **`FairGrad`** (Maheshwari and Perrot, 2022) is an in-processing iterative method that adjusts gradients for groups based on fairness levels; (iv) **`INLP`** (Ravfogel et al., 2020) is a post-processing approach that iteratively trains a

[1] In our preliminary set of experiments, we found this additional term brought more diversity in the generated examples.

[2] Please check the supplementary material. The final version will be released on GitHub with camera ready version.

classifier and then projects the representation on its null space; (v) **`Fair MixUp`** (Chuang and Mroueh, 2021) is a generative approach which enforces fairness by forcing the model to have similar predictions on samples generated by interpolating examples belonging to different sensitive groups; (vi) **DF Classifier** (Foulds et al., 2020) adds a regularization tailored to improve intersectional fairness. Our approach **`Augmented`** is same as **`Unconstrained`**, but trained on data generated via our proposed data generation mechanism.

In all experiments we employ a three-layer fully connected neural network with hidden layers of sizes 128, 64, and 32 as our classifier. Furthermore, we use ReLU as the activation with dropout fixed to 0.5. Cross-entropy loss is optimized in all cases, employing the Adam optimizer (Kingma and Ba, 2015) with its default parameters. Finally, for text-based datasets we encode the text using bert-base-uncased (Devlin et al., 2019) and for images we employ a pre-trained ResNet18[3] (He et al., 2016). Finally, we use equal sampling as shown effective in previous works (Maheshwari et al., 2023; Kamiran and Calders, 2009; González-Zelaya et al., 2021), ensuring equal number of examples for each group. The number of examples, treated as a hyperparameter, spans a spectrum from undersampling to oversampling regime. For more detailed description of hyperparameters and compute infrastructure, please refer to Appendix D.2.

To generate data for `Augmented`, we employ the generative function as described in Section 4.2. More specifically, our initial experiments suggest that employing a simpler model with fewer parameters (Equation 4) for the positive class, and a more complex model with a larger number of parameters for the negative class (Equation 5), leads to an enhanced fairness-accuracy trade-off, when using the False Positive rate as fairness measure. Consequently, for the positive class, we implement the function detailed in Equation 4, and for the negative class, we apply the model specified in Equation 5.

Fairness Metrics: To assess unfairness, we utilize two fairness definitions specifically designed for the context of intersectional fairness: α-Intersectional Fairness (IF_α) (Maheshwari et al., 2023) and Differential Fairness (DF) (Foulds et al., 2020). Detailed descriptions of these metrics are provided in the Appendix (Section D.3).

For the performance measure m associated with these definitions, we focus on False Positive Rate. Formally, for a group $\mathbf{g}$, m is given by:

$$m(h_\theta, \mathcal{T}_\mathbf{g}) = 1 - P(h_\theta(x) = 0 | (x, y) \in \mathcal{T}_\mathbf{g}, y = 1)$$

To estimate these empirical probabilities, we adopt the bootstrap estimation method proposed by Morina et al. (2019). We generate 1000 datasets by sampling from the original dataset with replacement. We then estimate the probabilities on this dataset using a smoothed empirical estimation mechanism and then average the results over all the sampled datasets. In addition to these fairness metrics, we report the performance measure for both the best and worst-performing groups.

Utility metric: In order to evaluate the utility of various methods, we employ balanced accuracy.

5.1 Quality of Generated Data

In this experiment, we assess the quality and diversity of data generated by our approach. Our goal is to generate data that resemble the overall distribution of real data, while ensuring the generated examples remain distinct from the original samples. To this end, we propose two evaluations:

- **Diversity:** for each generated example, we identify the most similar example in the real dataset. If the generated sample closely resembles a real one, the distance between the generated and real examples will be substantially smaller than between distinct real examples.
- **Distinguishability:** we train a classifier to differentiate between generated and real datasets. If the classifier's accuracy approaches that of a random guess, it suggests the empirical distributions of the generated and real data are similar.

In both experiments, we report metrics based on the entire dataset rather than computing averages for each group and then aggregating averages.

5.1.1 Diversity

In this experiment, we use cosine similarity as a measure of closeness. We generate 1000 examples and randomly select an equivalent number from the actual (real) dataset. For each real example, we find its nearest counterpart within the actual dataset to establish a baseline, termed 'R-R'. Then, for every generated example, we identify the closest match in the actual dataset, referred to as 'G-R'. To

[3] https://pytorch.org/vision/stable/models.html

Dataset	G-R	R-R	G-G
CelebA	0.46	0.48	0.44
Numeracy	0.51	0.58	0.45
Anxiety	0.51	0.59	0.46
Twitter Hate Speech	0.47	0.53	0.45

Table 1: Analyzing the similarity of a generated sample with existing sample. For clarity and ease of readability, we have omitted the standard deviation in our reporting, as it remained below $\pm$ 0.01 across all settings.

Dataset	Accuracy
CelebA	0.52 ± 0.011
Numeracy	0.64 ± 0.012
Anxiety	0.64 ± 0.019
Twitter Hate Speech	0.57 ± 0.022

Table 2: Accuracy of a classifier to distinguish between real and generated sample over various datasets. The value of 0.5 represents a random classifier, while 1.0 is a perfect classifier.

further assess diversity, we also present results of the closest match of each generated sample in the generated dataset, called 'G-G'. The results of this experiment are presented in Table 1.

Across all datasets, we observe that the distance between generated and real examples is similar to the distance observed between two real examples. In each dataset, the closeness between G-R pairs is less than that observed in R-R pairs. Moreover, the G-G pairs exhibit lower similarity scores compared to R-R pairs, suggesting greater diversity in the generated dataset. Based on these results, we conclude that the generated examples are diverse and not mere replicas of the real samples.

5.1.2 Distinguishability

We frame distinguishability as a binary classification task where we train a two-layer MLP classifier aimed at distinguishing between real and generated samples. Again, we compile a dataset by selecting 1000 real instances and 1000 generated samples. This dataset is subsequently partitioned into training and test sets with a ratio of 80% to 20%.

Results are presented in Table 2. The mean accuracy of the classifier is approximately 0.59, suggesting that the generated samples have a distribution similar, but not identical to, the real instances. In our preliminary experiments we found that by modulating the generator complexity (i.e by employing more complex models with more parameters), we could achieve near-random distinguishability. However, such adjustments led to an unfavorable fairness-accuracy trade-off. We conjecture this may arise because near-random indistinguishability in the generated samples causes them to inherit biases from the real data.

5.2 Fairness-Accuracy Trade-offs

In this experiment, we explore the impact of generated data on the fairness-privacy trade-off and compare our approach to existing fairness-promoting methods. We pay particular attention to the leveling down phenomenon: a method is considered to exhibit leveling down if its performance for the worst-off or best-off group is inferior to that of the unconstrained model.

The outcomes of this experiment is presented in Table 3. Detailed results for CelebA and Numeracy, both of which display a similar trend, are provided in Appendix D.4. In terms of accuracy, `Augmented` exhibits a slight drop for the Anxiety dataset. However, its accuracy is on par with the Unconstrained model when evaluated on Twitter Hate Speech. In terms of performance for both best-off and worst-off groups, `Augmented` outperforms competing methods. Notably, `Augmented` does not show any signs of leveling down across all datasets. When assessing IF_α with $\alpha = 0.5$, `Augmented` consistently achieves the best fairness results among the datasets. We also plot the complete trade-off between relative and absolute performance of groups by varying α in Figure 4 in Appendix D.4. For the Anxiety dataset, `Augmented` gives the best trade-off for every value of α. In the case of Twitter Hate Speech, `INLP` achieves comparable results, although with a noticeable drop in accuracy (14 points below `Augmented`). Overall, the results show that our `Augmented` gives a superior accuracy-fairness trade-off and successfully avoids leveling down.

5.3 Impact of Intersectionality

In this experiment, we examine the influence of intersectionality on our approach and its effect on worst-case performance. To this end, we iteratively introduce more sensitive axes and plot the worst case performance. For example, akin to the experiment in (Maheshwari et al., 2023) using CelebA, we initially consider gender as a single sensitive axis. In the subsequent step, we incorporate age

Method	BA ↑	Best Off ↓	Worst Off ↓	DF ↓	$IF_{0.5}$ ↓
Unconstrained	0.63 + 0.01	0.25 + 0.02	0.51 + 0.03	0.43 +/- 0.09	0.52 +/- 0.03
Adversarial	0.63 + 0.01	0.27 + 0.06	0.55 + 0.12	0.48 +/- 0.05	0.55 +/- 0.04
FairGrad	0.63 + 0.01	0.29 + 0.05	0.56 + 0.12	0.48 +/- 0.07	0.57 +/- 0.04
INLP	0.63 + 0.01	0.22 + 0.02	0.49 + 0.03	0.42 +/- 0.07	0.48 +/- 0.03
Fair MixUp	0.61 + 0.01	0.28 + 0.02	0.55 + 0.06	0.47 +/- 0.09	0.55 +/- 0.02
DF-Classifier	0.63 + 0.01	0.29 + 0.08	0.56 + 0.09	0.48 +/- 0.17	0.56 +/- 0.08
Augmented	0.6 + 0.0	0.13 + 0.08	0.35 + 0.12	0.29 +/- 0.32	0.39 +/- 0.11

(a) Results on Anxiety

Method	BA ↑	Best Off ↓	Worst Off ↓	DF ↓	$IF_{0.5}$ ↓
Unconstrained	0.81 + 0.0	0.18 + 0.01	0.46 + 0.01	0.42 +/- 0.05	0.46 +/- 0.02
Adversarial	0.79 + 0.01	0.18 + 0.01	0.48 + 0.04	0.46 +/- 0.08	0.47 +/- 0.02
FairGrad	0.8 + 0.0	0.17 + 0.01	0.49 + 0.03	0.49 +/- 0.1	0.44 +/- 0.02
INLP	0.66 + 0.0	0.08 + 0.02	0.26 + 0.02	0.22 +/- 0.25	0.29 +/- 0.04
Fair MixUp	0.81 + 0.01	0.18 + 0.02	0.46 + 0.02	0.42 +/- 0.09	0.45 +/- 0.04
DF-Classifier	0.81 + 0.0	0.13 + 0.01	0.45 + 0.02	0.46 +/- 0.1	0.39 +/- 0.03
Augmented	0.81 + 0.0	0.06 + 0.01	0.36 + 0.03	0.38 +/- 0.13	0.27 +/- 0.02

(b) Results on Twitter Hate Speech

Table 3: Test results on (a) *Anxiety*, and (b) *Twitter Hate Speech*. We select hyperparameters based on $IF_{0.5}$ value. The utility of various approaches is measured by balanced accuracy (BA), whereas fairness is measured by differential fairness (DF) and intersectional fairness ($IF_{0.5}$) on the False Positive Rate (FPR). For both fairness definitions, lower is better, while for balanced accuracy, higher is better. Best Off and Worst Off represent the min FPR and max FPR across groups (in both cases, lower is better). Results have been averaged over 5 different runs.

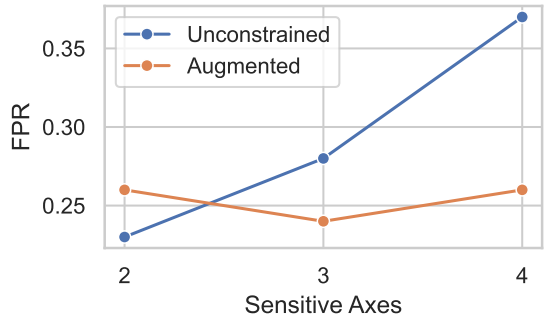


Figure 2: FPR of worst-off group on *CelebA* (the lower, the better) by varying the number of sensitive axes.

alongside gender. Similarly, we then add attractiveness, and finally skin color.

The results of this experiment can be found in Figure 2. With fewer groups (2 sensitive axes), the model's performance on the generated dataset closely matches that on the real dataset. However, as the number of axes increases, the performance difference becomes more pronounced. Furthermore, we find that the performance of the model remains relatively stable despite the increase in sensitive axes, further underscoring the effectiveness of our proposed approach.

5.4 Alternative Structures

Our proposed approach generates additional data for a target group with data from its corresponding parent groups. In this experiment, we explore alternative structures. Taking the target group $\mathbf{g}$ composed of {male, European American, under 45} as an example, we examine two distinct structures:

- `Alternate`: Here, we use examples from parent groups unrelated to the target group. More specifically, we follow an adversarial approach where we choose parents such that they share no examples with the target group. For instance, for group $\mathbf{g}$, we define the adversarial group $\neg\mathbf{g}$ as {Female, African American, above 45}. We then draw examples from parents of group $\neg\mathbf{g}$ for training our generative model for $\mathbf{g}$. We provide the exact formalism and setup in Appendix D.5.

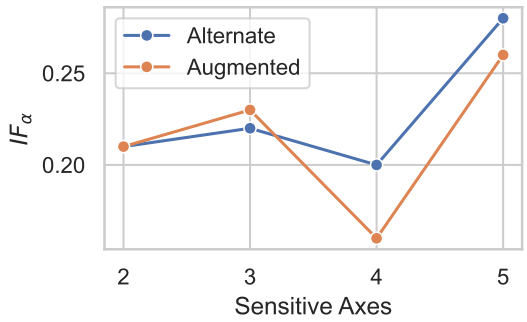


Figure 3: $IF_{0.5}$ comparison between `Augmented` and `Alternate` by varying the number of sensitive axes on CelebA. With a smaller number of sensitive axes, `Unconstrained` and `Alternate` exhibit comparable performance. However, as the number of sensitive axes increases, `Augmented` begins to outperform `Alternate`.

- `Abstract`: Here, we use examples from the parents of parents of the target group. For example, for g, the immediate parent groups are: ({male, European American}, {male, under 45}, {European American, under 45}). Instead of drawing examples from these immediate parent groups, we use examples from the parents of these parent groups, namely ({male}, {European American}, {under 45}).

Method	BA ↑	$IF_{0.5}$ ↓
`Unconstrained`	0.63	0.52
`Augmented`	0.60	0.39
`Alternate`	0.61	0.40
`Abstract`	0.59	0.43

(a) Results on Anxiety

Method	BA ↑	$IF_{0.5}$ ↓
`Unconstrained`	0.81	0.46
`Augmented`	0.81	0.27
`Alternate`	0.81	0.29
`Abstract`	0.81	0.32

(b) Results on Twitter Hate Speech

Table 4: Test results on (a) *Anxiety*, and (b) *Twitter Hate Speech* using False Positive Rate showcasing Balanced Accuracy (BA) and $IF_{0.5}$

The results of these experiments are provided in Table 4. For both experiments, we find that any form of data augmentation approach including the `Alternate` improves fairness. For instance, on Anxiety, `Alternate` significantly outperforms `Unconstrained` and reaches the same level of fairness as `Augmented`. Similarly, the `Abstract` approach outperforms `Unconstrained` on Anxiety. These observations indicate that data augmentation via combining from different groups is a viable strategy in general. However, when comparing `Abstract` performance with `Augmented`, we find that `Augmented` generally outperforms `Abstract`. We hypothesize that this occurs because considering more abstract groups approximates a scenario where no groups are considered, which is similar to an unconstrained.

Interestingly, we find that for Twitter Hate Speech and Anxiety, `Alternate` performs similarly to `Augmented` (with a small advantage to the latter in terms of fairness). We hypothesize the hierarchical structure leveraged in `Augmented` becomes more relevant with the increase in the number of sensitive axes as it provides better inductive bias. To test this hypothesis, we conducted an experiment akin to that in Section 5.3 where we gradually increase the number of sensitive axes in CelebA. The findings, illustrated in Figure 3, indicate that with a limited number of sensitive axes, both approaches yield comparable results. However, as the number of axes increases, `Augmented` generally outperforms the `Alternate`. It is important to note that these results might also be influenced by inherent dataset characteristics, such as modality, size, and diversity. A comprehensive exploration of how these characteristics interact with the optimal structure for generating augmented data is an interesting avenue for future research. In summary, our experiments show that data augmentation across groups is a viable strategy for enhancing the fairness of machine learning models in intersectional scenarios.

6 Conclusion

In this paper, we introduce a data augmentation mechanism that leverages the hierarchical structure inherent to intersectional fairness. Our extensive experiments demonstrate that this method not only generates diverse data but also enhances the classifier's performance across both the best-off and worst-off groups. In the future, we plan to extend our approach to a broader range of performance metrics, delve into zero-shot fairness, and explore more sophisticated sampling mechanisms.

7 Limitations

While appealing, our proposed data generation mechanism is not without limitations. Its primary constraint is the assumption of accurate sensitive annotations for each data point. Inaccurate or missing annotations could lead to scenarios where an otherwise fair model inadvertently harms groups with incorrect or missing annotations. Additionally, this mechanism adopts a static view of fairness, failing to account for issues like data drift, which may result in the model becoming unfair over time. Furthermore, despite our experiments indicating superior performance, our evaluation is confined to the specific datasets and settings we tested. We advise practitioners employing this approach to conduct thorough evaluations of the model, considering the unique aspects of their intended application.

References

Ahmed Abbasi, David G. Dobolyi, John P. Lalor, Richard G. Netemeyer, Kendall Smith, and Yi Yang. 2021. Constructing a psychometric testbed for fair natural language processing. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*, pages 3748–3758. Association for Computational Linguistics.

Solon Barocas, Elizabeth Bradley, Vasant Honavar, and Foster Provost. 2017. Big data, data science, and civil rights. *arXiv preprint arXiv:1706.03102*.

David Bau, Jun-Yan Zhu, Jonas Wulff, William Peebles, Hendrik Strobelt, Bolei Zhou, and Antonio Torralba. 2019. Seeing what a gan cannot generate. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4502–4511.

Elizabeth Buchanan. 2012. Ethical decision-making and internet research. *Association of Internet Researchers*.

Joy Buolamwini and Timnit Gebru. 2018. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on Fairness, Accountability and Transparency, FAT 2018, 23-24 February 2018, New York, NY, USA*, volume 81 of *Proceedings of Machine Learning Research*, pages 77–91. PMLR.

Jenna Burrell. 2016. How the machine 'thinks': Understanding opacity in machine learning algorithms. *Big data & society*, 3(1):2053951715622512.

Toon Calders and Sicco Verwer. 2010. Three naive bayes approaches for discrimination-free classification. *Data mining and knowledge discovery*, 21(2):277–292.

Flavio P Calmon, Dennis Wei, Bhanukiran Vinzamuri, Karthikeyan Natesan Ramamurthy, and Kush R Varshney. 2017. Optimized pre-processing for discrimination prevention. volume 30.

Nitesh V. Chawla, Kevin W. Bowyer, Lawrence O. Hall, and W. Philip Kegelmeyer. 2002. SMOTE: synthetic minority over-sampling technique. *J. Artif. Intell. Res.*, 16:321–357.

Ching-Yao Chuang and Youssef Mroueh. 2021. Fair mixup: Fairness via interpolation. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.

Evgenii Chzhen, Christophe Denis, Mohamed Hebiri, Luca Oneto, and Massimiliano Pontil. 2019. Leveraging labeled and unlabeled data for consistent fair binary classification. *arXiv preprint arXiv:1906.05082*.

European Commission. 2018. Communication artificial intelligence for europe.

Andrew Cotter, Heinrich Jiang, and Karthik Sridharan. 2019. Two-player games for efficient non-convex constrained optimization. In *Algorithmic Learning Theory*, pages 300–332. PMLR.

Kimberle Crenshaw. 1989. Demarginalizing the intersection of race and sex: A black feminist critique of antidiscrimination doctrine, feminist theory and antiracist politics. *The University of Chicago Legal Forum*, 140:139–167.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 4171–4186. Association for Computational Linguistics.

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

Giulio Filippi, Sara Zannone, and Adriano S. Koshiyama. 2023. Intersectional fairness: A fractal approach. *CoRR*, abs/2302.12683.

James R. Foulds, Rashidul Islam, Kamrun Naher Keya, and Shimei Pan. 2020. An intersectional definition of fairness. In *36th IEEE International Conference on Data Engineering, ICDE 2020, Dallas, TX, USA, April 20-24, 2020*, pages 1918–1921. IEEE.

Vladimiro González-Zelaya, Julián Salas, Dennis Prangle, and Paolo Missier. 2021. Optimising fairness through parametrised data sampling. In *EDBT*, pages 445–450.

Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron C. Courville, and Yoshua Bengio. 2014. Generative adversarial networks. *CoRR*, abs/1406.2661.

Arthur Gretton, Karsten M. Borgwardt, Malte J. Rasch, Bernhard Schölkopf, and Alexander J. Smola. 2012. A kernel two-sample test. *J. Mach. Learn. Res.*, 13:723–773.

Moritz Hardt, Eric Price, and Nati Srebro. 2016. Equality of opportunity in supervised learning. In *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, December 5-10, 2016, Barcelona, Spain*, pages 3315–3323.

Charles R. Harris, K. Jarrod Millman, Stéfan J. van der Walt, Ralf Gommers, Pauli Virtanen, David Cournapeau, Eric Wieser, Julian Taylor, Sebastian Berg, Nathaniel J. Smith, Robert Kern, Matti Picus, Stephan Hoyer, Marten H. van Kerkwijk, Matthew Brett, Allan Haldane, Jaime Fernández del Río, Mark Wiebe, Pearu Peterson, Pierre Gérard-Marchant, Kevin Sheppard, Tyler Reddy, Warren Weckesser, Hameer Abbasi, Christoph Gohlke, and Travis E. Oliphant. 2020. Array programming with NumPy. *Nature*, 585(7825):357–362.

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

Xiaolei Huang, Linzi Xing, Franck Dernoncourt, and Michael J. Paul. 2020. Multilingual Twitter corpus and baselines for evaluating demographic bias in hate speech recognition. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 1440–1448, Marseille, France. European Language Resources Association.

Vasileios Iosifidis, Besnik Fetahu, and Eirini Ntoutsi. 2019. Fae: A fairness-aware ensemble framework. In *2019 IEEE International Conference on Big Data (Big Data)*, pages 1375–1380. IEEE.

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

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

Jian Kang, Tiankai Xie, Xintao Wu, Ross Maciejewski, and Hanghang Tong. 2022. Infofair: Information-theoretic intersectional fairness. In *IEEE International Conference on Big Data, Big Data 2022, Osaka, Japan, December 17-20, 2022*, pages 1455–1464. IEEE.

Michael J. Kearns, Seth Neel, Aaron Roth, and Zhiwei Steven Wu. 2018. Preventing fairness gerrymandering: Auditing and learning for subgroup fairness. In *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018*, volume 80 of *Proceedings of Machine Learning Research*, pages 2569–2577. PMLR.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Hannah Rose Kirk, Yennie Jun, Filippo Volpin, Haider Iqbal, Elias Benussi, Frédéric A. Dreyer, Aleksandar Shtedritski, and Yuki M. Asano. 2021. Bias out-of-the-box: An empirical analysis of intersectional occupational biases in popular generative language models. In *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, pages 2611–2624.

John Lalor, Yi Yang, Kendall Smith, Nicole Forsgren, and Ahmed Abbasi. 2022. Benchmarking intersectional biases in NLP. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL 2022, Seattle, WA, United States, July 10-15, 2022*, pages 3598–3609. Association for Computational Linguistics.

Yitong Li, Timothy Baldwin, and Trevor Cohn. 2018. Towards robust and privacy-preserving text representations. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 25–30, Melbourne, Australia. Association for Computational Linguistics.

Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. 2015. Deep learning face attributes in the wild. In *2015 IEEE International Conference on Computer Vision, ICCV 2015, Santiago, Chile, December 7-13, 2015*, pages 3730–3738. IEEE Computer Society.

Michael Lohaus, Michaël Perrot, and Ulrike Von Luxburg. 2020. Too relaxed to be fair. In *International Conference on Machine Learning*, pages 6360–6369. PMLR.

Gaurav Maheshwari, Aurélien Bellet, Pascal Denis, and Mikaela Keller. 2023. Fair without leveling down: A new intersectional fairness definition. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 9018–9032, Singapore. Association for Computational Linguistics.

Gaurav Maheshwari and Michaël Perrot. 2022. Fairgrad: Fairness aware gradient descent. *CoRR*, abs/2206.10923.

Jacob Metcalf and Kate Crawford. 2016. Where are human subjects in big data research? the

emerging ethics divide. *Big Data & Society*, 3(1):2053951716650211.

Brent D. Mittelstadt, Sandra Wachter, and Chris Russell. 2023. The unfairness of fair machine learning: Levelling down and strict egalitarianism by default. *CoRR*, abs/2302.02404.

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

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Köpf, Edward Z. Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. 2019. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 8024–8035.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Shauli Ravfogel, Yanai Elazar, Hila Gonen, Michael Twiton, and Yoav Goldberg. 2020. Null it out: Guarding protected attributes by iterative nullspace projection. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 7237–7256. Association for Computational Linguistics.

Hoang Thanh-Tung and Truyen Tran. 2020. Catastrophic forgetting and mode collapse in gans. In *2020 international joint conference on neural networks (ijcnn)*, pages 1–10. IEEE.

Laura Weidinger, John Mellor, Maribeth Rauh, Conor Griffin, Jonathan Uesato, Po-Sen Huang, Myra Cheng, Mia Glaese, Borja Balle, Atoosa Kasirzadeh, et al. 2021. Ethical and social risks of harm from language models. *arXiv preprint arXiv:2112.04359*.

Forest Yang, Mouhamadou Cisse, and Oluwasanmi Koyejo. 2020. Fairness with overlapping groups; a probabilistic perspective. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

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

Dominik Zietlow, Michael Lohaus, Guha Balakrishnan, Matthäus Kleindessner, Francesco Locatello, Bernhard Schölkopf, and Chris Russell. 2022. Leveling down in computer vision: Pareto inefficiencies in fair deep classifiers. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 10400–10411. IEEE.

A Additional Related Work

With ML rapidly automating several key aspects of decsion making the potential for harm has sparked calls for greater accountability and transparency by researchers (Weidinger et al., 2021; Burrell, 2016; Metcalf and Crawford, 2016), government agencies (Commission, 2018; Barocas et al., 2017) and NGOs (Buchanan, 2012). This has peaked interest in fairness, with researchers responding in two primary ways: (i) Capturing and defining unfairness by proposing new metrics and evaluation suites, and (ii) developing mechanism to mitigate the unfairness.

The most prevalent approach to assessing intersectional unfairness involves comparing subgroup performances either with the overall population, as in subgroup fairness (Kearns et al., 2018), or with the best and worst performing subgroups, as in Differential Fairness (Foulds et al., 2020). Recent observations by (Maheshwari et al., 2023; Mittelstadt et al., 2023) suggest that solely focusing on relative performance among subgroups, without considering absolute performance, can lead to a phenomenon known as "leveling down." To address this, they recommend a hybrid metric, IF_α, which combines relative and absolute performance measures. Further details about these metrics are provided in the Appendix D.3.

Mitigation techniques can be typically categorized into three groups: (i) pre-processing, involving modifications at the dataset level (Kamiran and Calders, 2012; Feldman et al., 2015; Calmon et al., 2017); (ii) post-processing, which adjusts the outputs of pre-trained models that may exhibit biases (Iosifidis et al., 2019; Chzhen et al., 2019); and (iii) in-processing, entailing alterations to the training process and the model itself to enhance fairness (Cotter et al., 2019; Lohaus et al., 2020; Calders and Verwer, 2010).

In terms of approaches that specifically optimize intersectional fairness, Foulds et al. (2020) introduced an in-processing technique that incorporates a fairness regularizer into the loss function, balancing fairness and accuracy. Conversely, Morina et al. (2019) suggests a post-processing mechanism that adjusts the threshold of the classifier and randomizes predictions for each subgroup independently. InfoFair (Kang et al., 2022) adopts a distinct approach by minimizing mutual information between predictions and sensitive attributes. Recently, research has begun to explore the phenomenon of "leveling down" in fairness. Maheshwari et al. (2023); Mittelstadt et al. (2023) argue that the strictly egalitarian perspective of current fairness measures contributes to this phenomenon. Meanwhile, Zietlow et al. (2022) demonstrates leveling down in computer vision contexts and introduces an adaptive augmented sampling strategy using generative adversarial networks (Goodfellow et al., 2014) and SMOTE (Chawla et al., 2002). Our work aligns with these developments; however, we propose a modality-independent technique that effectively leverages the intrinsic structure of intersectionality.

B Background on Maximum Mean Discrepancy

Maximum Mean Discrepancy is an non-parametric kernel-based divergence used to assess the similarity between distributions. In a nutshell, it involves identifying an embedding function that, given two distributions $\mathcal{P}$ and $\mathcal{Q}$, yields larger values for samples drawn from $\mathcal{P}$ and smaller values for those from $\mathcal{Q}$. The difference in the mean value of this function for samples drawn from these two distributions provides an estimate of their similarity.

In this work, following the footsteps of Gretton et al. (2012), we use unit balls in characteristic reproducing kernel Hilbert spaces as the function class. Intuitively, the idea is to use the kernel trick to compute the differences in all moments of two distributions and then average the result. Formally, the MMD between two distributions $\mathcal{P}$ and $\mathcal{Q}$ is:

$$\begin{aligned} & MMD^2(\mathcal{P}, \mathcal{Q}) \\ = \;& \sup_{\|\Psi\|_H \leq 1} |E_{Z\sim\mathcal{P}}[\Psi(Z)] - E_{Z'\sim\mathcal{Q}}[\Psi(Z')]| \\ = \;& E_{Z\sim\mathcal{P}}[k(Z,Z)] - 2E_{Z\sim\mathcal{P},Z'\sim\mathcal{Q}}[k(Z,Z')] \\ + \;& E_{Z'\sim\mathcal{Q}}[k(Z',Z')] \end{aligned}$$

Here, k is the kernel derived from $\|\cdot\|_H$, the norm associated with corresponding Reproducing Kernel Hilbert Space H. In practice, we generally do not have access to true distributions but only samples, and thus the above equation is approximated as:

$$MMD^2(S_z, S_{z'}) = \frac{1}{m(m-1)}\big[\sum_i \sum_{j\neq i} k(z_i, z_j) + \sum_i \sum_{j\neq i} k(z'_i, z'_j)\big] + \frac{1}{m^2}\sum_i \sum_j k(z_i, z'_j)$$

where S_z (resp. $S_{z'}$) is a set of m samples drawn from $\mathcal{P}$ (resp. $\mathcal{Q}$). In this work, we use the radial basis function kernel $k : (z, z') \mapsto$

Algorithm 1 Training the Generative Models

Input: Groups $\mathcal{G}$, Dataset $\mathcal{T}$, batch size b, number of iterations l and batch size b

Output: K trained generative models $\{gen_{\theta,k}\}_{k=1}^{K}$ capable of generating data for each label k

for _ in l **do**
  Randomly sample a group $\mathbf{g}$ from $\mathcal{G}$
  for k in K **do**
    $S_{\mathbf{g},k} \leftarrow$ Sample b examples from $\mathcal{T}_{\mathbf{g}|Y=k}$
    $S_{\mathbf{g}^{\setminus i},k} \leftarrow$ Sample b examples from $\mathcal{T}_{\mathbf{g}^{\setminus i}|Y=k}\ \forall i \in \{1,\ldots,p\}$
    $S_{gen} \leftarrow$ Sample b examples from $gen_{\theta,k}(\mathcal{T},\mathbf{g})$
    Compute the MMD loss using these examples as stated in Equation 6
    Backpropagate this loss to update the parameters of the model $gen_{\theta,k}$
  end for
end for

$\exp(\|z-z'\|^2/2\sigma^2)$ where σ is the free parameter. In summary, MMD provides a simple way to compute the similarity between two distributions by using samples drawn from those distributions.

C Algorithm

The procedure to train our generative models is summarized in Algorithm 1.

D Experiments

D.1 Datasets

We benchmark our proposed generative approach over four datasets, and employ a similar setup as proposed by (Maheshwari et al., 2023). Note that all the datasets we experiment with are publicly available and can be used for research purpose.

- *CelebA* (Liu et al., 2015): It is composed of 202, 599 images of human faces. Additionally, each image is annotated with 40 binary attributes, such as 'eye glasses', 'bangs', and 'mustaches'. In our experiments, we set 'sex', 'Young', 'Attractive', and 'Pale Skin' attributes as the sensitive axis for the images and 'Smiling' as the class label. We split the dataset into 80% training of which 20% is used as validation, and the remaining 20% test split.

- *Twitter Hate Speech* (Huang et al., 2020): The dataset consists of tweets annotated with four race, age, gender, and country, We use the same pre-processing steps as employed by (Maheshwari et al., 2023), including binarizing the sensitive attributes, and focusing on English subset. After pre-processing, our train, validation and test sets consists of $22,818$, $4,512$, and $5,032$ tweets respectively.

- *Psychometric dataset* (Abbasi et al., 2021): The dataset consists of $8,502$ text responses alongside numerical scores provided by the physicians over several psychometric dimensions. Each response is also associated with four sensitive attributes, namely gender, race, and age. We focus on:
  - *Numeracy* which reflects the numerical comprehension ability of the patient.
  - *Anxiety* reflects the level of anxiety as described by the adult.

  We use same pre-processing as (Lalor et al., 2022) including binarizing the score. We use the same splitting procedure as described for CelebA dataset.

D.2 Hyperparameters

In all our experiments, we utilized an Intel Xeon CPU. Training a generative model on this platform typically takes about 15 minutes, whereas our fairness-accuracy experiments generally required about 30 minutes. For ease of replication, we will include the PyTorch model description in the README file accompanying the source code. All experiments were conducted using five different seeds: 10, 20, 30, 40, and 50. For the `Adversarial` approach, the λ parameter, which indicates the weight assigned to the adversarial branch, was set to the following values: 0.25, 0.5, 1.0, 5.0, 10.0, 50.0, 100.0. Similarly, for `Fair MixUp`, the mixup regularizer was assigned values of 0.25, 0.5, 1.0, 5.0, 10.0, 50.0, 100.0. For all other approaches, we used the default settings from the respective authors' codebases. The selection of optimal hyperparameters followed the procedure outlined in (Maheshwari and Perrot, 2022). In every experiment, we fixed the value of k at 0.03.

D.3 Fairness Definitions

In this work, we utilize two fairness definitions specifically formulated to assess intersectional fairness. Both definitions depend on group-wise performance measures, denoted as m, which can take

various forms, including Accuracy, True Positive Rate, and False Positive Rate. We focus on False Positive Rate for which the corresponding m is:

$$m(h_\theta, \mathcal{T}_\mathbf{g}) = 1 - P(h_\theta(x) = 0 | (x, y) \in \mathcal{T}_\mathbf{g}, y = 1)$$

- Differential Fairness: A model, denoted by h_θ, is considered to be ϵ-differentially fair (DF) wrt m, if

$$\text{DF}(h_\theta, m) \equiv \max_{\mathbf{g},\mathbf{g}' \in \mathcal{G}} \log \frac{m(h_\theta, \mathcal{T}_\mathbf{g})}{m(h_\theta, \mathcal{T}_{\mathbf{g}'})} \leq \epsilon.$$

- α-Intersectional Framework: A model h_θ is said to be (α, γ)-intersectionally fair (IF_α) with respect to m, if

$$\text{IF}_\alpha(h_\theta, m) \equiv \max_{\mathbf{g},\mathbf{g}' \in \mathcal{G}} I_\alpha(\mathbf{g}, \mathbf{g}', h_\theta, m) \leq \gamma.$$

  where $\mathbf{g}^w = \arg\min_{\mathbf{g} \in \mathcal{G}} m(h_\theta, \mathcal{T}_\mathbf{g})$ and $\mathbf{g}^b = \arg\max_{\mathbf{g} \in \mathcal{G}} m(h_\theta, \mathcal{T}_\mathbf{g})$. Here $I_\alpha(\mathbf{g}, \mathbf{g}', h_\theta, m)$ is defined as:

$$I_\alpha(\mathbf{g}, \mathbf{g}', h_\theta, m) = \alpha \Delta_{abs} + (1 - \alpha)\Delta_{rel}, \tag{7}$$

  where $\alpha \in [0, 1]$ and

$$\Delta_{abs} = \max\left(1 - m(h_\theta, \mathcal{T}_\mathbf{g}), 1 - m(h_\theta, \mathcal{T}_{\mathbf{g}'})\right),$$
$$\Delta_{rel} = \frac{1 - \max\left(m(h_\theta, \mathcal{T}_\mathbf{g}), m(h_\theta, \mathcal{T}_{\mathbf{g}'})\right)}{1 - \min\left(m(h_\theta, \mathcal{T}_\mathbf{g}), m(h_\theta, \mathcal{T}_{\mathbf{g}'})\right)}.$$

D.4 Results

We detail the additional experiments over the *CelebA* and *Numeracy* datasets. Table 5 shows results for fixed value of α. While Figure 4 plot the trade-off between relative and absolute performance over groups by varying α for all the datasets.

D.5 Alternate Structure

Recall that in `Alternate` approach, our aim is to draw examples from a different parent group set. More specifically, we follow an adversarial approach where we choose parents such that they share no examples with the group.

Formally, for a group $\mathbf{g}$ represented as $(a_1, \ldots, a_p)$, we define adversarial group as $\neg\mathbf{g}$ represented by $(\neg a_1, \ldots, \neg a_p)$. Note that, in this experiment we assume $\mathcal{A}_1, \ldots, \mathcal{A}_p$ to be binary discrete-valued. The generative function $gen_{\theta,k}(\mathcal{T}, \mathbf{g})$, akin to Equation 4, is defined as:

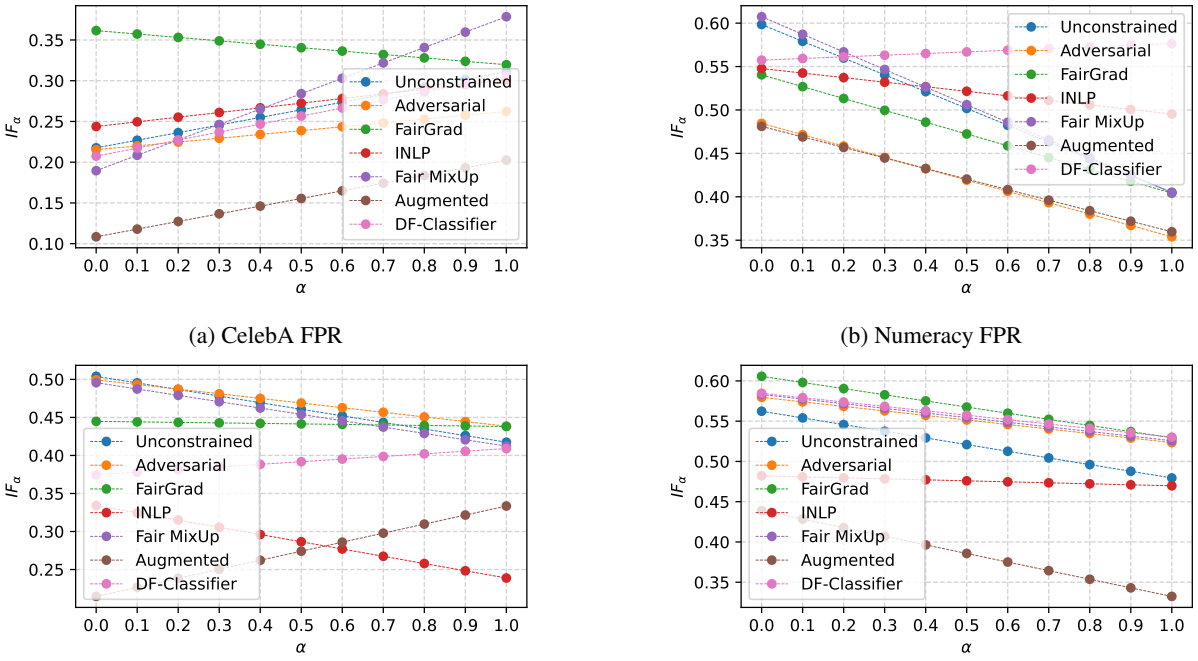
$$X_{gen} = \sum_{i=1}^{p} \lambda_i X_{\neg\mathbf{g}^{\setminus i}} \tag{8}$$

And the corresponding loss function akin to Equation 6 is:

$$L_{\mathbf{g},k}(\theta) = MMD(S_{gen}, S_{\mathbf{g},k}) + \sum_{i=1}^{p} MMD(S_{gen}, S_{\neg\mathbf{g}^{\setminus i},k}), \tag{9}$$

D.6 Tools

In all our experiments, we utilized Python and its associated machine learning libraries, including Numpy (Harris et al., 2020), PyTorch (Paszke et al., 2019), and scikit-learn (Pedregosa et al., 2011). Additionally, we employed ChatGPT for grammar correction.



(a) CelebA FPR

(b) Numeracy FPR

(c) Twitter Hate Speech FPR

(d) Anxiety FPR

Figure 4: Value of IF_α on the test set of various datasets by varying $\alpha \in [0, 1]$.

Method	BA	Best Off	Worst Off	DF	$\text{IF}_{0.5}$
Unconstrained	0.81 + 0.0	0.06 + 0.02	0.34 + 0.01	0.35 +/- 0.38	0.26 +/- 0.04
Adversarial	0.81 + 0.01	0.05 + 0.01	0.3 + 0.03	0.31 +/- 0.19	0.24 +/- 0.03
FairGrad	0.76 + 0.0	0.1 + 0.01	0.35 + 0.04	0.33 +/- 0.12	0.34 +/- 0.02
INLP	0.81 + 0.01	0.07 + 0.01	0.35 + 0.03	0.36 +/- 0.16	0.27 +/- 0.01
Fair MixUp	0.81 + 0.0	0.06 + 0.0	0.4 + 0.07	0.45 +/- 0.19	0.28 +/- 0.02
DF-Classifier	0.82 + 0.0	0.06 + 0.02	0.34 + 0.03	0.35 +/- 0.33	0.26 +/- 0.05
Augmented	0.76 + 0.01	0.02 + 0.0	0.21 + 0.03	0.22 +/- 0.21	0.16 +/- 0.01

(a) Results on CelebA

Method	BA	Best Off	Worst Off	DF	$\text{IF}_{0.5}$
Unconstrained	0.7 + 0.01	0.21 + 0.05	0.46 + 0.06	0.38 +/- 0.13	0.5 +/- 0.06
Adversarial	0.69 + 0.02	0.15 + 0.03	0.39 + 0.04	0.33 +/- 0.16	0.42 +/- 0.05
FairGrad	0.7 + 0.01	0.19 + 0.05	0.45 + 0.09	0.39 +/- 0.12	0.47 +/- 0.06
INLP	0.69 + 0.0	0.23 + 0.02	0.52 + 0.02	0.47 +/- 0.05	0.52 +/- 0.02
Fair MixUp	0.69 + 0.01	0.21 + 0.04	0.45 + 0.05	0.36 +/- 0.09	0.51 +/- 0.04
DF-Classifier	0.68 + 0.01	0.29 + 0.06	0.61 + 0.11	0.6 +/- 0.16	0.57 +/- 0.07
Augmented	0.69 + 0.02	0.14 + 0.05	0.39 + 0.11	0.34 +/- 0.24	0.44 +/- 0.07

(b) Results on Numeracy

Table 5: Test results on (a) *CelebA*, (b) *Numeracy*. We select hyperparameters based on $\text{IF}_{0.5}$ value. The utility of various approaches is measured by balanced accuracy (BA), whereas fairness is measured by differential fairness (DF) and intersectional fairness ($\text{IF}_{0.5}$) on the False Positive Rate (FPR). For both fairness definitions, lower is better, while for balanced accuracy, higher is better. The Best Off and Worst Off, in both cases lower is better, represents the min FPR and max FPR. Results have been averaged over 5 different runs.