Data Augmentation for Discrimination Prevention and Bias Disambiguation

Shubham Sharma
IBM Research and University of Texas
Yorktown Heights, New York, USA
Austin, Texas, USA
shubham_sharma@utexas.edu

Yunfeng Zhang
IBM Research
Yorktown Heights, New York, USA
zhangyun@us.ibm.com

Jesús M. Ríos Aliaga
IBM Research
Yorktown Heights, New York, USA
jriosal@us.ibm.com

Djallel Bouneffouf
IBM Research
Yorktown Heights, New York, USA
djallel.bouneffouf1@ibm.com

Vinod Muthusamy
IBM Research
Austin, Texas, USA
vmuthus@us.ibm.com

Kush R. Varshney
IBM Research
Yorktown Heights, New York, USA
krvarshn@us.ibm.com

ABSTRACT

Machine learning models are prone to biased decisions due to biases in the datasets they are trained on. In this paper, we introduce a novel data augmentation technique to create a fairer dataset for model training that could also lend itself to understanding the type of bias existing in the dataset i.e. if bias arises from a lack of representation for a particular group (sampling bias) or if it arises because of human bias reflected in the labels (prejudice based bias). Given a dataset involving a protected attribute with a privileged and unprivileged group, we create an "ideal world" dataset: for every data sample, we create a new sample having the same features (except the protected attribute(s)) and label as the original sample but with the opposite protected attribute value. The synthetic data points are sorted in order of their proximity to the original training distribution and added successively to the real dataset to create intermediate datasets. We theoretically show that two different notions of fairness: statistical parity difference (independence) and average odds difference (separation) always change in the same direction using such an augmentation. We also show submodularity of the proposed fairness-aware augmentation approach that enables an efficient greedy algorithm. We empirically study the effect of training models on the intermediate datasets and show that this technique reduces the two bias measures while keeping the accuracy nearly constant for three datasets. We then discuss the implications of this study on the disambiguation of sample bias and prejudice based bias and discuss how pre-processing techniques should be evaluated in general. The proposed method can be used by policy makers—who want to use unbiased datasets to train machine learning models for their applications—to add a subset of synthetic points to an extent that they are comfortable with to mitigate unwanted bias.

AIES '20, February 7–8, 2020, New York, NY, USA
© 2020 Association for Computing Machinery.
ACM ISBN 978-1-4503-7110-0/20/02…$15.00
https://doi.org/10.1145/3375627.3375865

CCS CONCEPTS

• **Social and professional topics → Computing / technology policy**; • **Applied computing → Law, social and behavioral sciences**; • **Computing methodologies →** *Machine learning*.

KEYWORDS

Fairness in Machine Learning, Responsible Artificial Intelligence, Discrimination Prevention

ACM Reference Format:
Shubham Sharma, Yunfeng Zhang, Jesús M. Ríos Aliaga, Djallel Bouneffouf, Vinod Muthusamy, and Kush R. Varshney. 2020. Data Augmentation for Discrimination Prevention and Bias Disambiguation. In *Proceedings of the 2020 AAAI/ACM Conference on AI, Ethics, and Society (AIES '20), February 7–8, 2020, New York, NY, USA.* ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/3375627.3375865

1 INTRODUCTION

Fairness in machine learning has been a growing and interesting field of research. The existence of unwanted discrimination by machine learning models (e.g. demonstrated in [5]), has led to the development of measures that can be used to quantify such bias, along with techniques to combat such biases. The definitions used to understand the bias in such models can be broadly categorized into three types: *independence*, *separation* and *sufficiency*. Specifically, a classifier satisfies independence if the protected attribute (such as race or gender) for which the model may be biased is independent of the classifier decision. Separation is satisfied if the classifier decision is independent of the protected attribute conditioned on the true label. Sufficiency is satisfied if the true label is independent of the protected attribute conditioned on the classifier prediction. It has been shown that unless ideal conditions are met, the three definitions are mutually incompatible [12]. Details on these fairness criteria, both mathematically and with respect to different worldviews, may be found in [2, 25] along with definitions of associated fairness metrics (such as statistical parity difference for independence, average odds difference for separation, and calibration for sufficiency).

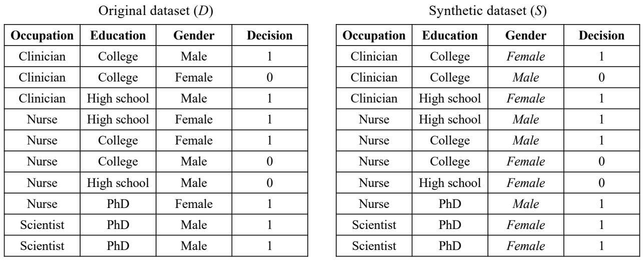
Using these definitions, several techniques have been proposed that help satisfy the various criteria of fairness for machine learning models to reduce bias, and such techniques can be broadly categorized into: pre-processing, in-processing and post-processing [8, 21]. In this paper, we focus on a pre-processing technique that aims to modify the dataset that is being used to train the model, to minimize the bias existing in the model. A simple pre-processing technique could be fairness through unawareness, where a protected attribute is simply not considered for training a model. However, this approach might still result in biases through other features that may be correlated with the protected attribute [14].

Previous methods that pre-process data include modifying the dataset by sampling or reweighing the training samples [17], changing individual data records [15], using t-closeness [22], optimizing using a multi-objective loss function [6], adversarial debiasing [27], fairness GAN [23], optimized score transformation [24], removing disparate impact [11], and learning fair representations [26]. However, many of these are either very complicated or not intuitive to understand, especially for a policy maker with limited knowledge of machine learning. We introduce a simple and effective technique that tries to minimize the bias by augmenting the dataset with synthetic points, such that the overall dataset represents a more equitable world, where the level of augmentation can be decided by domain experts or policy makers.

Apart from simply measuring bias, it is also important to understand the reason for the bias that arises in the data and hence the model, to be able to facilitate better dataset collection. Many types of biases can arise in a model [2, 3], and some effort has been undertaken towards understanding the reason behind them, both from an optimization perspective [7] and a causal inference perspective [20]. In this paper, we consider two out of the many possible types of bias 1) Prejudice-based bias: Also known as label bias, this exists when human biases affect the labels and 2) Sample bias: This type of bias may exist when a certain group is under-represented in the dataset, because of a non-uniform data collection strategy. We show experimentally how the data augmentation technique can lend itself towards answering if the bias arises because of prejudice or sample bias.

The data augmentation technique can be described as follows: given a dataset that contains a protected attribute (such as gender or race), we define an "ideal world dataset" as data where different groups within the protected attribute (such as male or female for gender) attain the same label, irrespective of other feature values.[1]

To demonstrate how such an ideal dataset would be created and useful, consider a hypothetical example, shown in Fig. 1, of the hiring decisions of applicants for a job posting, where the protected attribute is the applicant's gender. While this dataset is hypothetical, it is evident that since personal bias against certain protected attributes have been demonstrated from a psychological perspective [16], such datasets would exist in the real world. This dataset demonstrates examples of both prejudice and sample bias. An example of prejudice bias is that male and female clinicians with a college degree are treated differently. The same applies for male versus female nurses, both as a whole and when conditioned on the high school and college graduate sub-populations. An example of sample bias results from the absence of data on female scientists. Likewise, there is no data on male nurses with a PhD, nor female clinicians with a high school diploma. A model trained on this dataset would potentially be biased.



Original dataset (D)

Occupation	Education	Gender	Decision
Clinician	College	Male	1
Clinician	College	Female	0
Clinician	High school	Male	1
Nurse	High school	Female	1
Nurse	College	Female	1
Nurse	College	Male	0
Nurse	High school	Male	0
Nurse	PhD	Female	1
Scientist	PhD	Male	1
Scientist	PhD	Male	1

Synthetic dataset (S)

Occupation	Education	Gender	Decision
Clinician	College	*Female*	1
Clinician	College	*Male*	0
Clinician	High school	*Female*	1
Nurse	High school	*Male*	1
Nurse	College	*Male*	1
Nurse	College	*Female*	0
Nurse	High school	*Female*	0
Nurse	PhD	*Male*	1
Scientist	PhD	*Female*	1
Scientist	PhD	*Female*	1

Figure 1: Example with original dataset D and synthetic dataset S. Together they represent the final ideal dataset D^*.

To address these issues, we add synthetic points with different values for the protected gender attribute, as shown in Fig. 1. These synthetic points along with the original ones constitute the overall ideal dataset. This new dataset now has an equal number of males and females, and the label no longer depends on the gender, thereby potentially removing bias from the model that this overall dataset is trained on.

To minimize concerns over "polluting" the dataset with many synthetic points, we propose a data augmentation technique that selectively adds only a subset of the synthetic points to meet the fairness criteria while maintaining accuracy. The approach is to successively add a set of synthetic data points to create new augmented datasets, and then evaluate the model on the new datasets. We show that by augmenting data using this technique, it reduces bias based on two fairness definitions: statistical parity difference and average odds difference, while keeping the accuracy nearly constant. Theoretically, we show that the addition of any point to the dataset would result in the simultaneous increase or decrease of both statistical parity difference (independence) and average odds difference (separation). Other pre-processing techniques cannot make this guarantee. Secondly, we theoretically show that if the data is added in a greedy way to only reduce bias (and hence only favorable decision samples for the unprivileged group and unfavorable decision samples for the privileged group are considered for augmentation), the fairness definitions are submodular and hence we can greedily augment the dataset to reduce bias, while being computationally efficient.

We perform experiments on successive subsets of data for three datasets by sorting the synthetic points by a measure of how realistic they are with respect to the original data and then adding a k-percentage of these sorted synthetic points to the real data to create a k-augmented dataset. The realism of a synthetic data point is measured by finding cluster centers (using k-means) from the original data and measuring the distance of every synthetic point to these cluster centers. For example, a 1-augmented dataset would mean that 1% of the most realistic data points from the synthetic dataset have been added to the original dataset, while a 100-augmented dataset would be what we define as the ideal world dataset.

[1] Note that if other features exist that are directly representative of the protected attribute, and not a manifestation of indirect correlations, then these should also be changed to ensure that the ideal world dataset is feasible. In this paper, the datasets considered do not exhibit this property.

The reason for performing extensive experimentation on these subsets instead of just using the ideal world dataset (100-augmented dataset) is two fold. First, while such an ideal world dataset might intuitively make sense, it could raise several concerns with statisticians and policy makers. Chief among these concerns would be the thought of drastically manipulating the real-world data distribution (original data) such that the model that is trained on this new data may be highly inaccurate, or that the new model does not reflect how humans perceive the world. In this regard, an "ideal world" dataset from a fairness perspective would not be considered "ideal" from a statistical perspective: real-data distributions are being altered. k-augmented datasets offer model developers a hyper-parameter to decide the extent to which they can use the augmentation. We show in our experiments that accuracy is not significantly compromised, while bias is extensively reduced using such an augmentation scheme.

Secondly, we show that data augmentation has a surprising side-effect of potentially helping discover the type of bias existing in the data. If a small subset of the most realistic points being added results in a significant decrease in bias, a sampling bias may exist in the data, since collecting more realistic data could potentially fix the bias. If, however, adding a large set of synthetic points does not cause the bias to reduce, this could be indicative of prejudice based bias existing in the original data, since the points that lead to the decrease in bias are unrealistic with respect to the original dataset, and hence the unprivileged group has possibly been discriminated against by being assigned more unfavorable labels. To the best of our knowledge, this is the first such simple data augmentation strategy that mitigates bias while also suggesting the type of bias in the data.

The contributions of our work can be summarized as follows:

- A new data augmentation technique to mitigate bias and uncover the type of bias in the data.
- Theoretical guarantee on coincident change of two different fairness measures with this augmentation scheme.
- Proof of submodularity of fairness metrics under conditions.
- Discussion on the evaluation of pre-processing based techniques.

2 THEORY

Consider an original dataset D composed of features X, a binary protected attribute A, and a binary label Y that is used to train a machine learning model M for a classification problem.

Definition 1: For a dataset D with a binary protected attribute A, an ideal world dataset D^* is such that $D^* = D \cup S$, where S represents a set of synthetic points added such that $S_X = D_X, S_Y = D_Y$ and $S_A \neq D_A$.

That is, the ideal dataset contains the original dataset combined with the set of points (synthetic data) that have the same features X and label Y, but a different protected attribute A.

Lemma 1: A model trained on the ideal dataset with perfect accuracy will satisfy independence, separation, calibration, and counterfactual fairness [19].

The mathematical definitions and proofs for this lemma are given in the appendix[2], but it is intuitive to expect so since both independence and conditional independence in all forms would be satisfied for a classifier with perfect accuracy.

In this paper, we measure independence using statistical parity difference:

$$SPD = P(Y' = 1 \mid A = 0) - P(Y' = 1 \mid A = 1). \tag{1}$$

We measure separation using the average odds difference, i.e., the sum of the differences between the true positive and false positive rates between groups:

$$\begin{aligned} AOD =& 0.5(P(Y' = 1 \mid Y = 1, A = 0) \\ & - P(Y' = 1 \mid Y = 1, A = 1) \\ & + P(Y' = 1 \mid Y = 0, A = 0) \\ & - P(Y' = 1 \mid Y = 0, A = 1)) \end{aligned} \tag{2}$$

Where $A = 0$ represents an unprivileged group and $A = 1$ represents a privileged group, $Y = 0$ is an unfavorable outcome and $Y = 1$ is a favorable outcome and Y' represents the model prediction.

2.1 Creating multiple synthetic datasets

Creating such an ideal dataset D^* raises concerns in real-world deployment since the following questions arise: 1) Fabricating training points leads to a training distribution that does not represent the real world. Would a policy maker be comfortable with this? 2) Would creating such training points lead to a loss in accuracy? Hence, we do not pose this method as an automatic pre-processing step that can be carried out without human intervention. Instead, we augment the original dataset D in increments such that at every step k, we select the top $k\%$ points from S and create a new dataset D_k such that the $k\%$ points are the most realistic points with respect to the original training distribution, from S. Hence $D_0 = D$ and $D_{100} = D^*$. A policy maker can then see the effect of training the model on every D_k datasets on the accuracy and fairness metrics. Showing such an analysis is helpful to gain people's trust in the augmentation technique, select a possible level of augmentation, as well as to figure out the type of bias in the dataset, as shown in the experiments.

To sort the synthetic set S in order of most to least realistic data points, we use k-means on the original dataset D, such that a set of cluster centers are defined for every protected attribute value and every label value (i.e. for every combination of A and Y), and the inverse of the maximum distance to any cluster center for a point in the synthetic dataset having the same A and Y values is used as a score of how realistic the point is.

More formally, for every $A = a$ and $Y = y$ combination in D, we find l cluster centers $\{c_1, c_2, \ldots, c_l\}$ and then assign a realism score to a synthetic datapoint p from S having $A = a$ and $Y = y$ as:

$$Realism(p) = \frac{1}{max\{d(c_1, p), d(c_2, p), \ldots, d(c_l, p)\}} \tag{3}$$

Then, based on the score assigned to every datapoint, S is sorted, and to build an intermediate dataset D_k, the top $k\%$ points from

[2] The appendix can be found at: https://drive.google.com/open?id=1ZrEcNbH52801bLjyVMcWH-40PLas11du

this sorted synthetic dataset are selected to augment D. Note than any other measure of realism can also be considered (such as the distance between data distributions).

2.2 Effect of augmentation on fairness definitions

Theorem 1: If the augmentation of a dataset D by a point p leads to an increase in the statistical parity difference for a classifier trained on $D \cup p$, it would lead to an increase in the average odds difference, and vice versa. Conversely, if the statistical parity difference decreases for the classifier trained on the dataset $D \cup p$, the average odds difference will also decrease, and vice versa. That is, when pre-processing using data augmentation, independence and separation measures would increase or decrease simultaneously.

The proof for the theorem is given in the Appendix. This theorem is crucial, since using such a data augmentation technique does not require considering two types of fairness definitions (independence and separation): reducing one measure will always reduce the other, and hence from a pre-processing perspective, the two definitions are coincident and a policy maker who may be unfamiliar with the variety of fairness definitions would only need to be concerned about one of them without worrying about the other, something that other pre-processing techniques do not guarantee.

2.3 On greedy fairness-aware data augmentation

For certain critical applications such as criminal justice, a policy maker might be concerned with purely minimizing the bias in their dataset by adding a minimum subset of points from the synthetic set S. This case of greedy subset selection for bias reduction comes at a risk of being computationally expensive, since every possible subset would have to be considered and the model would have to be trained for each possible subset. We prove, however, that for SPD the function is submodular for the purpose of bias reduction, and hence we have guarantees on the computational complexity of this approach, allowing a feasible solution. The proof trivially extends to AOD as well.

Theorem 2: Let us denote the finite set of data points from S such that they belong to the favorable unprivileged or unfavorable privileged groups as α. If Ω is the set of all subsets that always contains D along with a random set from α, i.e., $\Omega = \{\omega : \omega = \{D, \alpha'\} s.t. \alpha' \subseteq \alpha\}$ and let $f = SPD, f : 2^{\Omega} - > \mathbf{R}$, then f is submodular.

The proof is given in the Appendix. This allows model developers to greedily find the minimum set of data points that can allow the overall model to be less biased significantly faster if it is important to consider a minimum subset, since submodular functions can lead to faster computation than pure brute force methods [13].

3 EXPERIMENTS

To perform experiments on the proposed method, three datasets are considered: UCI Adult dataset [18], COMPAS dataset [1] and the German Credit dataset [9]. Each of these datasets contains at least one protected attribute. All three datasets have been studied for demonstrating bias with respect to at least one protected attribute, specifically in [4] and the associated open-source notebooks, and we consider the same versions of the datasets and logistic regression models as is available in their open-source implementation. For the UCI Adult dataset, we consider both the gender and race protected attributes individually, where the race protected attribute is considered to demonstrate bias disambiguation. We consider the race protected attribute for the COMPAS dataset and the age protected attribute for the German Credit dataset. Details on datasets and for sorting the synthetic dataset using cluster centers are given in the appendix.

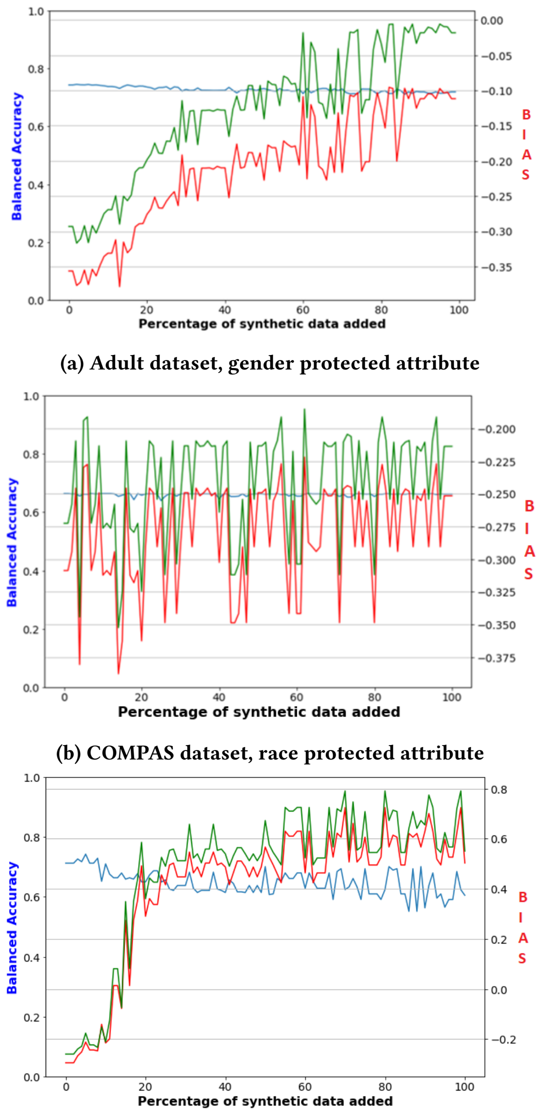
3.1 Bias mitigation

The plots for the three datasets are shown in Fig. 2. The x-axis represents the percentage of the most realistic points being added, where we train the model from scratch for every 1% increment. Hence, 0 represents the original model, 100 represents all real and synthetic points considered together (“ideal world dataset”), and as an example, the 32% mark represents a model trained with all real points and 32% of the most realistic points from the synthetic dataset. The two-sided y-axis represents a measure of balanced accuracy on the left, and fairness measure on the right. The blue line in each plot corresponds to accuracy, while the red and green lines correspond to AOD and SPD, respectively.

As we can see, the SPD and AOD follow a very similar plot with each percentage of synthetic data added. consistent with our theoretical proof, while accuracy decreases slightly for all three datasets. Note that the accuracy and bias metrics in these figures are measured by averaging over multiple runs by considering a holdout test set for one run and changing it for different runs. This holdout set is from the unaugmented data only and is not used to train any of the models. We discuss other ways of evaluating fairness based on different types of test sets in Section 3.4. A value of 0 for both AOD and SPD measures is optimal, indicating no bias. For a 100% accurate classifier, the value of these measures would always be 0 at the 100% x-axis mark. However, since the classifiers considered here are not perfect, the 100% mark isn’t fairness-optimal. We find that while for the German and Adult datasets, bias decreases and approaches near optimum with the addition of synthetic data points, the bias does not decrease significantly for the COMPAS dataset. To follow a pre-processing technique that can effectively reduce bias for this dataset, we consider the COMPAS dataset and add only favorable unprivileged and unfavorable privileged samples (as opposed to any type of synthetic data being added) from the synthetic dataset (and recommend model developers to do this for fairness critical applications). The result of doing this is shown in Fig. 3. As we can see, the bias is now significantly reduced without, once again, significantly compromising on accuracy.

3.2 Comparison to other pre-processing technqiues

We compare the proposed approach with other pre-processing techniques in Table 1. To run this analysis, we use the AIF360 [4] notebook implementations for optimized pre-processing, reweighing and adversarial debiasing. We report the accuracy, SPD and AOD for the percentage of augmentation which gives the least bias. For the German and Adult datasets, since just adding a subset of points in the standard fashion significantly reduces bias, we report



(a) Adult dataset, gender protected attribute

(b) COMPAS dataset, race protected attribute

(c) German Credit Dataset, age protected attribute

Figure 2: Plots for bias mitigation, where the x-axis represents the percentage of synthetic data added, the left side y-axis and corresponding blue line is the accuracy value, and the right side y-axis is the bias measure value where the green line represents AOD (Eqn. 2) and the red line represents SPD (Eqn. 1).

Table 1: Comparison of our data augmentation based bias disambiguation to other pre-processing techniques. For each dataset, we report the result corresponding to the best possible augmentation (maximum bias reduction) for two cases: one on a test set from the original unprocessed data called DataAug, and the second on a test set taken from the processed (augmented data), called DataAugp (Section 3.4). *For the COMPAS dataset, standard augmentation does not render a significant reduction in bias (Figure 2b), we consider the best result using only favorable unprivileged and unfavorable privileged sample augmentation (Figure 3).

	Method	Acc.	SPD	AOD
Adult	Raw	0.74	-0.36	-0.33
	Opt. Preprocessing	0.68	-0.01	-0.05
	Reweighing	0.71	-0.09	-0.03
	Adv. Debiasing	0.67	-0.08	-0.04
	DataAug	**0.70**	**-0.09**	**-0.01**
	DataAugp	**0.73**	**-0.02**	**0**
German	Raw	0.71	-0.32	-0.33
	Opt. Preprocessing	0.57	-0.65	-0.63
	Reweighing	0.66	-0.27	-0.29
	Adv. Debiasing	0.58	0.18	0.22
	DataAug	**0.65**	**0**	**0**
	DataAugp	**0.70**	**0**	**0**
COMPAS*	Raw	0.68	-0.3125	-0.27
	Opt. Preprocessing	0.67	-0.09	-0.05
	Reweighing	0.63	0.05	0.10
	Adv. Debiasing	0.64	0.03	0.08
	DataAug	**0.67**	**0**	**0.05**
	DataAugp	**0.67**	**0**	**0**

numbers concurrent to Fig. 2c and Fig. 2a, respectively. However, since we notice that the COMPAS dataset does not have a significant decrease using the standard method (the reasons for which is discussed in the next section), we consider just adding the favorable unprivileged and unfavorable privileged samples from the synthetic set to mitigate the bias, as in Fig. 3, and report the best bias reduction measures (least bias) and corresponding accuracy for that approach. DataAug corresponds to a holdout test set from the original unproccesed data, and DataAugp corresponds to a holdout test set from the augmented data. We discuss why we report both in the subsection on evaluating pre-processing techniques.

As we can see, our method significantly reduces both SPD and AOD, while keeping the accuracy nearly the same. Compared to other methods, our approach towards adding such points to retrain models is easier to understand, especially for policy makers, compared to methods such as optimized pre-processing (does not work well with a smaller dataset like German Credit) and adversarial debiasing (complicated implementation). While reweighing is also easy to understand and implement, it involves weighting existing points and does not perform well on particular datasets, as can be seen for the German credit dataset in Table 1. While we report the best case results assuming that a model developer is comfortable with any level of augmentation, a model developer gets an understanding of the extent to which the data is augmented to reduce bias using the plots discussed before, and hence they can select the extent to which modification is acceptable, thereby selecting a level of augmentation that may not hinder accuracy significantly if that is more critical. In fact, we can see how the extent of augmentation can vary for different datasets: for the German credit dataset,

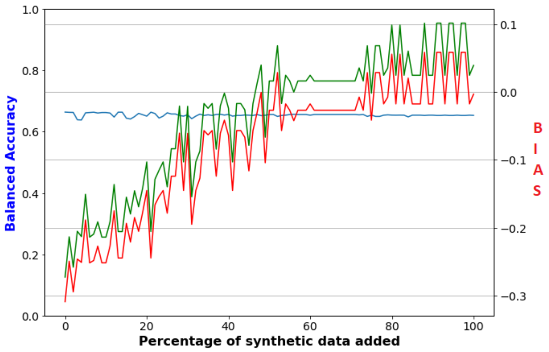


Figure 3: COMPAS dataset (race protected attribute), positive unprivileged and negative privileged examples added

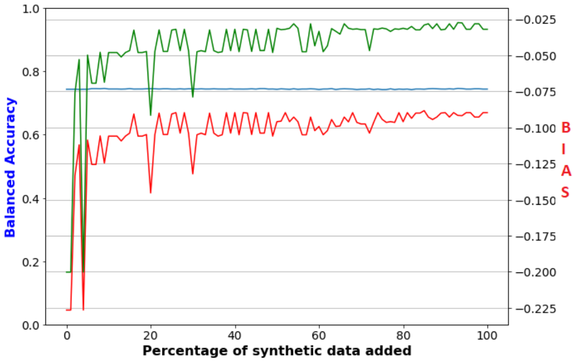


Figure 4: Adult dataset for the race protected attribute

adding just around 17% most realistic points from the synthetic dataset reduces bias to 0 and this is not true for other datasets.

3.3 Bias disambiguation

We believe that this method can also lend itself to analysing the type and extent of bias existing within the dataset. When synthetic points are added from most realistic to least realistic in the way that we have done, if the addition of a few percentage of synthetic points causes a significant decrease in bias, this could be indicative of just a sampling based bias, since collecting more realistic data that adheres to a similar distribution would have resulted in less discrimination. If however, the bias does not decrease significantly with a smaller subset of realistic synthetic points being added, or decreases predominantly with more unrealistic synthetic points, this could be indicative of prejudice, since collecting more points in a similar environment would not have yielded any improvements.

To demonstrate this, consider the COMPAS dataset examples in Fig. 2 and Fig. 3 and consider the Adult dataset example in Fig. 2 and in Fig. 4. In the original plot for the COMPAS dataset Fig. 2, the bias hardly decreases with the addition of synthetic points. Even for Fig. 3, the bias decreases slowly with the addition of points from most to least realistic. The COMPAS dataset has been known to be influenced from prejudice based bias [3], and this is evident from these plots, since standard augmentation did not help, and adding only favorable unprivileged and unfavorable privileged points also resulted in a slow decrease with augmentation. In contrast, consider the UCI Adult dataset with race as the protected attribute, shown in Fig. 4, or even the German Credit dataset in Fig. 2c. As we can see, the addition of around 20% of the most realistic synthetic points causes the bias to significantly reduce. Hence, collecting more data for the unprivileged group would have potentially fixed biases arising in the model, since adding a few realistic points helped in reducing bias. Through these experiments, we motivate the use of our method for disambiguating bias, potentially even at the data collection stage, to create an evenly distributed dataset.

3.4 Evaluating pre-processing techniques

While all pre-processing techniques consider accuracy in the way we have for the plots and the DataAug metrics (where a holdout test set from the unprocessed data is used to evaluate the accuracy and other measures), there are multiple ways in which accuracy can be calculated, and this should be an important consideration when reporting evaluations. Since the goal here is to remove the bias in the model by pre-processing the dataset, a natural question arises: should accuracy be evaluated using a random set from the processed or unprocessed data?

While the first thought would be to consider test points from unprocessed data only (since these represent true data points and this has been followed in the literature as well), it is worth noting that the model is now trained on a modified distribution of data, where the modified distribution is representative of a "fairer world" [10]. However, the original dataset remains biased. Consider a simple case where the original dataset has a sampling bias (lack of data for a particular group *G*). If we use any form of pre-processing, that should account for such a bias. However, when considering any random test set, that would more likely contain fewer points belonging to group *G*, since the original dataset had fewer samples for this group to begin with. Hence even though the model is relatively unbiased, that bias would still show in the evaluation, because of a bias in the test set which is drawn from the real biased dataset (data vs. model bias).

To address such concerns, we recommend two ways in which accuracy should also be reported to understand the effect of the pre-processing technique under consideration: 1) Report accuracy on the entire processed dataset along with accuracy on a holdout test set from the real data. We recommend reporting both since both evaluations have their merits, as was discussed above. 2) Create a curated test set from the real dataset. If the real dataset contains sampling bias, we can potentially create test sets that do not contain a disparity in samples to evaluate the model. We report the measures for the three datasets by considering the test set from the processed dataset in Table 1 as **DataAugp**, and we encourage researchers to report both such measures in future work.

4 CONCLUSION AND DISCUSSION

In this paper, we present a simple data augmentation technique for bias reduction and bias disambiguation. We show how the approach guarantees a simultaneous increase or decrease in two different notions of fairness (independence and separation), and show that these definitions are submodular for bias mitigation and hence a minimum subset for augmentation can be chosen efficiently. Experiments are performed on three datasets and we see that our method performs better than previous approaches to mitigate bias while keeping accuracy nearly similar. We also show experimentally that this method can be used to decipher the type of bias in the data.

A possible concern may be that successive augmentations and training a model might be computationally expensive if the dataset contains many features. For such cases, we simply recommend augmenting with only favorable unprivileged and unfavorable privileged samples if the application is fairness critical. We intend to now conduct a user-study to show such plots and understand the extent to which model developers and policy makers are comfortable with such an augmentation.

REFERENCES

[1] [n. d.]. ProPublica COMPAS. https://www.propublica.org/datastore/dataset/compas-recidivism-risk-score-data-and-analysis.

[2] Solon Barocas, Moritz Hardt, and Arvind Narayanan. 2019. *Fairness and Machine Learning*. fairmlbook.org. http://www.fairmlbook.org.

[3] Solon Barocas and Andrew D Selbst. 2016. Big data's disparate impact. *Calif. L. Rev.* 104 (2016), 671.

[4] Rachel K. E. Bellamy, Kuntal Dey, Michael Hind, Samuel C. Hoffman, Stephanie Houde, Kalapriya Kannan, Pranay Lohia, Jacquelyn Martino, Sameep Mehta, Aleksandra Mojsilovic, Seema Nagar, Karthikeyan Natesan Ramamurthy, John Richards, Diptikalyan Saha, Prasanna Sattigeri, Moninder Singh, Kush R. Varshney, and Yunfeng Zhang. 2019. AI Fairness 360: An Extensible Toolkit for Detecting and Mitigating Algorithmic Bias. *IBM Journal of Research and Development* 63, 4/5 (2019).

[5] Toon Calders and Indrė Žliobaitė. 2013. Why unbiased computational processes can lead to discriminative decision procedures. In *Discrimination and privacy in the information society*. Springer, 43–57.

[6] Flavio Calmon, Dennis Wei, Bhanukiran Vinzamuri, Karthikeyan Natesan Ramamurthy, and Kush R Varshney. 2017. Optimized Pre-Processing for Discrimination Prevention. In *Advances in Neural Information Processing Systems 30*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (Eds.). Curran Associates, Inc., 3992–4001. http://papers.nips.cc/paper/6988-optimized-pre-processing-for-discrimination-prevention.pdf

[7] Irene Chen, Fredrik D Johansson, and David Sontag. 2018. Why is my classifier discriminatory?. In *Advances in Neural Information Processing Systems*. 3539–3550.

[8] Brian d'Alessandro, Cathy O'Neil, and Tom LaGatta. 2017. Conscientious classification: A data scientist's guide to discrimination-aware classification. *Big data* 5, 2 (2017), 120–134.

[9] Dheeru Dua and Casey Graff. 2017. UCI Machine Learning Repository. http://archive.ics.uci.edu/ml

[10] Sanghamitra Dutta, Dennis Wei, Hazar Yueksel, Pin-Yu Chen, Sijia Liu, and Kush R Varshney. 2019. An Information-Theoretic Perspective on the Relationship Between Fairness and Accuracy. *arXiv preprint arXiv:1910.07870* (2019).

[11] 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*. ACM, 259–268.

[12] Sorelle A Friedler, Carlos Scheidegger, and Suresh Venkatasubramanian. 2016. On the (im) possibility of fairness. *arXiv preprint arXiv:1609.07236* (2016).

[13] Satoru Fujishige. 2005. *Submodular functions and optimization*. Vol. 58. Elsevier.

[14] Pratik Gajane and Mykola Pechenizkiy. 2017. On formalizing fairness in prediction with machine learning. *arXiv preprint arXiv:1710.03184* (2017).

[15] Sara Hajian and Josep Domingo-Ferrer. 2012. A methodology for direct and indirect discrimination prevention in data mining. *IEEE transactions on knowledge and data engineering* 25, 7 (2012), 1445–1459.

[16] Carol Isaac, Barbara Lee, and Molly Carnes. 2009. Interventions That Affect Gender Bias in Hiring: A Systematic Review. *Academic medicine : journal of the Association of American Medical Colleges* 84 (10 2009), 1440–6. https://doi.org/10.1097/ACM.0b013e3181b6ba00

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

[18] Ron Kohavi. 1996. Scaling up the accuracy of naive-bayes classifiers: A decision-tree hybrid.. In *Kdd*, Vol. 96. Citeseer, 202–207.

[19] Matt J Kusner, Joshua Loftus, Chris Russell, and Ricardo Silva. 2017. Counterfactual fairness. In *Advances in Neural Information Processing Systems*. 4066–4076.

[20] David Madras, Elliot Creager, Toniann Pitassi, and Richard Zemel. 2019. Fairness through causal awareness: Learning causal latent-variable models for biased data. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*. ACM, 349–358.

[21] Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. 2019. A survey on bias and fairness in machine learning. *arXiv preprint arXiv:1908.09635* (2019).

[22] Salvatore Ruggieri. 2014. Using t-closeness anonymity to control for non-discrimination. *Trans. Data Privacy* 7, 2 (2014), 99–129.

[23] Prasanna Sattigeri, Samuel C. Hoffman, Vijil Chenthamarakshan, and Kush R. Varshney. 2019. Fairness GAN: Generating Datasets with Fairness Properties using a Generative Adversarial Network. *IBM Journal of Research and Development* 63, 4/5 (2019).

[24] Dennis Wei, Karthikeyan Natesan Ramamurthy, and Flavio du Pin Calmon. 2019. Optimized Score Transformation for Fair Classification. *arXiv preprint arXiv:1906.00066* (2019).

[25] Samuel Yeom and Michael Carl Tschantz. 2018. Discriminative but Not Discriminatory: A Comparison of Fairness Definitions under Different Worldviews. *arXiv preprint arXiv:1808.08619* (2018).

[26] Rich Zemel, Yu Wu, Kevin Swersky, Toni Pitassi, and Cynthia Dwork. 2013. Learning fair representations. In *International Conference on Machine Learning*. 325–333.

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