Accepted at the ACM International Conference on Information and Knowledge Management (CIKM) 2022

Cascaded Debiasing : Studying the Cumulative Effect of Multiple Fairness-Enhancing Interventions

Bhavya Ghai
Stony Brook University
New York, USA
bghai@cs.stonybrook.edu

Mihir Mishra
Los Altos High School
California, USA
mihir.mishra@gmail.com

Klaus Mueller
Stony Brook University
New York, USA
mueller@cs.stonybrook.edu

ABSTRACT

Understanding the cumulative effect of multiple fairness enhancing interventions at different stages of the machine learning (ML) pipeline is a critical and underexplored facet of the fairness literature. Such knowledge can be valuable to data scientists/ML practitioners in designing fair ML pipelines. This paper takes the first step in exploring this area by undertaking an extensive empirical study comprising 60 combinations of interventions, 9 fairness metrics, 2 utility metrics (Accuracy and F1 Score) across 4 benchmark datasets. We quantitatively analyze the experimental data to measure the impact of multiple interventions on fairness, utility and population groups. We found that applying multiple interventions results in better fairness and lower utility than individual interventions on aggregate. However, adding more interventions do no always result in better fairness or worse utility. The likelihood of achieving high performance (F1 Score) along with high fairness increases with larger number of interventions. On the downside, we found that fairness-enhancing interventions can negatively impact different population groups, especially the privileged group. This study highlights the need for new fairness metrics that account for the impact on different population groups apart from just the disparity between groups. Lastly, we offer a list of combinations of interventions that perform best for different fairness and utility metrics to aid the design of fair ML pipelines.

CCS CONCEPTS

• **Computing methodologies → Machine learning**.

KEYWORDS

fairness, debiasing, machine learning, fair ML pipeline

1 INTRODUCTION

Algorithmic bias is a complex socio-technical problem whose impact can be felt in all sub-disciplines of machine learning [16, 20–22, 24, 35]. Recent years have seen a huge surge of fairness enhancing interventions that operate at different stages of the ML pipeline. Some of these interventions are more effective than others at reducing bias as captured by a specific fairness metric. However, the problem is far from being solved if that is even possible [19]. Hence, there is a need for better interventions to reduce bias even further. Moreover, Algorithmic bias can virtually emerge from any single or multiple stage(s) of the machine learning pipeline, right from problem formulation, dataset selection/creation to model formulation, deployment, and so on. [25]. The existing literature focuses on curbing algorithmic bias by intervening at *a* particular stage of the ML pipeline (see Figure 1). However, algorithmic bias might still flourish via other stages/components of the ML pipeline. So, our focus should be on ensuring fairness across the ML pipeline instead of a single stage of the pipeline. This issue is also backed by a recent study with ML practitioners that elaborated on the disconnect between academic research and real world needs [25]. One of the findings was to consider fairness as a system level property where the focus is on evaluating the impact of ML system as a whole instead of monitoring individual components.

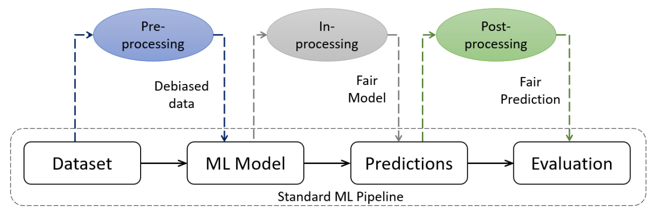

Figure 1: Three different types of fairness enhancing interventions and how they fit into the standard ML pipeline.

An intuitive solution to enhance fairness across the ML pipeline can be to apply multiple fixes (interventions) at different stages of the ML pipeline where bias can emerge from. We will refer to such a series of fairness enhancing interventions as cascaded interventions. For example, one might choose to debias the dataset, train a fairness aware classifier over it and then post-process the model's predictions to achieve more fairness. This approach is inline with the real world where different laws/policies/guidelines try to alleviate social inequality by intervening at multiple stages of life like education, employment, promotion, etc. Examples include Affirmative action in the US and Caste based reservation in India. This begs the question if it were possible to achieve more fairness in the ML world by intervening at multiple different stages of the ML pipeline and what might be its possible fallouts.

In this work, we undertake an extensive empirical study to understand the impact of individual interventions as well as the cumulative impact of cascaded interventions on utility metrics like accuracy, different fairness metrics and on the privileged/unprivileged groups. Here, we have focused on the binary classification problem over tabular datasets with a single sensitive (protected) attribute. We have considered 9 different interventions where 2 operate at the data stage, 4 operate at the modeling stage and 3 operate at the post modeling stage. We also consider all possible combinations of these interventions as shown in Figure 3. To execute multiple interventions in conjunction, we feed the output of one intervention as input to the next stage of the ML pipeline. We simulate multiple 3 stage ML pipelines that are acted upon by different combinations

Conference'17, July 2017, Washington, DC, USA
2022. ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

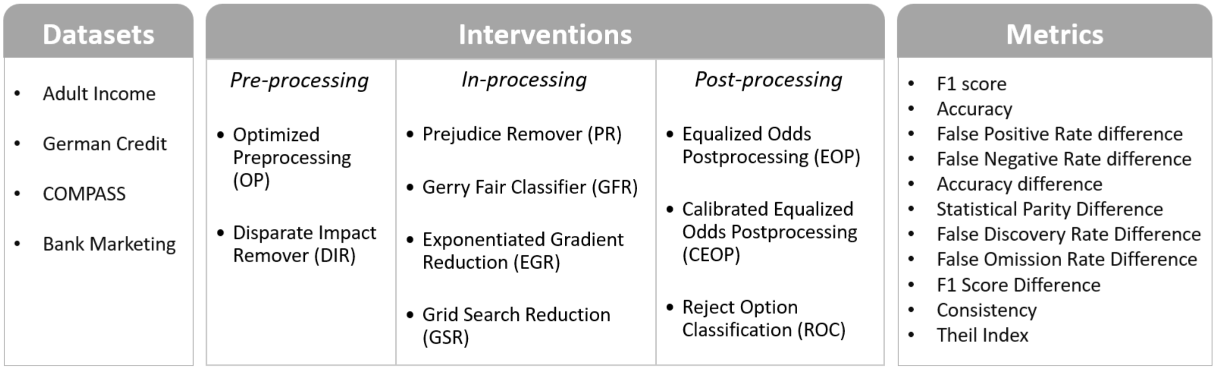

Figure 2: Experimental Setup - Different datasets, interventions and metrics considered for the empirical study

of interventions. We measure the impact of all these interventions on 9 fairness metrics and 2 utility metrics over 4 different datasets. Thereafter, we perform quantitative analyses on the results and try to answer the following research questions:

- **R1.** Effect of Cascaded Interventions on Fairness metrics
  Does intervening at multiple stages reduce bias even further? If so, does it always hold true? What is the impact on Group fairness metrics and Individual fairness metrics?
- **R2.** Effect of Cascaded Interventions on Utility metrics
  How do utility metrics like accuracy and F1 score vary with different number of interventions? Existing literature discusses the presence of Fairness Utility tradeoff for individual interventions. Does it hold true for cascaded interventions?
- **R3.** Impact of Cascaded Interventions on Population Groups
  How are the privileged and unprivileged groups impacted by cascaded interventions in terms of F1 score, False negative rate, etc.? Are there any negative impacts on either groups?
- **R4.** How do different cascaded interventions compare on fairness and utility metrics?

2 BACKGROUND AND LITERATURE REVIEW

2.1 Fairness Enhancing Interventions

Bias Mitigation techniques can be broadly classified into 3 stages :- Pre-processing, In processing and Post-processing (Fig. 1). In the following, we discuss a few interventions that we have considered in this work, in the context of the intervention stage they operate.

2.1.1 *Pre-processing.*

Interventions at the Pre-processing stage operate on the raw dataset to generate its debiased version. The debiased dataset can then be fed back into the standard ML pipeline for fairer predictions. Specifically:

Optimized Preprocessing (OP) — uses convex optimization to transform the underlying dataset such that fairness is enhanced and utility is preserved with limited data distortion [10].

Disparate Impact Remover (DIR) — edits the feature set of a given dataset such that the predictability of the protected variable is impossible while preserving rank ordering within groups [18].

2.1.2 *In-processing.*

Interventions in this stage operate at the data modeling stage to train a fair ML model. Specifically:

Gerry Fair Classifier (GFC) — formulates fairness as a zero-sum game between a Learner (the primal player) and an Auditor (the dual player) to compute an equilibrium for this game [28].

Prejudice Remover (PR) — adds a specialized regularization term to the learning objective such that the classifier becomes independent of the sensitive information [27].

Exponential Gradient Reduction (EGR) — reduces fair classification to a sequence of cost-sensitive classification problems, returning a *randomized classifier* with the lowest empirical error subject to fair classification constraints [1].

Grid Search Reduction (GSR) — reduces fair classification to a sequence of cost-sensitive classification problems, returning the *deterministic classifier* with the lowest empirical error subject to fair classification constraints [1, 2].

2.1.3 *Post-processing.*

Such interventions operate at the model's predictions to yield more fair predictions. Specifically:

Calibrated Equalized Odds Postprocessing (CEOP) — changes classifier results based on calibrated score outputs and an equalized odds goal [37].

Equalized Odds Postprocessing (EOP) — solves a linear program to find probabilities whose corresponding labels will optimize the equalized odds goal [23, 37].

Reject Option Classification (ROC) — reduces discrimination by assigning positive labels to the unprivileged groups and negative labels to the privileged groups for the data points that lie close to the the decision boundary i.e. the points for which the classifier is uncertain about [26].

Existing literature has studied the above mentioned interventions in isolation. In this work, we explore if a combination of these interventions can lead to enhanced fairness across the ML pipeline.

2.2 Measuring Fairness

Quantifying fairness in Algorithms is an active research area. Numerous fairness metrics have been proposed in the literature which mathematically encode different facets of fairness like Group fairness, Individual fairness, Counterfactual fairness, etc. [4, 5, 11, 13, 14, 30, 36]. For eg., Group fairness implies that members of one group should receive a similar proportion of positive/negative outcomes as other groups [4, 11], Individual fairness implies that similar individuals should be treated similarly [13, 14], etc. Another way to

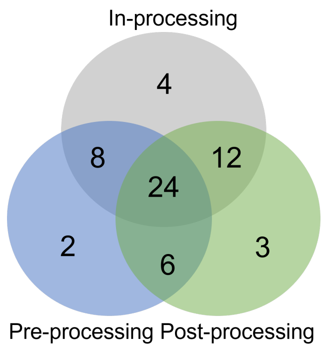

Figure 3: Distribution of fairness enhancing interventions considered in this paper. This includes 9 individual interventions and 50 different combinations of interventions.

classify fairness metrics can be on the level they operate on. For eg., dataset based metrics are solely calculated on the basis of the dataset and are independent of the classifier. On the other hand, classifier based metrics are dependent on the predictions of the classifier like the False negative rate difference. In this work, we have opted for a diverse set of fairness metrics to paint a more comprehensive picture. Here, we have not used any dataset based metrics due to their inability to capture the impact of in-processing and post-processing interventions.

The efficacy of different fairness enhancing interventions is typically measured using different fairness metrics. However, these metrics do not present the impact on different population groups. For example, the impact on different groups is irrelevant for fairness metrics following the notion of individual fairness. This even holds true for multiple fairness metrics based on the notion of group fairness. Such metrics focus on measuring the disparity between groups without much regard for the impact on specific groups. For example, the fairness metric, False Negative Rate difference, reports the difference in false negativity rate between groups. An increase/decrease in this metric does not tell us anything about the specific impact on the privileged or unprivileged groups. In this work, we analyze how interventions impact different fairness metrics and population groups.

2.3 Fairness across ML Pipeline

Research that focuses on fairness in a multi-stage ML system has received some attention and is still in its early stages [8, 32, 38, 39]. Biswas et al. studied the impact of data preprocessing techniques like standardization, feature selection, etc. on the overall fairness of the ML pipeline [8]. They found certain data transformations like sampling to enhance bias. Hirzel et al. also focus on the data preprocessing stage [32]. They present a novel technique to split datasets into train/test that is geared towards fairness. Wang et al. focused on fairness in the context of multi-component recommender systems [38]. They found that overall system's fairness can be enhanced by improving fairness in individual components. Our work focuses on how different combinations of interventions at 3 stages of the ML pipeline can be leveraged to enhance fairness across the ML pipeline.

There is a related line of work that discusses fairness in the context of compound decision making processes [9, 15, 17]. In such data systems, there is a sequence of decisions to be made where each decision can be thought of as a classification problem. For eg., in a two stage hiring process, candidates are first filtered for the interview stage and the remaining candidates are again filtered to determine who gets hired. This line of work focuses on fairness over multiple tasks and does not pay much attention towards enhancing fairness of an individual task. Here, a single task (classification problem) can be thought of as a ML pipeline. This is where our work comes in. Our work studies the different combinations of interventions that can together enhance fairness for a single decision making process.

3 EXPERIMENT SETUP

We have used IBM's AIF 360 [6] open source toolkit to conduct all experiments for this paper. More specifically, we leveraged 4 datasets, 9 fairness enhancing interventions and 11 evaluation metrics from this toolkit as shown in Fig. 2. To have a more even comparison, we have used the same ML model i.e., logistic regression (linear model) across the board. Moreover, we only selected those in-processing interventions that are based on or compatible with linear models.

Interventions. Among the 9 interventions, 2 belong to the pre-processing stage, 4 belong to the in-processing stage and 3 belong to the post-processing stage. Apart from these individual interventions, we also execute different combinations of these interventions in groups of 2 and 3. For example, one might choose to intervene at any 2 stages (say a pre-processing intervention followed by a post-processing intervention) or choose to intervene at all 3 stages of the ML pipeline. To form all possible combinations, we cycle through all available options (interventions) for a given ML stage along with a 'No Intervention' option and repeat it for all the 3 stages. This results in 8 combinations of pre-processing and in-processing interventions, 12 combinations of in-processing and post-processing interventions, 6 combinations of pre-processing and post-processing interventions and 24 combinations of of all 3 types of interventions (see Fig. 3). In totality, we perform 9 individual interventions, 50 different combinations of interventions and a baseline case (No intervention for all stages) for each of the 4 datasets. Here, we have used the default set of hyperparameters for all interventions. In this paper, we will refer to the different interventions by their acronyms like PR for Prejudice Remover as defined in subsection 2.1. For cascaded interventions, we will concatenate the respective acronyms with a '+' sign. For example, OP + PR means that we performed the Optimized Preprocessing (OP) intervention followed by the Prejudice Remover (PR) intervention. The baseline case is referred as 'Logistic Regression'.

Evaluation Metrics. The impact of the different interventions is captured using a diverse set of 11 evaluation metrics. Two of them, namely Accuracy and F1 score, are utility metrics that measure the ability of a ML model to learn the underlying patterns from the training dataset. Here, we have included the F1 score as it can better deal with imbalanced output class distributions. Both of these metrics range between 0 and 1. Higher values mean better performance. The remaining 9 metrics each capture some facet of

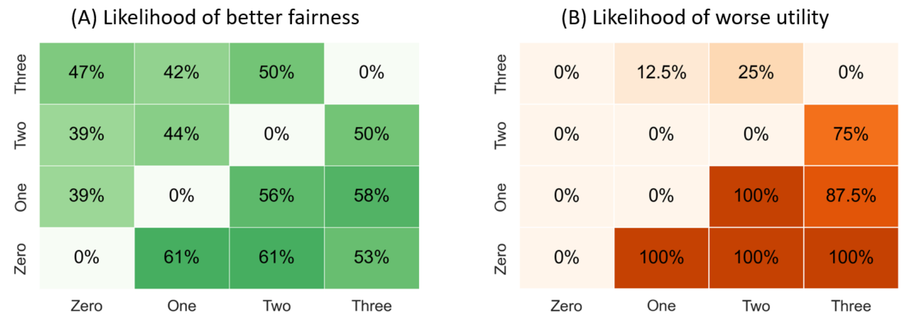

Figure 4: Heatmaps for Fairness and Utility metrics across different numbers of interventions. In Figure (A), a cell (i,j) represents the percentage of cases where j interventions yielded better fairness metrics than i interventions. In Figure (B), a cell (i,j) represents the percentage of cases where j interventions yielded worse utility metrics than i interventions. Here, i represents rows and j represents columns.

fairness. Two of the fairness metrics, namely Consistency and Theil index, subscribe to the notion of individual fairness. Higher values for Consistency and lower values for the Theil index mean more fairness. All other fairness metrics subscribe to the notion of group fairness, namely False Positive Rate Difference (FPR Diff), False Negative Rate Difference (FNR Diff), Statistical Parity Difference (SPD), False Discovery Rate Difference (FDR Diff), False Omission Rate Difference (FOR Diff), Accuracy Difference (Accuracy Diff) and F1 Score Difference (F1 Score Diff). All group fairness metrics measure disparity between groups based on some measure such as False Positive Rate (FPR). A lower absolute value for the group fairness metrics means more fairness. The sign of these metrics represents the group that is getting the upper/lower hand. A value of 0 means perfect fairness.

Datasets. Each of the 4 tabular datasets used in this paper have been used extensively in the fairness literature. They deal with a binary classification problem and typically contain one or more binary protected attributes such as gender, race, etc. For each of these datasets, we have used the default preprocessing procedure as provided by the AIF360 package (not to be confused with preprocessing interventions). It should be noted that the default preprocessing often involves one hot encoding to deal with categorical variables; this inflates the number of columns compared to the original dataset. We describe the datasets briefly as follows:

Adult Income Dataset. After pre-processing, this dataset consists of 45,222 rows and 99 columns that are derived from the 1994 Census database. Each row represents a person characterized by variables like education, gender, race, workclass, etc. These attributes are used to predict if an individual makes more than $50k a year. Here, we have used gender as the sensitive attribute with males as the privileged group and females as the unprivileged group.

German Credit Dataset. After pre-processing, this dataset consists of 1,000 rows and 59 columns which was originally prepared by Prof. Hofmann. The task is to predict if an individual has good or bad credit risk based on features like credit amount, credit history, personal status, sex, etc. Here, the sensitive attribute is age. Individuals older than 25 years belong to the privileged group and vice versa.

COMPAS Recidivism Dataset. After pre-processing, this dataset contains 6,167 rows and 402 columns which pertains to the COMPAS algorithm used for scoring defendants in Broward County, Florida. The task is to predict if an individual will recommit a crime within a two year period based on personal attributes like charge degree, prior count, etc. Here, the sensitive attribute is race with Caucasians as the privileged group and non-Caucasians as the unprivileged group.

Bank Marketing Dataset. After pre-processing, this dataset consists of 30,488 rows and 53 columns; it pertain to a direct marketing campaign of a Portuguese banking institution. The classification task is to predict if a client will buy a term deposit based on features like type of job, marital status, education, etc. Here, the sensitive attribute is age. Individuals (clients) younger than 25 years belong to the unprivileged group and vice versa.

For each of these datasets, the positive outcome label refers to the favorable outcome for the recipient. For example, the positive outcome label for the adult income dataset refers to an income greater than $50k. Similarly, for the COMPAS dataset, positive outcome label refers to *not* recommitting a crime in 2 years. This information will help interpret measures such as false positive rate, false negative rate, etc.

Method. After default pre-processing, we standardize different features of the dataset so that all non-protected features have the same mean and standard deviation. Thereafter, each dataset is randomly divided into train and test dataset in the ratio 70:30. For the baseline case, we train a logistic regression model on the training dataset and then compute different evaluation metrics using the test data. Next, we execute all individual and cascaded interventions using the train dataset and record their impact on different utility and fairness metrics using the test dataset. Apart from these metrics, we also record statistics like false negative rate (FNR), base rate, etc. for the privileged and unprivileged groups. This entire process is repeated 3 times for each dataset with different random splits between train and test dataset to counter sampling bias. Lastly, we compute the mean values of all evaluation metrics across the 3 iterations for each intervention. For each dataset, these results can be represented in tabular format with 60 rows and 11 columns where

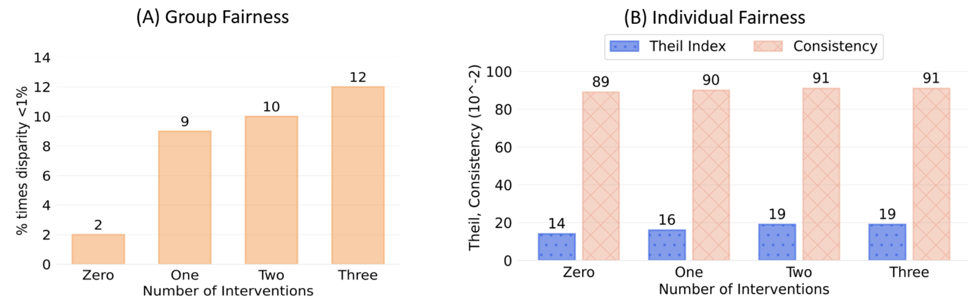

Figure 5: Effect of individual and cascaded interventions on fairness metrics (A) Percentage of times the disparity between the privileged and unprivileged groups was less than 1% across all group fairness metrics. A higher value means more group fairness. (B) Mean values for Theil index and Consistency across different numbers of interventions. Lower values of the Theil index and higher values for Consistency means more individual fairness.

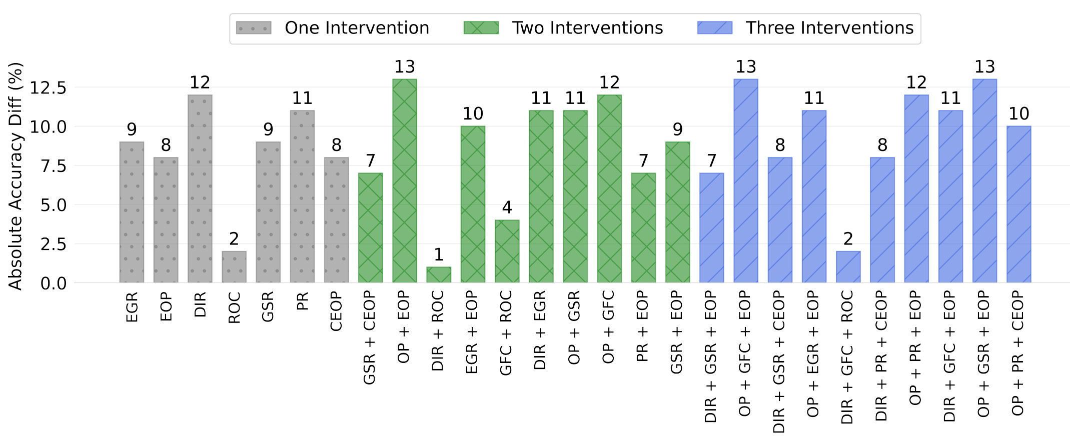

Figure 6: Absolute values for the Accuracy difference metric across different interventions for the Adult Income dataset. Here, lower values are desirable. This plot shows multiple cases where values corresponding to more number of interventions are larger than lower number of interventions. Hence, more interventions does not always lead to more fairness.

each row represents an intervention and each column represents an evaluation metric.

4 RESULTS

In this section, we analyze the empirical data from our experiments to understand the possible effect of different cascaded interventions on fairness, utility and population groups.

4.1 Effect of Cascaded Interventions on Fairness Metrics

We first gauge the impact of cascaded interventions on fairness as a whole (across fairness metrics). We start by grouping all interventions into 4 buckets i.e., 0 intervention, 1 intervention, 2 interventions and 3 interventions, respectively. For each bucket, we compute the average score for different fairness metrics and repeat this process for all datasets. It is important to note that different fairness metrics are not directly comparable as they are based on different interpretations of fairness and also vary in terms of their numerical distribution (range, mean, standard deviation, etc.). So, we will compare the mean value of a fairness metric with its counterpart for a different bucket. We count the percentage of times one bucket performs better than another across fairness metrics and datasets. This data is visualized using a heatmap of size 4 x 4 in Figure 4(A). Each row and column represents a bucket (number of interventions). Here, a cell (i,j) represents the percentage of times j interventions performs better than i interventions. For example, the cell (2,1) is labeled 44%. It means that a single intervention yielded better fairness scores than two interventions for 44% of cases. A bucket j will be considered favorable over another bucket i if the value for the cell (i,j) is greater than 50% and vice versa.

It should be noted that different fairness metrics might be incompatible with each other. So, the net trend (cell values) might appear a bit faded as some fairness metrics might cancel the effect of another. Looking at the row i=0, we find that any number of interventions greater than 0 provide better overall fairness than having no interventions. Looking at the row i=1, we find that the

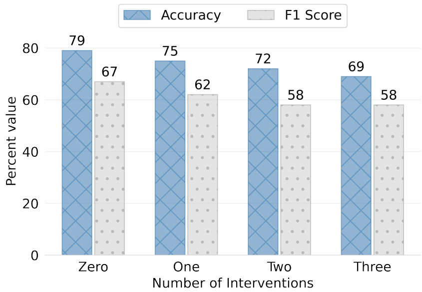

Figure 7: Mean Accuracy and F1 Score for different number of interventions across all datasets.

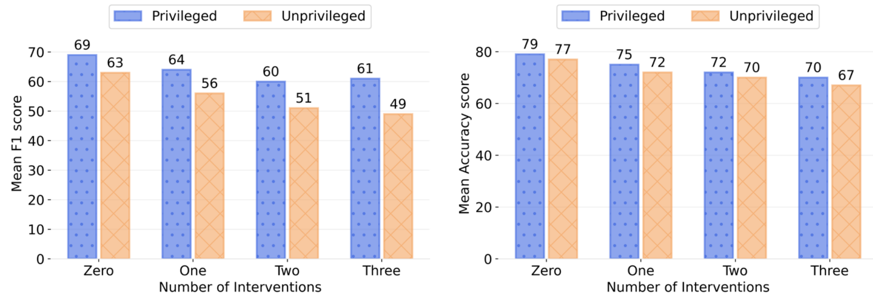

Figure 8: Mean F1 Score (Left) and Accuracy (Right) for different number of interventions. We observe that both metrics decrease for the privileged and the unprivileged groups with more number of interventions.

columns j=2 and j=3 have values more than 50% i.e., two or three interventions yielded better fairness than a single intervention. Moving to the row i=2, we find that the value of the cell (2,3) is 50%. Perhaps surprisingly, this means that it is equally likely for either buckets to outperform each other. Overall, it appears that fairness improves from 0 to 2 interventions and becomes constant thereafter. However, it is important to note that the heatmap encodes frequency and not the magnitude of difference between fairness metrics. So, it is possible that three interventions reduce bias significantly more (in terms of magnitude and not the count of fairness metrics) than the two interventions case and might still appear to be no better than the two interventions case.

The heatmap provides an aggregate picture of how fairness metrics vary for different numbers of interventions. Now, let us dig a bit deeper and gauge the impact of cascaded interventions on group fairness and individual fairness. For individual fairness, we plot the mean values for the Theil index and Consistency for different numbers of interventions. For the group fairness metrics, we compute the percentage of times the absolute value of each constituting metric is less than 0.01. As we can see from Figure 5 (A), the percent of times the group fairness metrics are below a threshold increases steadily with higher numbers of interventions from 2% to 12%. In other words, *group fairness improves with more interventions on aggregate.* This observation largely concurs with our findings from Figure 4(A). On the other hand, we get mixed signals from the individual fairness metrics. The Consistency metric shows a slight improvement in fairness while the Theil index shows a downfall in fairness with higher number of interventions.

It is important to note that all of these patterns reflect the aggregate trend and may not apply for all cases. For example, Figure 6 shows the absolute values for the Accuracy difference metric across different interventions for the Adult Income dataset. In this case, we observe multiple instances where a larger number of interventions did not lead to more fairness. On the other hand, we observed multiple instances where lower number of interventions performed better than higher numbers of interventions. This observation is contrary to the aggregate trend for group fairness that we observed in Figure 5(A). So, *it is not always the case that more interventions will result in more fairness. One needs to choose the right combination of interventions to get the best results.* We will discuss which combinations work for different metrics in subsection 4.4.

4.2 Effect of Cascaded Interventions on Utility Metrics

We start off by analyzing how different number of interventions compare against each other on utility metrics as a whole. Following a similar procedure as defined in subsection 4.1, we plot a heatmap for utility metrics instead of fairness metrics (see Figure 4 (B)). Here, a cell (i,j) represents the percentage of cases where j interventions yielded lower utility than i interventions. As expected, we observe that any non-zero number of interventions results in lower utility than the baseline case (see row i=0). Similarly, we observe that two interventions yields worse utility metrics than one intervention (see cell(1,2)) and three interventions yields worse utility metrics than two interventions (see cell(2,3)). Overall, this reveals a strong downward trend for utility metrics with more number of interventions. Looking at Figure 4 (A) and (B) in conjunction, we observe that three interventions perform on par with two interventions on fairness. However, 75% of the times three interventions performed worse on utility metrics than two interventions. This observation hints that one should typically opt for two interventions and go for the third intervention only in specific contexts.

To quantitatively understand the effect on specific utility metrics, we analyzed how Accuracy and F1 score vary across different number of interventions. We computed the mean accuracy and F1 score for different number of interventions across datasets. These mean scores are visualized in Figure 7. In line with our findings in Figure 4(B), we observe that both accuracy and F1 score steadily decrease as the number of interventions increase. This downward trend is more pronounced in the beginning than the end. For eg., the mean F1 score drops by 5% going from no intervention to one intervention and later stabilizes going from two interventions to three interventions. Overall, this trend shows that there is a cost to be paid for adding more interventions. So, *one should not blindly opt for more interventions.* ML practitioners should consider the potential loss in utility metrics while designing fair ML pipelines.

We have looked at the effect of cascaded interventions on utility metrics and fairness metrics in isolation. Now, let us investigate the effect of cascaded interventions on the bivariate relationship between utility metrics and fairness metrics. We start off by grouping all experimental data across datasets by the number of interventions. For each group, we compute the spearman correlation coefficient between different fairness metrics and F1 across as shown in Table 1.

Table 1: Spearman correlation coefficient between F1 score and fairness metrics for different number of interventions (represented as rows). We observe that the correlation coefficient decreases as the number of interventions increase across metrics.

	FPR Diff	FNR Diff	Accuracy Diff	FOR Diff	FDR Diff	SPD	F1 Score Diff	Theil Index	Consistency
0	0.748	0.42	0.385	0.42	0.329	0.678	0.58	0.708	-0.986
1	0.343	-0.054	0.211	0.354	0.176	0.263	0.253	0.148	-0.628
2	0.228	-0.11	0.15	0.24	0.148	0.109	0.115	-0.144	-0.481
3	0.119	-0.185	0.189	0.058	-0.112	-0.056	0.073	-0.282	-0.167

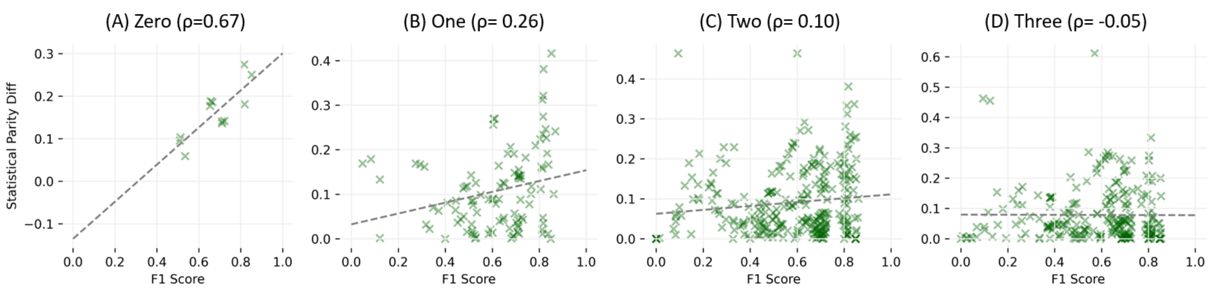

Figure 9: Relation between Statistical Parity difference and F1 score for different numbers of interventions from 0 to 3 (A - D). Each green 'x' marker corresponds to a specific intervention executed on one of the 3 random subsets of a given dataset. The grey line represents the regression line that best fits all the points. It visually indicates the strength of the correlation.

We observe that there is a significant positive correlation between fairness metrics and F1 score for the baseline case. For all fairness metrics except for the consistency metric, a higher value means more bias (less fairness). Hence, a positive correlation suggests that F1 score and fairness are negatively linked. In other words, interventions with high F1 score generally result in poor fairness and vice versa. This observation is in line with existing literature which discusses a tradeoff between accuracy and fairness for individual interventions [34]. As the number of interventions increases, we observe a steady decline in the correlation coefficient across fairness metrics. Here, the correlation coefficient for consistency moves in the opposite direction as unlike all other fairness metrics higher values means more fairness. *The decrease in correlation suggests that the likelihood of attaining a high F1 score along with high fairness increases with higher numbers of interventions.* As an example, we plot the bivariate relation between Statistical parity difference and F1 score for different numbers of intervention (see Figure 9). We observe that three interventions are able to achieve high F1 score and low bias scores more consistently than one or two interventions. Here, the decrease in correlation coefficient ρ is evident from the decrease in the slope of the regression line as we move towards higher numbers of interventions. If the reduction in F1 score caused by different interventions was in proportion to the corresponding increase in fairness, the correlation coefficient would have remained constant across different number of interventions. Hence, the decrease in correlation coefficient shows the efficacy of cascaded interventions in reducing bias without sacrificing too much on performance (F1 Score).

4.3 Impact on Population groups

In our experimental setup, we kept a log of different statistics like false positive rate, false negative rate, F1 score, base rate, etc. for the privileged and unprivileged groups across all interventions. We analyzed this data to understand the impact of different interventions on these two groups. Figure 8 shows the aggregate impact of different number of interventions on Accuracy and F1 Score. In line with our earlier finding (see Figure 7), we observe that these utility metrics deteriorate for both groups with more number of interventions. However, the impact on the underprivileged group is more severe than the privileged group for the F1 metric. The F1 score for the privileged group dropped 8 percent points from 69% to 61% while it dropped 14 percent points for the underprivileged group from 63% to 49%. This disproportionate impact also lead to an increase in disparity between the groups in terms of F1 Score from 6% to 12%. In the case of accuracy, the impact on both groups is roughly even and the disparity between groups remains almost constant (~3%) across different number of interventions.

The decrease in utility metrics signal an increase in error rates. So, let us look at the impact on the False positive rate (FPR) and False negative rate (FNR). Figure 10 shows the percentage of times we observed an increase in FNR and FPR compared to the baseline for different number of interventions. As we can see, there is a large percentage of cases where individual interventions resulted in higher error rates for both the privileged and the unprivileged group than we started out with. As we go for higher numbers of interventions, the percentage of such cases generally increases further. This trend is in agreement with the decreasing trend in utility metrics for more number of interventions. On comparing between groups, we find that interventions are more likely to result in higher FNR for the privileged group than the unprivileged one. It means that individuals from the privileged group are more likely to be misclassified with the unfavorable outcome than the unprivileged group. This trend flips for FPR where the unprivileged group are more likely to have a higher FPR. In other words, individuals from

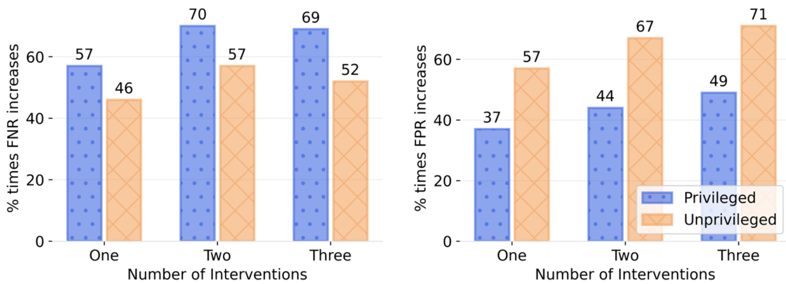

Figure 10: Percentage of times False Negative Rate (Left) and False Positive Rate (Right) increases compared to the baseline (No Intervention) across different number of interventions for all datasets.

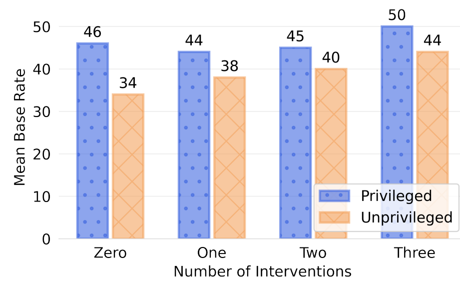

Figure 11: Mean base rate for the privileged and unprivileged groups across different numbers of interventions.

the unprivileged group are more likely to be misclassified with the favorable outcome than the privileged group. Both of these trends persist for different number of interventions and generally deepens with more number of interventions. Looking at these patterns in conjunction with Figure 8, it appears that the loss in Accuracy/F1 score can atleast be partially explained by the tendency of the interventions to assign more positive outcomes to the unprivileged group and negative outcomes to the privileged group.

Next, let us look at the impact on base rate for different groups (see Figure 11). Here, base rate is defined as the proportion of positive outcomes for different groups. It is computed over model's prediction for the test data post all relevant interventions. For the no intervention case, we observe a 12% disparity in favor of the privileged group. With more interventions, the base rate for the unprivileged group steadily increases from 34% to 44% (10 % jump). On the other hand, the base rate for the privileged group decreases a bit for the one and two interventions case and increases by 4% for the three interventions case. Overall, this leads to a decrease in disparity between groups from 12% to 5% for the two intervention case. In a context where equality between base rates is a priority, two interventions seems to be the way to go. It is also important to note the some of the interventions can negatively impact the privileged group. This is evident from the drop in base rate for the privileged group for the one and two interventions case. If we look at base rate over the entire population, we find that the base rate undergoes a modest increment for the one (1%) and two interventions case (2%). However, it increases significantly for the three interventions case (7%). So, ML practitioners should exercise caution while adding the third intervention, especially for contexts where the number of favorable outcomes is fixed such as hiring.

Fairness metrics provide elegant mathematical representations for different notions of fairness but might obscure the specific effect on different population groups. Figure 5 shows that group fairness metrics decrease with more interventions on aggregate. However, we do not observe a similar reduction in error rates for individual groups as shown in Figure 10. To investigate this further, let us look at a specific example. Figure 12 shows the difference in false negative rates (FNR) compared to the baseline for the privileged (males) and the unprivileged (females) groups across different interventions. Here, we have only considered interventions that have reduced the magnitude of the FNR difference metric between males and females compared to the baseline for the Adult Income dataset. So, as per the FNR Diff metric, all of these interventions are effective at reducing bias. However, we observe that for many of these interventions FNR has actually increased for one or both groups. In Figure 12, we observe 10 cases (interventions) where FNR has increased for either or both groups. So, the reduction in the FNR Diff metric is due to the uneven increase in FNR for different population groups. It can be argued whether it is desirable to reduce disparity between groups by increasing error rates unevenly for different groups. *This finding points to a need for new fairness metrics that account for the specific impact on individual groups apart from just the gap between those groups.*

There can be different ways to interpret the empirical findings based on one's value system. One way to interpret these numbers can be from a pure ML perspective where the focus is to train a ML model that best fits the underlying dataset. Here, the objective of an intervention is to ensure that the model performs equally well for different groups in terms of Accuracy, False positive rate, False negatives rate, etc. From this viewpoint, many of the interventions are counterproductive as they increase error rates and decrease the Accuracy/F1 score for different groups (see Figure 8 and Figure 10). Going from individual interventions to cascaded interventions makes things worse as the error rates for population groups increases further and the Accuracy/F1 score further deteriorates. Ideally, one would reduce disparity by reducing the error rates for different groups but not making it worse for any of them. For example, 3 interventions in Figure 12 reduce disparity (FNR Diff metric) by reducing FNR for both groups. Similarly, mitigating discrimination against one group at the cost of another is self-defeating and unethical. For illustration, we find 5 interventions in Figure 12 where FNR decreases for the unprivileged group and increases for the privileged group compared to the baseline. In other words, some high income individuals from the privileged group were misclassified as low income in an effort to increase fairness. This is not a one off case. Our experiments show an aggregate trend across datasets where privileged groups were disproportionately misclassified with unfavorable outcomes and unprivileged groups were disproportionately misclassified with favorable outcomes (see Figure 10). These observations highlight how interventions can negatively impact the privileged group. Future interventions should

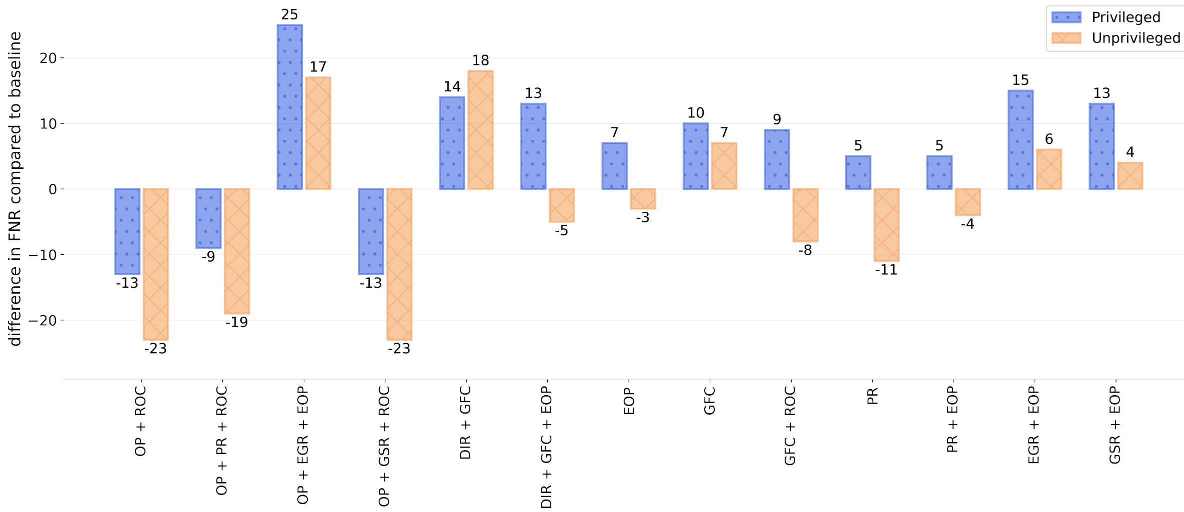

Figure 12: Difference in False Negative Rate compared to the baseline (No Intervention) across different Interventions for the Adult Income dataset. Here, Negative values are desirable and vice versa.

be more considerate towards their possible negative impact on the different population groups.

Another way to interpret these numbers can be from the perspective of social justice where interventions should not only reduce disparity in error rates but also serve as an instrument to right historical wrongs. For example, interventions should enforce equality/equity of outcomes i.e., ML models should assign positive outcomes to different groups at the same rate or even prioritize the unprivileged group irrespective of the patterns in the dataset [33]. In the process of achieving these goals, the loss incurred in terms of accuracy and F1 score is secondary as these metrics are computed over output labels (ground truth) polluted with historical biases. Similarly, the disproportionate increase in FNR for the privileged group is incidental/imperative to serve the larger goal of equal representation. Under this viewpoint, cascaded interventions are desirable as they help bridge the gap in base rates.

4.4 Comparison between different Interventions

So far, we have largely focused on the aggregate trends which might or might not apply for a given intervention. In this section, we focus on specific interventions and compare how they perform against different utility and fairness metrics. Such knowledge can assist practitioners and researchers in choosing interventions based on their specific context. We begin by computing the mean scores for all evaluation metrics across 4 datasets corresponding to each intervention. This provided a mean score for each evaluation metric across 60 different interventions. For each evaluation metric, we ranked all interventions from the best performing to the worst performing. For the Accuracy, F1 score and Consistency metric, higher values are desirable so we sorted interventions based on descending order of their corresponding values. For all other metrics, we used ascending order. It is important to note that we used absolute values for all group fairness metrics as we are primarily concerned with the magnitude of bias.

Table 2 contains the 10 best and worst performing interventions for each evaluation metric. From this table, we can make a few important observations. Logistic regression (by which we mean 'No intervention') tops the list for Accuracy. This is in line with existing literature that comments on the accuracy fairness tradeoff [7]. As all interventions optimize for some aspect of fairness, they might sacrifice a bit on accuracy. So, one should not use any intervention for achieving the best accuracy. For the F1 score, we observe that a few interventions rank higher than Logistic Regression, such as DIR + EGR + ROC. However, the difference between them was quite slim (0.3%) which might be attributed to imbalanced output class distribution. The broader point is that *applying more interventions does not always lead to a significant loss in utility.* In the case of DIR + EGR + ROC, we observe the best performance for the F1 score and the 4th best for accuracy. Among the 10 bottom ranked interventions across all fairness metrics, Logistic Regression occurs only twice. This shows that there are several individual and cascaded interventions that perform worse than the baseline case for at least some fairness metric. Hence, it is important to choose interventions wisely. ML practitioners/researchers can leverage resources like Table 2 and prioritize interventions that have worked well for other datasets and hopefully save some time in the process.

For the fairness metrics, we observe that the best performing intervention is mostly unique for each one of them. In other words, there is no silver bullet for all fairness metrics. This observation is in line with the existing literature which proves that no intervention can simultaneously optimize for all fairness metrics [29]. From a practical standpoint, this implies that ML practitioners need to prioritize which metrics are more important to them and then choose interventions accordingly. It is also worth noting that the best performing intervention for any metric is either Logistic Regression (No Intervention) or a combination of two or more interventions.

Table 2: Ranking of 10 best and worst performing interventions for different evaluation metrics. Here, we have ranked interventions in ascending order of their corresponding absolute values for all metrics except for Accuracy, F1 Score and Consistency that are sorted in descending order. This ordering schema ensures that desirable interventions are ranked higher for all metrics. Interventions with '+' sign represents combinations of multiple interventions. Here, 'Logistic Regression' represents the baseline case i.e., no internvetions for all stages.

Rank	Accuracy	F1 Score	Theil Index	Consistency	FPR Diff
1	Logistic Regression	DIR + EGR + ROC	DIR + EGR + ROC	OP + GFC + ROC	EGR + EOP
2	DIR	DIR	DIR	OP + GFC + CEOP	OP + GFC + ROC
3	GSR	DIR + GSR	DIR + EGR	OP + GFC	PR + EOP
4	DIR + EGR + ROC	DIR + EGR	DIR + GSR	OP + GSR	OP + EGR + EOP
5	DIR + EGR	Logistic Regression	DIR + PR + EOP	OP	OP + GFC + EOP
6	DIR + GSR	OP + CEOP	DIR + ROC	OP + EGR + ROC	OP + GFC
7	EGR	OP + GSR + CEOP	DIR + CEOP	OP + PR	DIR + GFC
8	DIR + PR + CEOP	OP	Logistic Regression	OP + GSR + CEOP	GFC + EOP
9	CEOP	OP + GSR	DIR + PR + ROC	OP + CEOP	DIR + EOP
10	PR + CEOP	GSR	DIR + GSR + CEOP	OP + PR + CEOP	DIR + GSR + EOP
51	OP + GFC + CEOP	OP + PR	OP + PR + ROC	PR + EOP	Logistic Regression
52	OP + PR + ROC	OP + GFC + EOP	OP + PR	DIR + PR + ROC	GSR + CEOP
53	DIR + GFC + EOP	DIR + GFC + EOP	OP + ROC	DIR + EGR + EOP	OP
54	DIR + GFC	GFC + EOP	OP + GFC	DIR + EOP	PR + CEOP
55	OP + ROC	OP + GFC	DIR + GFC	DIR + GFC + EOP	GFC + CEOP
56	OP + GFC + EOP	DIR + GFC	GFC + CEOP	EOP	DIR + GSR + CEOP
57	OP + GSR + ROC	GFC + ROC	OP + GFC + EOP	OP + GSR + EOP	DIR + PR + CEOP
58	DIR + PR + ROC	DIR + GFC + CEOP	GFC + ROC	OP + EOP	OP + CEOP
59	DIR + ROC	GFC + CEOP	OP + GSR + ROC	OP + PR + EOP	DIR + CEOP
60	OP + PR	DIR + GFC + ROC	DIR + GFC + ROC	OP + GFC + EOP	CEOP

Rank	FNR Diff	Accuracy Diff	FOR Diff	FDR Diff	SPD
1	OP + GFC + EOP	DIR + GFC + ROC	OP + PR	Logistic Regression	OP + GFC + EOP
2	GFC + EOP	DIR + GFC	DIR + GFC + CEOP	DIR	OP + EGR + ROC
3	OP + EGR + EOP	OP + PR + ROC	CEOP	GSR + CEOP	OP + EGR
4	OP + EOP	OP + ROC	PR + CEOP	DIR + EGR + CEOP	OP + GSR + EOP
5	OP + GSR + ROC	ROC	DIR + CEOP	GFC + CEOP	OP + GFC + ROC
6	OP + GFC + ROC	EOP	PR	GFC	GSR + ROC
7	OP + PR + EOP	DIR + GFC + EOP	DIR + PR + CEOP	DIR + GFC	DIR + GFC
8	DIR + GSR + EOP	OP + GSR + ROC	OP + GFC + ROC	DIR + PR	OP + EOP
9	GSR + EOP	GFC + CEOP	DIR + GSR + CEOP	DIR + EGR	EGR + ROC
10	OP + GFC	PR + ROC	GFC + CEOP	PR	OP + PR + EOP
51	GFC + CEOP	OP + EGR	GSR + ROC	OP + GFC + CEOP	GFC + CEOP
52	DIR + EGR + CEOP	OP + PR	GFC + EOP	DIR + GSR + EOP	PR
53	GSR + CEOP	OP + PR + CEOP	OP + ROC	OP + EGR	OP + CEOP
54	DIR + GFC + CEOP	OP + EGR + ROC	OP + CEOP	OP + EGR + ROC	GSR + CEOP
55	EGR + CEOP	OP + GSR	OP + GSR + ROC	OP + PR	DIR + PR + CEOP
56	DIR + GSR + CEOP	OP + GSR + CEOP	OP + PR + ROC	OP + EOP	Logistic Regression
57	DIR + PR + CEOP	OP + GFC + ROC	OP + EGR + EOP	OP + EGR + EOP	DIR + GSR + CEOP
58	PR + CEOP	OP + EGR + CEOP	DIR + GFC + EOP	OP + PR + EOP	PR + CEOP
59	DIR + CEOP	OP + GFC	OP + EGR + ROC	OP + GFC + EOP	DIR + CEOP
60	CEOP	OP + GFC + CEOP	OP + EGR	OP + GSR + EOP	CEOP

Apart from the top performing interventions, we also observe that the top 10 interventions for all fairness metrics are predominantly cascaded interventions. For example, 9 out of the top 10 interventions for the Consistency metric are cascaded interventions. These observations further motivate the efficacy of cascaded interventions over individual interventions. Among the top 10 interventions across all 10 metrics, OP + GFC + ROC occurs the most number of times (5). Similarly, OP + GFC + EOP occurs the most number of times among the bottom 10 interventions across metrics. It is interesting to see that both of these interventions have much in common (OP and GFC). This shows that certain intervention are more compatible/incompatible with another. Changing an ingredient can drastically impact the outcome. For instance, swapping ROC with EOP resulted in the entire combination (OP + GFC + ROC) to change from being one of the top ranked to one of the worst ranked interventions. It should be noted that ranking abstracts the real difference in magnitude. For brevity, we have used ranking in the table. We encourage the readers to refer to the source code/experimental data for more details.

5 DISCUSSION

This work explores different research questions in the realm of cascaded debiasing based on comprehensive experiments using multiple benchmark datasets, fairness metrics and interventions. The scope of this paper is limited to the binary classification problem for tabular datasets. Future work might conduct similar studies for other data types like text, images, etc. and consider other problems types such as regression, clustering, etc. It should be noted that all the insights and analyses presented in this paper are based on empirical evidence and so they may or may not generalize to other datasets, interventions or metrics. This study was facilitated by the AIF 360 package that provided easy access to different datasets, interventions and metrics. On the flip side, this package can also be considered as a limiting factor because our choices were limited to the different options it provided. Moreover, we were unable to execute certain interventions like the Optimized Preprocessing intervention for the Bank Marketing dataset due to limited technical support. In the future, researchers might also consider other fairness packages [31] like Fairlearn and possibly include more fairness metrics, datasets, interventions, etc.

On the fairness front, we have considered a respectable set of fairness metrics but there are other popular metrics like counterfactual fairness [30] or more recent metrics like Statistical Equity [33] that can be explored in future studies. Moreover, this work deals with fairness at a group level (say males and females) and at an individual level (through the individual fairness metrics). Future work might study the cumulative effect of fairness enhancing interventions on different subgroups say poor black females, young white males, etc. As far as datasets go, we have conducted experiments on 4 datasets that have been used extensively in the fairness literature to benchmark different fairness enhancing interventions. Recent research efforts have questioned the validity of some these datasets like the COMPAS [3] and Adult Income dataset [12]. Future work might include more recent datasets [12] and other stages of intervention, such as the data curation stage into the analysis [22]. Hyperparameters for different interventions can significantly impact the results. In the interest of reducing computational complexity, this work largely uses the default set of hyperparameters. Future work might go for a deeper analysis by optimizing hyperparameters for different interventions [39]. Lastly, the source code and experimental data has been made publicly available at *github.com/bhavyaghai/Cascaded-Debiasing* for easy reproducibility and for anyone to analyze the data in their own different ways.

ACKNOWLEDGMENTS

This research was partially supported by NSF grants IIS 1527200 and 1941613.

REFERENCES

[1] Alekh Agarwal, Alina Beygelzimer, Miroslav Dudík, John Langford, and Hanna Wallach. 2018. A reductions approach to fair classification. In *International Conference on Machine Learning*. PMLR, 60–69.

[2] Alekh Agarwal, Miroslav Dudík, and Zhiwei Steven Wu. 2019. Fair regression: Quantitative definitions and reduction-based algorithms. In *International Conference on Machine Learning*. PMLR, 120–129.

[3] Michelle Bao, Angela Zhou, Samantha Zottola, Brian Brubach, Sarah Desmarais, Aaron Horowitz, Kristian Lum, and Suresh Venkatasubramanian. 2021. It's COMPASlicated: The Messy Relationship between RAI Datasets and Algorithmic Fairness Benchmarks. *arXiv preprint arXiv:2106.05498* (2021).

[4] Yahav Bechavod and Katrina Ligett. 2017. Penalizing unfairness in binary classification. *arXiv preprint arXiv:1707.00044* (2017).

[5] Rachel KE Bellamy, Kuntal Dey, Michael Hind, Samuel C Hoffman, Stephanie Houde, Kalapriya Kannan, Pranay Lohia, Jacquelyn Martino, Sameep Mehta, Aleksandra Mojsilovic, et al. 2018. AI Fairness 360: An extensible toolkit for detecting, understanding, and mitigating unwanted algorithmic bias. *arXiv* (2018).

[6] Rachel KE Bellamy, Kuntal Dey, Michael Hind, Samuel C Hoffman, Stephanie Houde, Kalapriya Kannan, Pranay Lohia, Jacquelyn Martino, Sameep Mehta, Aleksandra Mojsilović, et al. 2019. AI Fairness 360: An extensible toolkit for detecting and mitigating algorithmic bias. *IBM Journal of Research and Development* 63, 4/5 (2019), 4–1.

[7] Richard Berk, Hoda Heidari, Shahin Jabbari, Matthew Joseph, Michael Kearns, Jamie Morgenstern, Seth Neel, and Aaron Roth. 2017. A convex framework for fair regression. *arXiv preprint arXiv:1706.02409* (2017).

[8] Sumon Biswas and Hridesh Rajan. 2021. Fair Preprocessing: Towards Understanding Compositional Fairness of Data Transformers in Machine Learning Pipeline. *arXiv preprint arXiv:2106.06054* (2021).

[9] Amanda Bower, Sarah N Kitchen, Laura Niss, Martin J Strauss, Alexander Vargas, and Suresh Venkatasubramanian. 2017. Fair pipelines. *arXiv preprint arXiv:1707.00391* (2017).

[10] Flavio P Calmon, Dennis Wei, Bhanukiran Vinzamuri, Karthikeyan Natesan Ramamurthy, and Kush R Varshney. 2017. Optimized pre-processing for discrimination prevention. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*. 3995–4004.

[11] Kate Crawford. 2017. The Trouble with Bias. (2017). Neural Information Processing Systems 2017 Keynote talk.

[12] Frances Ding, Moritz Hardt, John Miller, and Ludwig Schmidt. 2021. Retiring Adult: New Datasets for Fair Machine Learning. *Advances in Neural Information Processing Systems* 34 (2021).

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

[14] Cynthia Dwork and Christina Ilvento. [n. d.]. Group fairness under composition. In *Conference on Fairness, Accountability, and Transparency (FAT* 2018)*.

[15] Cynthia Dwork, Christina Ilvento, and Meena Jagadeesan. 2020. Individual fairness in pipelines. *arXiv preprint arXiv:2004.05167* (2020).

[16] Michael D Ekstrand, Mucun Tian, Mohammed R Imran Kazi, Hoda Mehrpouyan, and Daniel Kluver. 2018. Exploring author gender in book rating and recommendation. In *Proceedings of the 12th ACM Conference on Recommender Systems*. ACM, 242–250.

[17] Vitalii Emelianov, George Arvanitakis, Nicolas Gast, Krishna Gummadi, and Patrick Loiseau. 2019. The price of local fairness in multistage selection. *arXiv preprint arXiv:1906.06613* (2019).

[18] 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*. 259–268.

[19] Sorelle A Friedler, Carlos Scheidegger, and Suresh Venkatasubramanian. 2021. The (im) possibility of fairness: Different value systems require different mechanisms for fair decision making. *Commun. ACM* 64, 4 (2021), 136–143.

[20] Nikhil Garg, Londa Schiebinger, Dan Jurafsky, and James Zou. 2018. Word embeddings quantify 100 years of gender and ethnic stereotypes. *Proceedings of the National Academy of Sciences* 115, 16 (2018), E3635–E3644.

[21] Bhavya Ghai, Md Naimul Hoque, and Klaus Mueller. 2021. WordBias: An Interactive Visual Tool for Discovering Intersectional Biases Encoded in Word Embeddings. In *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–7.

[22] Bhavya Ghai, Q. Vera Liao, Yunfeng Zhang, and Klaus Mueller. 2020. Measuring Social Biases of Crowd Workers using Counterfactual Queries. arXiv:2004.02028 [cs.HC]

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

[24] Drew Harwell. 2018. The Accent Gap. (2018). https://www.washingtonpost.com/graphics/2018/business/alexa-does-not-understand-your-accent/?noredirect=on&utm_term=.8ca67c7f13b9 The Washington Post.

[25] Kenneth Holstein, Jennifer Wortman Vaughan, Hal Daumé III, Miro Dudik, and Hanna Wallach. 2019. Improving fairness in machine learning systems: What do industry practitioners need?. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. 1–16.

[26] Faisal Kamiran, Asim Karim, and Xiangliang Zhang. 2012. Decision theory for discrimination-aware classification. In *2012 IEEE 12th International Conference on Data Mining*. IEEE, 924–929.

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

[28] Michael Kearns, Seth Neel, Aaron Roth, and Zhiwei Steven Wu. 2018. Preventing fairness gerrymandering: Auditing and learning for subgroup fairness. In *International Conference on Machine Learning*. PMLR, 2564–2572.

[29] Jon Kleinberg, Sendhil Mullainathan, and Manish Raghavan. 2016. Inherent trade-offs in the fair determination of risk scores. *arXiv preprint arXiv:1609.05807* (2016).

[30] Matt J Kusner, Joshua R Loftus, Chris Russell, and Ricardo Silva. 2017. Counterfactual fairness. *arXiv preprint arXiv:1703.06856* (2017).

[31] Michelle Seng Ah Lee and Jat Singh. 2021. The landscape and gaps in open source fairness toolkits. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–13.

[32] Kiran Kate Martin Hirzel and Parikshit Ram. 2021. Engineering fair machine learning pipelines. *target* 73, 2.2 (2021), 1–028.

[33] Ninareh Mehrabi, Yuzhong Huang, and Fred Morstatter. 2020. Statistical Equity: A Fairness Classification Objective. *arXiv preprint arXiv:2005.07293* (2020).

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

[35] Vidya Muthukumar, Tejaswini Pedapati, Nalini Ratha, Prasanna Sattigeri, Chai-Wah Wu, Brian Kingsbury, Abhishek Kumar, Samuel Thomas, Aleksandra Mojsilovic, and Kush R Varshney. 2018. Understanding Unequal Gender Classification Accuracy from Face Images. *arXiv preprint arXiv:1812.00099* (2018).

[36] Arvind Narayanan. 2018. 21 Fairness Definitions and Their Politics. (2018). Conference on Fairness, Accountability, and Transparency, NYC.

[37] Geoff Pleiss, Manish Raghavan, Felix Wu, Jon Kleinberg, and Kilian Q Weinberger. 2017. On fairness and calibration. *arXiv preprint arXiv:1709.02012* (2017).

[38] Xuezhi Wang, Nithum Thain, Anu Sinha, Flavien Prost, Ed H Chi, Jilin Chen, and Alex Beutel. 2021. Practical Compositional Fairness: Understanding Fairness in Multi-Component Recommender Systems. In *Proceedings of the 14th ACM International Conference on Web Search and Data Mining*. 436–444.

[39] Qingyun Wu and Chi Wang. 2021. Fair AutoML. *arXiv preprint arXiv:2111.06495* (2021).