ECAI 2024
U. Endriss et al. (Eds.)

doi:10.3233/FAIA240585

FairUS - UpSampling Optimized Method for Boosting Fairness

Nurit Cohen-Inger[a,1], Guy Rozenblatt[a,1], Seffi Cohen[a], Lior Rokach[a] and Bracha Shapira[a]

[a]Ben-Gurion University of The Negev

Abstract. The increasing application of machine learning (ML) in critical areas such as healthcare and finance highlights the importance of fairness in ML models, challenged by biases in training data that can lead to discrimination. We introduce 'FairUS', a novel pre-processing method for reducing bias in ML models utilizing the Conditional Generative Adversarial Network (CTGAN) to synthesize upsampled data. Unlike traditional approaches that focus solely on balancing subgroup sample sizes, FairUS strategically optimizes the quantity of synthesized data. This optimization aims to achieve an ideal balance between enhancing fairness and maintaining the overall performance of the model. Extensive evaluations of our method over several canonical datasets show that the proposed method enhances fairness by 2.7 times more than the related work and 4 times more than the baseline without mitigation, while preserving the performance of the ML model. Moreover, less than a third of the amount of synthetic data was needed on average. Uniquely, the proposed method enables decision-makers to choose the working point between improved fairness and model's performance according to their preferences.

1 Introduction

ML models have become essential tools in various decision-making fields. The reliability and effectiveness of these models are highly dependent on the quality and representativeness of the data used for training. A critical aspect of training ML models involves ensuring fairness [14]. Minimizing bias is essential in ML models to guarantee that automated decisions are equitable and do not perpetuate existing societal inequalities, particularly in sensitive domains like healthcare, finance, and criminal justice. Biases related to protected attributes such as race, gender, occupation and age can result in unfair outcomes, disproportionately affecting specific groups. This could lead to discrimination and other negative impacts on individuals from these groups.

Bias mitigation techniques in ML can be broadly categorized into three types [25]: pre-processing, in-processing, and post-processing. Pre-processing techniques mitigate bias in the training data before AI models are affected. In-processing techniques address bias during the training process of the model, while post-processing techniques attempt to mitigate bias in the models' results. Pre-processing techniques for bias mitigation are instrumental in ensuring fairness by reducing bias in training data, regardless of their model's training process or architecture. Bias in datasets can considerably influence the predictions of the model, leading to unjust outcomes, especially for underrepresented minority groups. These techniques focus on identifying and rectifying biases in the data before the training process commences. By addressing the root cause of bias, these methods foster transparency and accountability in AI systems. Pre-processing methods encompass several techniques, including transforming protected attributes to balance group representation [11, 5, 23], fair representation learning [4, 12, 33, 30, 31, 7, 34], relabeling and perturbation methods [16, 24, 11, 21], and sampling strategies such as upsampling and downsampling [20, 9, 1]. These techniques aim to eliminate biases from training data and balancing group representation, which are both the focus of our work.

Our motivation was initiated by identifying gaps in how traditional sampling techniques affect fairness, particularly evident in real-world datasets such as COMPAS [18]. We observed a unique trend: while initial incremental upsampling of minority sub-groups in the dataset leads to improved fairness (as gauged by the Equalized Odds metric, where lower values signify greater fairness), this trend does not consistently hold. As shown in Figure 1, with each upsampling incremental, the fairness begins to decline rather than improve. This pattern is crucial, because, in some cases, attempting to balance sample sizes between groups can paradoxically reduce fairness. Driven by these insights, we developed a novel pre-processing method for bias mitigation. Our approach is distinctively designed to fine-tune the upsampling process, strategically optimizing the balance between minority and majority groups in the dataset to enhance fairness.

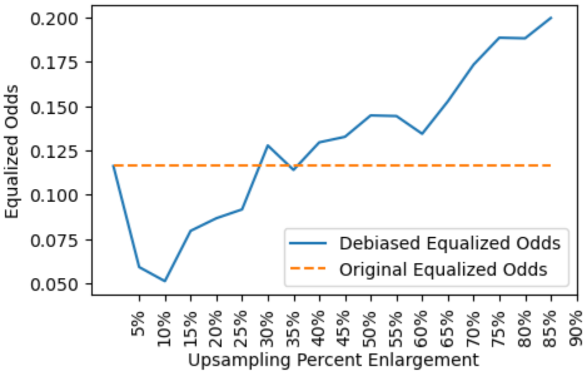


Figure 1. Equalized Odds, which represents the measured bias, is worsen (higher values) in the upsampling process for COMPAS dataset.

While the idea of equalizing group sizes may appear intuitive, it does not always guarantee the most optimal fairness and accuracy outcomes as is exemplified in Figure 1. To better achieve fairness

[1] Equal contribution.

in this context, a more nuanced upsampling approach is needed. By adopting an optimized upsampling strategy, we have been able to strike a balance between fairness and performance, achieving improved model generalization while mitigating bias and promoting fairness effectively.

We have proposed and implemented an upsampling method that employs CTGAN to synthesize samples in a refined manner, aiming to enhance the fairness of the data while maintaining the model's performance. We investigated whether varying ratios between privileged and unprivileged groups can result in enhanced fairness while maintaining the model's performance, compared to mere group size equalization which is the common practice of related work, as far as we know. To undertake this investigation, we delved into a multi-optimization problem, utilizing Tree structure Parzen Estimator, with the specific aim of identifying optimal values that amplify the representation of smaller groups within each label, as demonstrated in Figure 2 with the goal of minimizing the bias of the training dataset while maximizing the accuracy.

1.1 Main contributions of our research

- We introduce 'FairUS' - a novel method, for pre-processing bias mitigation using upsampling with CTGAN synthetic data.
- A notable contribution of our study is the careful and optimized alteration of the original training set. Unlike existing methods that may significantly modify the training data, our approach introduces a fine-tuned amount of synthetic data upsampled. This approach ensures that the modifications serve the specific purpose of enhancing fairness while preserving the underlying structure and quality of the original data.
- The proposed method empowers decision-makers to improve fairness and simultaneously determine their preferred balance with accuracy, allowing for a customized tradeoff between these critical factors.
- Our method can be seamlessly integrated into ML pipelines for model-agnostic applicability. The reproducible code of our method is available on GitHub.
- We demonstrated superior results in fairness metrics and accuracy through evaluations on various canonical datasets, both in comparison to a baseline without bias mitigation and in comparison to related work.

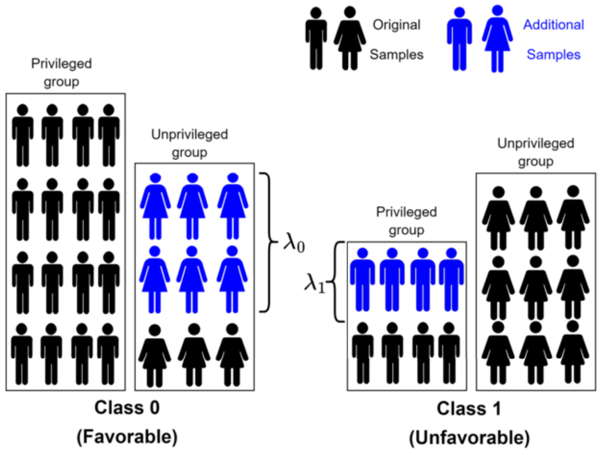


Figure 2. Optimization of the upsampling process per class and sub-group

2 Related Work

In this section, we provide a concise review of the foundational and contemporary work in our research area, primarily focusing on bias detection metrics, and pre-processing bias mitigation methods.

2.1 Bias Detection Metrics

Canton et al.'s comprehensive survey [6] broadly classifies bias metrics into two main categories: outcome probabilities-based metrics, such as Statistical Parity [4] and Disparate Impact [4], and confusion matrix-based metrics, including Equalized Odds, Equalized Opportunity [13], and Accuracy Rate Difference [3]. Our research uses Equalized Odds, a metric that considers the delta of True Positive and False Positive Rates measured in sub-groups, as suggested by Hardt et al. [13]. This metric particularly overcomes the pitfalls of the former bias detection measures. It is important to mention that the metrics of Statistical Parity and Disparate Impact, by their definitions, are inherently enhanced by upsampling techniques. Therefore, they do not present meaningful or insightful measures for evaluating our approach.

2.2 Sampling Pre-processing Bias Mitigation Methods

Bias mitigation methods can be divided into pre-processing, in-processing, and post-processing approaches [15]. Our study focuses on pre-processing methods, mainly balancing groups by sampling [11, 5, 23].

Sampling techniques have been used extensively to address data imbalance, which could lead to biased models. Two recent upsampling methods, FairSMOTE [8] and FairGAN [28], produce synthetic data generation to balance class distribution and mitigate bias. FairSMOTE employs the Synthetic Minority Over-sampling Technique (SMOTE) [9] to equalize group sizes, creating synthetic samples by interpolating minority samples. FairGAN leverages Conditional Tabular GAN (CTGAN) [32] to generate synthetic samples that preserve statistical properties and dependencies in tabular data. A recent study [22] emphasizes the importance of measuring the contribution of each subgroup and its size to the fairness measured, but mitigates this disparity with re-weight samples in an optimized manner to the relevant subgroup.

Our proposed method, FairUS, differs from existing approaches such as FairSMOTE and FairGAN, which aim to directly equalize group sizes. Instead, FairUS employs an optimization-centric strategy to achieve the optimized balance between groups for each label.

By solving this optimization problem, FairUS identifies a balance that enhances both fairness and accuracy. This approach allows FairUS to cater to specific instances of imbalanced data and fairness requirements, leading to improved fairness, accuracy, and better generalization across various real-world applications.

2.3 CTGAN for synthetic data generation

Xu et al. introduced Conditional GAN (CTGAN), a tool that uses Generative Adversarial Networks (GANs) to generate synthetic tabular data by modeling its distribution [32]. CTGAN handles complex distributions of both discrete and continuous data columns through a process called "mode-specific normalization". Each data column is processed separately. For continuous data, a Gaussian mixture model estimates distinct peaks in the distribution, assigns probabilities to each data point from these modes, and normalizes the data point based on a randomly sampled mode. This results in data points represented as a mix of the mode, which is one-hot encoded, and the original value. For discrete columns with imbalanced data, a 'conditional generator' creates synthetic rows based on the logarithmic frequency of each category ('training-by-sampling'). Generating discrete columns data is one of the advantages of CTGAN over other synthetic data generation tools, such as SMOTE or other methods that leverage synthetic data for bias mitigation [28], especially relevant for handling protected attributes, such as gender, religion, marital status or race, which are mostly discrete, by their nature. In our study, we employed CTGAN to produce new synthetic samples for each group.

3 Method

FairUS method aimed at improving fairness in a given input dataset D without compromising the model's performance. This enhancement is realized through an innovative pre-processing multi-objective optimization of upsampling. The upsampling is done utilizing CTGAN technique to synthesize data efficiently.

Our method coordinates two principal algorithms: Upsampling Dataset Algorithm (Algorithm 1), and Finding Optimal Lambdas Algorithm (Algorithm 2). The process initiates with Algorithm 1, tasked with upsampling the dataset via predetermined lambda values, thereby creating new samples through CTGAN. Following this, Algorithm 2 is engaged to refine these lambda values. This refinement is optimized iteratively through the Tree-structured Parzen Estimator (TPE) optimization method, which incorporates Algorithm 1 at each step. The aim is to construct an objective that leverages the outcomes of the upsampling, both throughout the iterative process and in the final analysis. The high level description of the method is illustrated in Figure 3.

The resulting output is an optimized by sub-groups upsampled dataset D'. To achieve fairer results for the protected attribute, Algorithm 1 utilizes λ_0 as the proportion of the sub-groups of the protected attribute within class "0" and λ_1 as the proportion of the sub-groups of the protected attribute within class "1" to balance the dataset with synthesized data utilizing CTGAN. Algorithm 2 focuses on determining the optimal parameters, denoted by λ_0 and λ_1, that enhance the representation of minority groups within each label.

Algorithm 1: Optimized Upsampling Dataset

```
1 Input: $D$ - dataset, $PA$ - Protected Attribute, $PAV$ - Privileged
    Attribute Value, $\lambda_0$, $\lambda_1$ - Lambdas for upsampling
2 Output: $D'$ - Upsampled dataset
  1: $p_1 \leftarrow |\{i \in D \mid \text{PA}(i) = \text{PAV} \wedge \text{label}(i) = 1\}|$
  2: $p_0 \leftarrow |\{i \in D \mid \text{PA}(i) = \text{PAV} \wedge \text{label}(i) = 0\}|$
  3: $np_1 \leftarrow |\{i \in D \mid \text{PA}(i)! = \text{PAV} \wedge \text{label}(i) = 1\}|$
  4: $np_0 \leftarrow |\{i \in D \mid \text{PA}(i)! = \text{PAV} \wedge \text{label}(i) = 0\}|$
  5: $adjustment\_factor_0 \leftarrow \lambda_0 \times \max(0, p_0 - np_0, np_0 - p_0)$
  6: $adjustment\_factor_1 \leftarrow \lambda_1 \times \max(0, p_1 - np_1, np_1 - p_1)$
  7: $syn\_nonpriv\_0 \leftarrow CTGAN_0(adjustment\_factor_0)$
  8: $syn\_nonpriv\_1 \leftarrow CTGAN_1(adjustment\_factor_1)$
  9: $syn\_priv\_0 \leftarrow CTGAN_0(adjustment\_factor_0)$
 10: $syn\_priv\_1 \leftarrow CTGAN_1(adjustment\_factor_1)$
 11: $D' \leftarrow D \cup syn\_nonpriv\_0 \cup syn\_nonpriv\_1 \cup$
     $syn\_priv\_0 \cup syn\_priv\_1$
 12: return $D'$
```

3.1 Algorithm 1 - Upsampling Function

Algorithm 1 aims to balance the protected attribute by upsampling the dataset with synthetic samples. The input to this algorithm include the dataset (D), the protected attribute (PA), the privileged attribute value (PAV), and lambda values λ_0 and λ_1. Initially, the algorithm splits the dataset into different groups based on labels and the privileged attribute value (line 1-4). Subsequently, it calculates the number of samples to generate for each group, considering the λ_0 and λ_1 and the difference in group sizes (line 5-6). Utilizing CTGAN, it generates synthetic samples for the minority group out of the non-privileged and privileged sub-groups separately (line 7-10). Finally, it merges the synthetic samples with the original dataset (line 11), creating an upsampled dataset (D') as the output (line 12). The incorporation of CTGAN, a sophisticated generative modeling technique, facilitates the generation of realistic samples, thereby contributing to the overall effectiveness of the approach. While we use binary values in this example, this method can be easily applied to any discrete protected attribute by applying one-hot encoding.

3.2 Algorithm 2 - Finding Optimized Lambdas for Upsampling Dataset

Algorithm 2 is tasked with the challenge of improving the dataset fairness while simultaneously preserving its accuracy.
This is achieved by meticulously determining the most appropriate λ_0 and λ_1 values. The balancing act between fairness enhancement and accuracy maintenance is moderated by the user by choosing the optimal trial from the TPE results. A trial close to the first one will represent a preference for greater fairness.

The algorithm requires a dataset (D), a protected attribute (PA) with a privileged attribute value (PAV), and N optimization trials.

Initiating with lambda values set to zero ensures a neutral commencement (line 1). The subsequent phase involves the calculation of baseline metrics for Equalized Odds and accuracy, utilizing a Random Forest model for this purpose. These baseline metrics lay the foundation for comparative assessment following lambda adjustments (lines 2-3).

The algorithm then enters a critical phase of iterative exploration, spanning a range of lambda values from 0 to 2 utilizing Tree structure Parzen Estimator (TPE), $Parzen\_T$ (line 5) . In each iteration, for every distinct combination of λ_0 and λ_1, the algorithm generates an upsampled dataset, using Algorithm 1. It then evaluates the efficacy of each lambda pairing through the lens of the fairness and accuracy metrics of the upsampled dataset, juxtaposed against the baseline values (lines 6-10). The scaling of these comparisons is pivotal, providing a proportionate measure of the metrics' deviations from their baseline states.

The result of each optimization trial is updated in the $Parzen\_T$ based on multi optimization objectives (line 12).

Conclusively, the algorithm yields the optimally upsampled dataset from the $Parzen\_T$, identified by the λ_0 and λ_1 values that most effectively conform to the user's balance preferences between fairness and accuracy (line 14-15). Algorithm 2 thus presents a sophisticated, customizable methodology for navigating the often intricate tradeoff between fairness and accuracy in dataset sampling, a challenge in equitable ML models development.

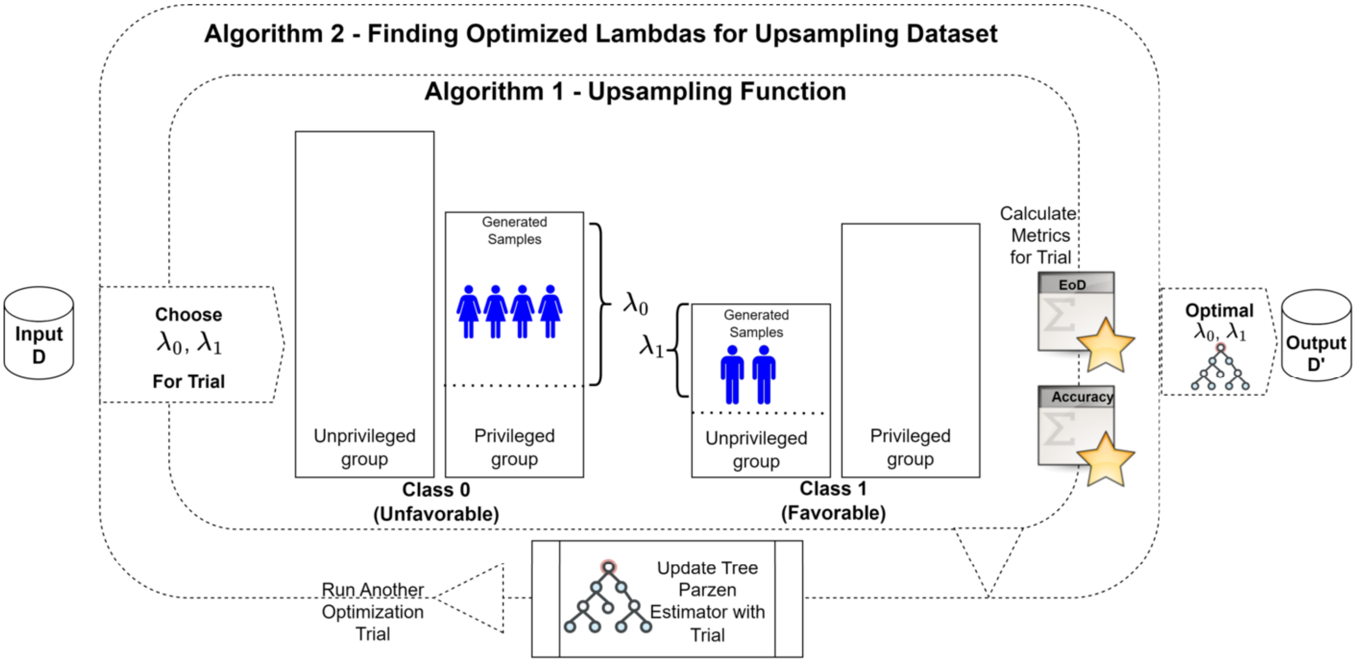

Algorithm 2 - Finding Optimized Lambdas for Upsampling Dataset
Algorithm 1 - Upsampling Function
Input D
Choose λ_0, λ_1 For Trial
Generated Samples
λ_0
λ_1
Generated Samples
Unprivileged group
Privileged group
Class 0 (Unfavorable)
Unprivileged group
Privileged group
Class 1 (Favorable)
Calculate Metrics for Trial
EoD
Accuracy
Optimal λ_0, λ_1
Output D'
Run Another Optimization Trial
Update Tree Parzen Estimator with Trial

Figure 3. High-level description of FairUS: Algorithm 2 finds the optimized lambdas using Algorithm 1 to upsample the training set

```
Algorithm 2: Finding Optimized Lambdas for Upsampling Dataset
Input: $D$ - dataset, $PA$ - Protected Attribute, $PAV$ - Privileged Attribute Value
Parameters: $N$ - number of optimization trials
Output: $D'_{opt}$ - Optimized Dataset with $\lambda_{0_{opt}}, \lambda_{1_{opt}}$ - Optimized upsampling parameters

$Parzen\_T \leftarrow$ Initialize TPE-based optimization model
$OrigFair\_score \leftarrow EoD(D, PA, PAV)$
$OrigAcc\_score \leftarrow Accuracy(D)$
for $i \leftarrow 1$ to $N$ do
    $(\lambda_0, \lambda_1) \leftarrow$ Select parameters using TPE model $Parzen\_T$
    $D' \leftarrow UpsampleDataset(D, PA, PAV, \lambda_0, \lambda_1)$
    $CurrFair\_score \leftarrow EoD(D', PA, PAV)$
    $CurrAcc\_score \leftarrow Accuracy(D')$
    $Scaled\_Fair\_score \leftarrow CurrFair\_score/OrigFair\_score$
    $Scaled\_Acc\_score \leftarrow CurrAcc\_score/OrigAcc\_score$
    $Optimization\_Objective \leftarrow Scaled\_Fair\_score, Scaled\_Acc\_score$
    $Parzen\_T \leftarrow$ Update with $(\lambda_0, \lambda_1, Optimization\_Objective)$
end for
$(\lambda_{0_{opt}}, \lambda_{1_{opt}}) \leftarrow$ Optimal user-defined trial from TPE $Parzen\_T$
$D'_{opt} \leftarrow UpsampleDataset(D, PA, PAV, \lambda_{0_{opt}}, \lambda_{1_{opt}})$
return $D'_{opt}$
```

3.3 Lambdas Optimization with TPE

We apply the TPE algorithm for optimizing lambda parameters, specifically λ_0 and λ_1, to balance fairness and accuracy in predictive models. TPE, an effective hyperparameter optimization technique, is particularly suited for this task due to its iterative and probabilistic approach. It excels in scenarios where optimizing hyperparameters to balance competing objectives, such as accuracy and latency in quantization, is crucial. TPE operates iteratively, using a history of evaluated hyperparameters to construct a probabilistic model that guides the selection of new hyperparameters. The process involves defining a search space domain, creating an objective function, and employing Parzen Estimators for modeling densities based on observed scores.

Our implementation is inherently multi-objective, hence combines equalized odds with model accuracy. This composite measure allows for a nuanced approach to optimizing the upsampling parameters λ_0 and λ_1. The TPE algorithm iteratively adjusts these parameters, assessing their impact on both fairness and accuracy. Each iteration leverages historical performance data to refine the parameters, guiding the optimization towards a balance that minimizes equalized odds and maximize accuracy. The algorithm divides observations into two groups based on scores, models two densities using Parzen Estimators, and samples hyperparameters from these densities. This method aims to identify a set of lambdas that optimally balance the trade-off between fairness and accuracy, ensuring the most equitable representation of data groups while maintaining robust and high predictive quality.

3.4 Pareto Front Optimization

Within the scope of this study, our methodology employs the concept of a Pareto front to examine the trade-offs between fairness and accuracy, two pivotal objectives in the context of bias mitigation in ML datasets. The Pareto front is an essential tool for multi-objective optimization, delineating a set of optimal solutions where any attempt to improve one objective would lead to a deterioration in another. This frontier is critical for our analysis as it captures the inherent trade-offs involved when aiming to enhance dataset fairness without compromising the accuracy of the models.

In deploying the Pareto front, each point along the front represents a distinct combination of fairness and accuracy, achieved through our bias mitigation strategy. This graphical representation allows us to systematically explore the efficiency frontier of these two competing objectives. By comparing various points along the Pareto front, we can discern the extent to which improvements in fairness may influence the accuracy of the models and vice versa.

3.5 Intuition for the Method Effectiveness

As discussed in the related work, the notion of fairness can be quantified by the Equalized Odds metric, examining the difference in TPRs and FPRs across different groups [13], and expressed as:

$$EoD^{-} = \frac{1}{2}(|TPR_{PA=0} - TPR_{PA=1}| + |FPR_{PA=0} - FPR_{PA=1}|) \quad (1)$$

Here, the term PA represents the protected attribute. For simplicity, we consider PA to have two distinct values: 0,1, one represents the unprivileged group, and the other denotes the privileged group. In this equation, TPR represents the true positive rate, and FPR represents the false positive rate.

A classifier achieves equalized odds if both TPR and FPR are the same between groups, which means that it is equally accurate for all groups. Any deviation from zero in the EoD^{-} value indicates a disparity in the classifier's performance between the two groups, signaling potential unfairness.

Suppose the classifier's False Positive Rate ($FPR = \frac{FP}{FP+TN}$) for group $PA = 1$ is greater than that for group $PA = 0$:

$$FPR_{PA=1} > FPR_{PA=0} \quad (2)$$

This indicates that group $PA = 1$ experiences an unfair situation of more False Positives relative to class 1, which increases the EoD^{-}.

Similarly, suppose the classifier's True Positive Rate ($TPR = \frac{TP}{TP+FN}$) for group $PA = 1$ is greater than that for group $PA = 0$:

$$TPR_{PA=1} > TPR_{PA=0} \quad (3)$$

This indicates that group $PA = 0$ experiences an unfair situation of more False Negatives relative to class 0, which increases the EoD^{-}.

The EoD^{-} is reduced when a classifier's predictive performance, specifically TPR and FPR, become more homogeneous across protected and non-protected groups. The Upsampling Dataset Algorithm 1 addresses this by correcting imbalances within the training data, which are often the cause of unequal classifier performance metrics.

Balancing Representation to Reduce False Negatives (FN) When a protected group is underrepresented in the positive class, there is an increased risk that the classifier predicts more negative outcomes for this group, leading to a higher FN rate. Upsampling the positive class for the protected group using λ_1 augments its presence in the training data. This enriched representation equips the classifier with a broader spectrum of positive examples from the protected group, improving its ability to recognize such cases accurately and thus reducing FN rates.

Balancing Exposure to Reduce False Positives (FP) Similarly, if the protected group is underrepresented in the negative class, the classifier may incorrectly predict more positive outcomes, resulting in a higher FP rate. Upsampling the negative class for the protected group using λ_0 bolsters the number of negative examples from this group within the training set. This expanded exposure helps the classifier to better understand the characteristics of negative cases within the protected group, decreasing the propensity for FPs.

The strategy of upsampling to reduce EoD^{-} is predicated on the construction of a training set that mirrors a more equitable distribution of outcomes for the protected group. By adjusting the dataset to provide the classifier with a balanced learning environment, the performance metrics between the protected and non-protected groups are harmonized. This leads to a reduction in EoD^{-} values and contributes to the advancement of fairness in ML algorithms.

4 Experiments

The experiments that were conducted, compared EoD^{-} and accuracy of our method against a baseline without any bias mitigation, and against two related works, namely FairSMOTE and FairGAN.

4.1 Datasets

We have selected eight well-known tabular datasets from various domains: COMPAS [2], ADULT [17], German Credit [10], Bank Marketing [26], Diabetes [2], Utrecht Fairness Recruit[3], Nursery [27] and Default Credit[4] as summarized in Table 1, due to their relevance in ML research and the presence of fairness concerns [19]. These datasets contain various attributes that can potentially introduce biases and discrimination, such as race, occupation, age, and gender. Pre-processing of these datasets included: the removal of rows with missing values and the encoding of categorical attributes.

4.2 Experimental Setup

We performed experiments to evaluate the performance of our proposed method, FairUS, in terms of fairness and accuracy. We applied FairUS to the training data, generating an upsampled training set. We then trained a model on this upsampled training set and evaluated its evaluation metrics as discussed next.

[2] Medical dataset that was published in 2023 on Kaggle. The data set is created by Mohammed Mustafa.

[3] Students' admission to university dataset that was published in 2022 on Kaggle. The data set is created by Sieuwert van Otterloo.

[4] Contains information on default payments, demographic factors, credit data, history of payment, and bill statements of credit card clients in Taiwan from April 2005 to September 2005.

Dataset	#Sample	Protected	Privileged
COMPAS	7,217	Race Sex	Caucasian Male
ADULT	32,561	Race Sex	Caucasian Male
GERMAN CREDIT	1,000	Sex	Male
BANK MARKETING	11,162	Age	25-65
RECRUIT	4,000	Sex	Male
DIABETES	99,982	Gender	Female
NURSERY	12,960	Parents Occupation	Great_pret
Default Credit	30,000	Sex	Male

Table 1. Experiments' datasets with a variety of protected attributes.

4.2.1 Models Setup

To calculate the Equalized Odds, we used Random Forest from Sklearn [29] library, with the default parameters. We used the "Synthetic Data Vault Project" (SDV) CTGAN implementation[5] with default parameters. We used the "Optuna" Python package[6] for TPE implementation, with the following settings n_startup_trials=20, n_ei_candidates=24, multivariate=True, n_trials=200. For the multi-objective, the directions were defined as "minimize" for Equalized Odds and "maximize" for accuracy.

4.3 Comparison Methods

We compared our method against three alternatives:

- **Baseline:** A model trained without any bias mitigation. This provides a reference point to gauge the improvement brought by the other methods as well as FairUS.
- **FairSMOTE:** A well-known pre-processing method that uses a variant of SMOTE for fair upsampling.
- **FairGAN:** A method that employs Generative Adversarial Networks for fairness-enhanced data generation.

4.4 Evaluation Metric

In our experiments, we selected the Equalized Odds metric to evaluate our method's effectiveness, as detailed in the Related Work and Method sections. Our research has dual optimization goals: (1) to minimize the value of the Equalized Odds metric, as indicated by Equation 1, and (2) to maximize the accuracy of the test set. It should be noted that our method is suitable for any fairness metric and can be easily adjusted accordingly. It should be noted that our method can support any fairness metric as a parameter to the multi-objective optimization algorithm.

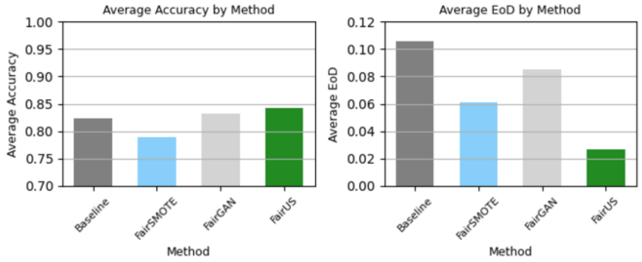
5 Results

The results of our experiments are comprehensively presented in Table 2. We demonstrate the efficacy of our method, FairUS. For each dataset and protected attribute, we compare the performance of our method against a baseline without bias mitigation, and against related work - FairSMOTE, and FairGAN. The metrics used for comparison are Accuracy and EoD^{-}. We show the superiority of FairUS in enhancing fairness in nine out of nine experiments, all of which on the first trial. Interestingly, when focusing on accuracy, FairUS maintained its leading position in eight out of nine experiments, three of which were attributed to the first trial as well. Five experiments, where FairUS showcased the highest accuracy, were part of advanced trials. This illustrates the nuanced capability of FairUS to navigate the fairness-accuracy landscape effectively.

These findings underscore the Pareto front's conceptual importance in our analysis. By demonstrating that FairUS can either lead or closely align with the optimal points on the Pareto front, we highlight its utility in balancing the dual objectives of fairness and accuracy.

On average, FairUS presented **enhanced fairness** and improved accuracy as shown in Figure 4, indicating its effectiveness in promoting fairness.

Figure 4. Comparison of average accuracy and EoD^{-} across all methods



Furthermore, our results validate the capability of FairUS to allow decision-makers to effectively balance fairness and accuracy. This is illustrated in Figure 5, where FairUS exhibits superior performance over other methods in the accuracy-fairness space, presenting a Pareto-Dominate of fairness-accuracy combinations. FairUS consistently achieved a lower Equalized odds score, indicating enhanced fairness, while maintaining competitive accuracy. During the first trials, all FairUS experiments demonstrated effective reduction of EoD^{-} scores while maintaining or enhancing accuracy.

A key contribution of our method is the optimized alteration of the original training set. This approach is evident in our results, showing that FairUS has the least amount of data upsampled, hence, preserving the integrity of the original data. This is reflected in the balanced dataset growth across all experiments, as detailed in Table 2 and strengthening our study's motivation as presented in Figure 1.

5.1 Discussion

- **Lower STD**: FairUS presents lower STD averages of 0.0076 compared to 0.0111, as calculated from Table 2 for the Equalized Odds and similar STD for the accuracy, suggesting enhanced stability in its performance compared to benchmark methods.
- **Trade-off**: As illustrated in Figure 5, FairUS trials consistently demonstrated superior performance (indicated by light blue bullets) over other methods, and creating a Pareto-Dominate, achieving higher accuracy and lower bias. For all experiments, the first

[5] https://github.com/sdv-dev/CTGAN
[6] https://optuna.org/

Dataset	Protected	Method	Accuracy (± STD)	EoD^- (± STD)	Dataset Growth	Time (min)
COMPAS	Race	Baseline	0.682 ± 0.014	0.278 ± 0.059	0%	0
		FairSMOTE	0.674 ± 0.010	0.154 ± 0.034	19.83%	0.662
		FairGAN	0.691 ± 0.010	0.130 ± 0.016	19.83%	4.262
		FairUS (trial = 1st)	**0.712 ± 0.013**	**0.029 ± 0.014**	20.54%	4.38
ADULT	Race	Baseline	0.847 ± 0.003	0.108 ± 0.037	0%	0
		FairSMOTE	0.797 ± 0.007	0.063 ± 0.005	167.92%	3.925
		FairGAN	0.891 ± 0.003	0.119 ± 0.006	167.92%	31.426
		FairUS (trial = 1st)	0.879 ± 0.003	**0.032 ± 0.009**	32.69%	12.64
		FairUS (trial = 5th)	**0.895 ± 0.002**	0.071 ± 0.022	55.61%	12.64
ADULT	Sex	Baseline	0.847 ± 0.003	0.125 ± 0.042	0%	0
		FairSMOTE	0.747 ± 0.004	0.041 ± 0.017	85.45%	1.628
		FairGAN	0.899 ± 0.003	0.117 ± 0.003	85.45%	31.59
		FairUS (trial = 1st)	0.878 ± 0.002	**0.038 ± 0.004**	37.75%	11.02
GERMAN	Sex	Baseline	0.727 ± 0.013	0.065 ± 0.034	0%	0
		FairSMOTE	0.747 ± 0.024	0.071 ± 0.041	99.60%	0.398
		FairGAN	0.781 ± 0.019	0.073 ± 0.040	99.60%	2.279
		FairUS (trial = 1st)	0.769 ± 0.005	**0.044 ± 0.021**	59.4%	3.41
		FairUS (trial = 6th)	**0.818 ± 0.009**	0.068 ± 0.026	57.8%	3.41
BANK	Age	Baseline	0.832 ± 0.005	0.227 ± 0.085	0%	0
		FairSMOTE	0.790 ± 0.007	0.043 ± 0.006	104.90%	1.097
		FairGAN	0.831 ± 0.006	0.127 ± 0.011	104.90%	4.424
		FairUS (trial = 1st)	0.823 ± 0.007	**0.021 ± 0.012**	23.16%	7.47
		FairUS (trial = 6th)	**0.838 ± 0.008**	**0.068 ± 0.019**	79.51%	7.47
RECRUIT	Age	Baseline	0.766 ± 0.009	0.061 ± 0.028	0%	0
		FairSMOTE	0.791 ± 0.014	0.036 ± 0.013	91.6%	1.39
		FairGAN	0.761 ± 0.015	0.045 ± 0.020	91.6%	3.59
		FairUS (trial = 1st)	**0.793 ± 0.008**	**0.029 ± 0.010**	54.47%	5
DIABETES	Gender	Baseline	0.970 ± 0.001	0.007 ± 0.006	0%	0
		FairSMOTE	0.902 ± 0.005	0.014 ± 0.007	116.3%	9.61
		FairGAN	0.909 ± 0.003	0.056 ± 0.006	116.3%	65.45
		FairUS (trial = 1st)	**0.975 ± 0.002**	**0.007 ± 0.003**	22.35%	25.21
NURSERY	Parents Occupation	Baseline	0.932 ± 0.007	0.067 ± 0.015	0%	0
		FairSMOTE	0.915 ± 0.004	0.092 ± 0.010	104.1%	2.59
		FairGAN	0.869 ± 0.004	0.076 ± 0.010	104.1%	3.67
		FairUS (trial = 1st)	0.922 ± 0.003	**0.039 ± 0.020**	9.13%	5.28
		FairUS (trial = 10th)	**0.934 ± 0.005**	**0.075 ± 0.006**	3.85%	5.28
Default Credit	Sex	Baseline	0.815 ± 0.003	0.017 ± 0.008	0%	0
		FairSMOTE	0.742 ± 0.003	0.042 ± 0.004	91.23%	2.72
		FairGAN	0.858 ± 0.005	0.033 ± 0.006	91.23%	29.58
		FairUS (trial = 1st)	0.830 ± 0.004	**0.007 ± 0.009**	10.06%	17.2
		FairUS (trial = 9th)	**0.859 ± 0.002**	0.018 ± 0.012	34.77%	17.2

Table 2. Experiments results show that in all experiments, FairUS had the best EoD^- and Accuracy measured in comparison to baseline without bias mitigation and related work. Only in one experiment FairGAN had better Accuracy with worsened EoD^- than the FairUS.

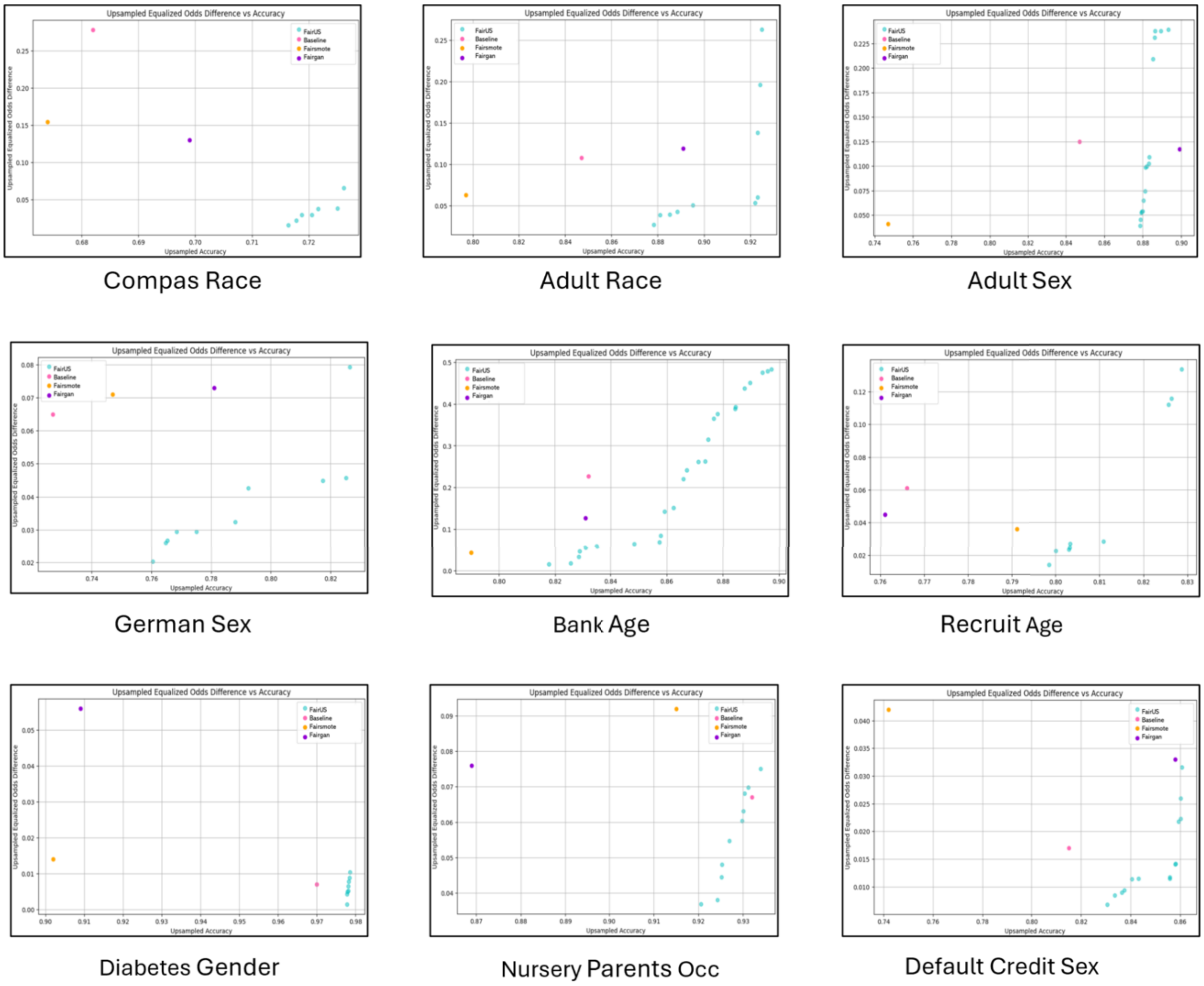
trial yielded maximal fairness, underscoring our initial motivation. While choosing different trials gradually shifts the focus towards enhancing accuracy, thus creating a more balanced optimization between these two objectives.

- **Optimized Synthetic Data Generation**: FairUS generated the least amount of additional synthetic data compared to other methods in all experiments (but, COMPAS with race, which has similar dataset growth), as shown in Figure 6. We observed that benchmark methods had, in average, over three times more data than FairUS (29.95% vs. 97.89%), showcasing FairUS efficient data usage, as calculated from Table 2.
- **Running Time**: FairUS shows an increased running time, 4.09 times longer than FairSMOTE due to parameter optimization, but 0.51 times shorter than FairGAN, attributed to less upsampled data. This one-time optimization in the model development lifecycle is deemed acceptable for a pre-processing method.

6 Conclusions

This research addressed the pivotal issue of bias in ML models, particularly those originating from training data. We developed FairUS, a novel pre-processing technique that utilizes CTGAN for optimized upsampling, as opposed to merely equalizing group sizes. Empirical evaluations conducted on several canonical datasets demonstrated the effectiveness of FairUS. It not only mitigated bias and improved fairness but also preserved the accuracy. Moreover, compared to benchmarks, FairUS displayed superior performance. The results validate FairUS as an effective and practical tool for bias mitigation across various domains and contexts. The **main limitation of FairUS** lies in its running time complexity, related to the chosen number of trials for the optimization process (N) in Algorithm 2. However, this limitation is acceptable, given that the method is applied only once in a pre-processing manner. In terms of **future research directions**, we aim to adapt FairUS to handle different do-

Figure 5. Comparison of EoD^{-} vs. accuracy between methods across different datasets and protected attributes. Different trials of FairUS presented the Pareto-Dominate



Compas Race

Adult Race

Adult Sex

German Sex

Bank Age

Recruit Age

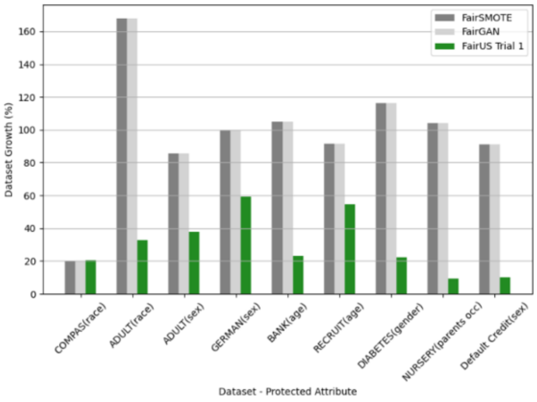
Diabetes Gender

Nursery Parents Occ

Default Credit Sex

mains such as computer vision training sets. We believe that our work lays a solid foundation for further exploration in the field of AI fairness and cultivating more equitable AI systems.

Figure 6. Comparison of dataset growth between the methods. FairUS that generated the least amount of data in comparison to other methods.



7 Data and Code Availability

The reproducible code of our method is available on: https://github.com/GuyRozenblatt/FairUS

Acknowledgement

This work was supported by the Israel National Cyber Directorate.

References

[1] A. Abusitta, E. Aïmeur, and O. A. Wahab. Generative adversarial networks for mitigating biases in machine learning systems. *arXiv preprint arXiv:1905.09972*, 2019.

[2] J. Angwin, J. Larson, S. Mattu, and L. Kirchner. Machine bias risk assessments in criminal sentencing. *ProPublica, May*, 23, 2016.

[3] R. Berk, H. Heidari, S. Jabbari, M. Kearns, and A. Roth. Fairness in criminal justice risk assessments: The state of the art. *Sociological Methods & Research*, 50(1):3–44, 2021.

[4] T. Calders and S. Verwer. Three naive bayes approaches for discrimination-free classification. *Data mining and knowledge discovery*, 21:277–292, 2010.

[5] F. Calmon, D. Wei, B. Vinzamuri, K. Natesan Ramamurthy, and K. R. Varshney. Optimized pre-processing for discrimination prevention. *Advances in neural information processing systems*, 30, 2017.

[6] S. Caton and C. Haas. Fairness in machine learning: A survey. *arXiv preprint arXiv:2010.04053*, 2020.

[7] J. Chai and X. Wang. Self-supervised fair representation learning without demographics. *Advances in Neural Information Processing Systems*, 35:27100–27113, 2022.

[8] J. Chakraborty, S. Majumder, and T. Menzies. Bias in machine learning software: Why? how? what to do? In *Proceedings of the 29th ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering*, pages 429–440, 2021.

[9] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer. Smote: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, 16:321–357, 2002.

[10] D. Dheeru and E. K. Taniskidou. Uci machine learning repository. http:« archive. ics. uci. edu< ml. *arXiv preprint arXiv:1710.11342*, 2017.

[11] M. Feldman, S. A. Friedler, J. Moeller, C. Scheidegger, and S. Venkatasubramanian. Certifying and removing disparate impact. In *proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining*, pages 259–268, 2015.

[12] M. Gupta, A. Cotter, M. M. Fard, and S. Wang. Proxy fairness. *arXiv preprint arXiv:1806.11212*, 2018.

[13] M. Hardt, E. Price, and N. Srebro. Equality of opportunity in supervised learning. *Advances in neural information processing systems*, 29, 2016.

[14] K. Holstein, J. Wortman Vaughan, H. Daumé III, M. Dudik, and H. Wallach. Improving fairness in machine learning systems: What do industry practitioners need? In *Proceedings of the 2019 CHI conference on human factors in computing systems*, pages 1–16, 2019.

[15] M. Hort, Z. Chen, J. M. Zhang, F. Sarro, and M. Harman. Bia mitigation for machine learning classifiers: A comprehensive survey. *arXiv preprint arXiv:2207.07068*, 2022.

[16] F. Kamiran and T. Calders. Classifying without discriminating. In *2009 2nd international conference on computer, control and communication*, pages 1–6. IEEE, 2009.

[17] R. Kohavi et al. Scaling up the accuracy of naive-bayes classifiers: A decision-tree hybrid. In *Kdd*, volume 96, pages 202–207, 1996.

[18] J. Larson, S. Mattu, L. Kirchner, and J. Angwin. How we analyzed the compas recidivism algorithm. *ProPublica (5 2016)*, 9(1):3–3, 2016.

[19] T. Le Quy, A. Roy, V. Iosifidis, W. Zhang, and E. Ntoutsi. A survey on datasets for fairness-aware machine learning. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, page e1452, 2022.

[20] P. Li and H. Liu. Achieving fairness at no utility cost via data reweighing with influence. In *International Conference on Machine Learning*, pages 12917–12930. PMLR, 2022.

[21] T. Li, Z. Tang, T. Lu, and X. M. Zhang. 'propose and review': Interactive bias mitigation for machine classifiers. *Available at SSRN 4139244*, 2022.

[22] T. Li, Z. Li, A. Li, M. Du, A. Liu, Q. Guo, G. Meng, and Y. Liu. Fairness via group contribution matching. In *International Joint Conference on Artificial Intelligence*, 2023. URL https://api.semanticscholar.org/CorpusID:26...

[23] K. Lum and J. Johndrow. A statistical framework for fair predictive algorithms. *arXiv preprint arXiv:1610.08077*, 2016.

[24] B. T. Luong, S. Ruggieri, and F. Turini. k-nn as an implementation of situation testing for discrimination discovery and prevention. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 502–510, 2011.

[25] N. Mehrabi, F. Morstatter, N. Saxena, K. Lerman, and A. Galstyan. A survey on bias and fairness in machine learning. *arXiv preprint arXiv:1908.09635*, 2019.

[26] S. Moro, P. Cortez, and P. Rita. A data-driven approach to predict the success of bank telemarketing. *Decision Support Systems*, 62:22–31, 2014.

[27] M. Olave, V. Rajkovic, and M. Bohanec. An application for admission in public school systems. *Expert Systems in Public Administration*, 1: 145–160, 1989.

[28] A. M. Patrikar, A. Mahenthiran, and A. Said. Leveraging synthetic data for ai bias mitigation. In *Synthetic Data for Artificial Intelligence and Machine Learning: Tools, Techniques, and Applications*, volume 12529, pages 174–179. SPIE, 2023.

[29] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, et al. Scikit-learn: Machine learning in python. *the Journal of machine Learning research*, 12:2825–2830, 2011.

[30] M. Rateike, A. Majumdar, O. Mineeva, K. P. Gummadi, and I. Valera. Don't throw it away! the utility of unlabeled data in fair decision making. In *2022 ACM Conference on Fairness, Accountability, and Transparency*, pages 1421–1433, 2022.

[31] C. Wu, F. Wu, T. Qi, and Y. Huang. Semi-fairvae: Semi-supervised fair representation learning with adversarial variational autoencoder. *arXiv preprint arXiv:2204.00536*, 2022.

[32] L. Xu, M. Skoularidou, A. Cuesta-Infante, and K. Veeramachaneni. Modeling tabular data using conditional gan. *Advances in Neural Information Processing Systems*, 32, 2019.

[33] R. Zemel, Y. Wu, K. Swersky, T. Pitassi, and C. Dwork. Learning fair representations. In *International conference on machine learning*, pages 325–333. PMLR, 2013.

[34] Y. Zhang, F. Zhou, Z. Li, Y. Wang, and F. Chen. Fair representation learning with unreliable labels. In *International Conference on Artificial Intelligence and Statistics*, pages 4655–4667. PMLR, 2023.