Mitigating Algorithmic Bias Through Sampling: The Role of Group Size and Sample Selection

Maliheh Heidarpour Shahrezaei(✉), Róisín Loughran, and Kevin Mc Daid

RSRC, Dundalk Institute of Technology, Dundalk, Ireland
{Maliheh.heidarpour,Roisin.Loughran,kevin.mcdaid}@dkit.ie

Abstract. This study proposes a structured framework for mitigating algorithmic bias through sampling-based preprocessing techniques, with particular attention to the roles of group size adjustment and sample selection strategies. We focus on SMOTE-based methods and introduce a 3×3 matrix to categorize bias mitigation techniques. This matrix combines three group size strategies, Equalized Representation, UP-Focused Equalized Representation, and Balanced, group sizes and three sample selection strategies. This framework enables systematic evaluation of each technique's impact on fairness metrics, including Demographic Parity and Equalized Odds, as well as predictive performance. Evaluations across ten diverse datasets show that methods focusing on the unprivileged positive group and leveraging decision-boundary-aware sampling yield significant fairness improvements without substantial accuracy loss. These results highlight the efficacy of targeted oversampling strategies in achieving equitable outcomes in machine learning applications. State-of-the-art methods like preferential sampling continue to excel in optimizing Demographic Parity, while uniform sampling remains superior for achieving Equalized Odds.

Keywords: Bias · Fairness · Mitigation Techniques · Sampling Techniques

1 Introduction

Unwanted algorithmic bias in Machine Learning (ML) systems can result in unfair treatment of certain demographic groups, particularly in classification tasks [1]. These biases can lead to discriminatory outcomes based on sensitive attributes like race, gender, or socioeconomic status [2]. Such disparities often stem from historical prejudices embedded in training data, underrepresentation of certain groups, or design choices within algorithms themselves [3]. In an interesting study [4], the critical role of data preprocessing in mitigating algorithmic bias was highlighted, emphasizing that biases present in training data can lead to discriminatory outcomes in ML systems. Similar studies also categorize preprocessing techniques into methods such as label modification, sampling, and feature modification, noting that these approaches aim to adjust the data distribution to promote fairness [4–6]. These studies underscore the importance of selecting appropriate preprocessing strategies tailored to specific bias scenarios to enhance the fairness of ML models [7]. Prior research has broadly explored preprocessing sampling

A. Bellogin et al. (Eds.): IR4U2 2025/BIAS 2025, CCIS 2786, pp. 30–44, 2026.
https://doi.org/10.1007/978-3-032-12717-4_3

interventions, ranging from reweighting, random over/under-sampling, and SMOTE, to fairness-aware oversampling techniques based on generative models and causal inference frameworks [8–13]. However, Friedler et al. [14] demonstrate that fairness-enhancing interventions through these techniques can behave inconsistently across datasets and fairness definitions. This variability underscores the need for more structured evaluations of sampling-based techniques.

This paper addresses this gap by exposing how adjusting group sizes during the preprocessing stage, specifically within sampling-based bias mitigation techniques, can influence fairness outcomes. To do this, we introduce a framework for bias mitigation that jointly examines two critical preprocessing dimensions: group size adjustment and sample selection strategies. Our approach is built around a 3×3 matrix that systematically combines three methods of group size adjustment with three sampling strategies. We evaluate these effects using a range of fairness metrics, focusing primarily on Demographic Parity while also considering Equalized Odds to assess whether these techniques inadvertently introduce new forms of bias. By doing so, we provide a more comprehensive view of fairness-performance trade-offs. The remainder of this paper is organized as follows: Sect. 2 delves into foundational concepts, fairness metrics, and an analysis of dataset imbalances pertinent to our study. Section 3 details the methodology employed, highlighting the novel preprocessing techniques developed. Section 4 presents an analysis of the results, evaluating the effectiveness of the proposed methods. Finally, Sect. 5 concludes the paper with key findings and offers suggestions for future research directions. To support reproducibility and transparency, the source code and datasets used in this study are publicly available at: https://github.com/MaliHeidarpourSh/Group_size.

2 Fairness Evaluation Framework

2.1 Concept of Bias and Discrimination

Bias, in the context of decision-making, is an inherent tendency or inclination that influences the way information is interpreted or decisions are made [15]. It is a necessary element for classification and differentiation between instances. Bias allows systems, whether human or machine, to make distinctions and categorizations based on various features or criteria [16]. Discrimination, on the other hand, refers to the adverse effects or unfair treatment that can result from bias [16]. In other words, discrimination occurs when biased decisions lead to unequal or unfavorable outcomes for certain individuals or groups [15]. In ML, discrimination can be measured as a difference in the probability of receiving a favorable outcome (positive classification rates) between privileged and unprivileged groups [17].

2.2 Metrics

Fairness in ML is often associated with principles of non-discrimination and equitable treatment [16, 18]. However, fairness is a broad social and ethical concept that cannot be fully captured by any single formal definition [19]. In practice, to mitigate unwanted algorithmic bias, researchers use fairness metrics, quantitative tools that approximate

specific notions of fairness within ML systems [14, 20]. These metrics provide operational tools to evaluate disparities between groups but only reflect particular fairness definitions, highlighting the complexity and context-dependency of fairness in real-world applications [1, 14, 21]. There are two popular categories of group-based concepts of fairness: Demographic Parity and Equalized Odds [22]. These metrics aim to evaluate disparities between privileged and unprivileged groups based on protected attributes. Fairness can be evaluated in the context of unequal distribution of different groups in the training set of the model by Demographic Parity metrics which originates from discrimination-aware modeling practices. Specifically, these fairness metrics examine whether different demographic groups receive equal treatment in terms of favorable outcomes [23]. Disparate Impact (DI) and Statistical Parity (SP) are two widely used fairness metrics in this concept of fairness.

SP: In this metric the likelihood of a positive outcome (Y = 1) should be the same for the privileged group (S = 1) and unprivileged group (S = 0) [24] therefore the ideal value for that is 0 as shown in Eq. 1.

$$SP = P(Y = 1|S = 0) - P(Y = 1|S = 1) \tag{1}$$

DI: This metric resembles SP but instead of using the difference, the ratio is taken [25]. Therefore, according to Eq. 2 the ideal value for this metric is 1.

$$\text{DI} = \frac{P(Y = 1|S = 0)}{P(Y = 1|S = 1)} \tag{2}$$

In this paper we aim to improve fairness with respect to the above metrics. However, we consider other group-based fairness strategies, known as equalized odds, to assess whether these techniques inadvertently introduce new forms of bias. By doing so, we provide a more comprehensive view of fairness-performance trade-offs. Equalized Odds is a fairness criterion that assesses whether a ML model's predictions are equally accurate across different demographic groups [26]. In this context, two widely used metrics are Average Odds Difference (AOD) and Equal Opportunity Difference (EO).

AOD: This metric quantifies the average disparity in both the true positive rate and false positive rate between unprivileged and privileged groups [27]. The ideal result is zero.

EO: This measure ensures everyone is treated similarly and satisfies the same requirements [28]. It mandates that the privileged and unprivileged groups should have similar true positive rates. The ideal result is zero.

By utilizing both Demographic Parity and Equalized Odds metrics, a more comprehensive assessment of bias mitigation techniques can be achieved. The trade-off between fairness and predictive performance remains a critical consideration, as increasing fairness often results in a decline in model accuracy, a phenomenon extensively discussed in the algorithmic fairness literature [19, 29].

Accuracy (Acc) is the most common performance measure which calculates the number of correct predictions divided by the number of total predictions [30]. However, in dealing with an imbalanced dataset, using Acc alone may not be sufficient. **Balanced Accuracy (BAC)**, which calculates the average of sensitivity and specificity, offers a

more robust evaluation [31]. BAC provides a more reliable evaluation metric for imbalanced datasets by considering both the true positive rate and true negative rate [31]. Notably, Acc and BAC will yield the same value if the dataset is balanced.

2.3 Group Definitions

Each sample in the training set is categorized based on its protected attribute (privileged or unprivileged) and class label (positive or negative) [8, 32]. For instance, the Unprivileged Positive (UP) group is defined according to Eq. 3. Similarly, the other groups, Privileged Positive (PP), Unprivileged Negative (UN), and Privileged Negative (PN), follow the same structural definition.

$$\mathrm{UP} = (\mathrm{S} = 0,\ \mathrm{Y} = 1) \tag{3}$$

If the dataset includes more than one protected attribute, to mitigate bias toward multiple protected attributes simultaneously the number of groups will be increased [33, 34]. For instance, combining two protected attributes with the class label results in eight distinct demographic groups. Addressing bias in this manner is crucial for ensuring fairness in ML models [35], as focusing on a single attribute may overlook complex interdependencies between different attributes, potentially leading to unintended discriminatory outcomes [36].

2.4 Dataset

In our study, we utilized ten tabular datasets widely used in fairness-aware ML research [37]. These datasets span diverse domains such as finance, healthcare, education, and criminology, and each includes at least one protected attribute, making them suitable for evaluating bias mitigation strategies. Notably, five of these datasets feature multiple protected attributes. To systematically analyze the datasets, we categorized them based on four key types of imbalances, using threshold-based criteria to assess severity levels:

Class Label Imbalance: refers to the disproportionate distribution between positive and negative class labels. A dataset is considered highly imbalanced if one class makes up more than 70% of the data, moderately imbalanced if it falls between 60%–70%, and less imbalanced if both classes represent at least 40% of the data [38].

Protected Attribute Imbalance: captures unequal representation among different groups defined by a sensitive attribute. If the majority group exceeds 80% of the population, the imbalance is high; if it falls between 65%–80%, it is moderate; and if it is under 65%, the distribution is considered balanced.

Group-Based Imbalance: accounts for disparities in the sizes of subgroups formed by intersecting class labels and protected attributes. This is evaluated as High, when there are substantial differences in group sizes; Moderate, when group size disparities are noticeable but less extreme; Low, when group sizes are relatively uniform.

Positive Class Label Ratio Imbalance: Measures differences in the rates of positive outcomes across protected groups. A difference greater than 20% indicates high imbalance, 10%–20% is moderate, and below 10% is considered low.

Based on these criteria, we assigned each dataset–attribute combination an imbalance level of High, Moderate, or Low, treating each as a distinct evaluation scenario. These results are summarized in Fig. 1, which presents a heatmap illustrating the imbalance levels across all four dimensions for each dataset.

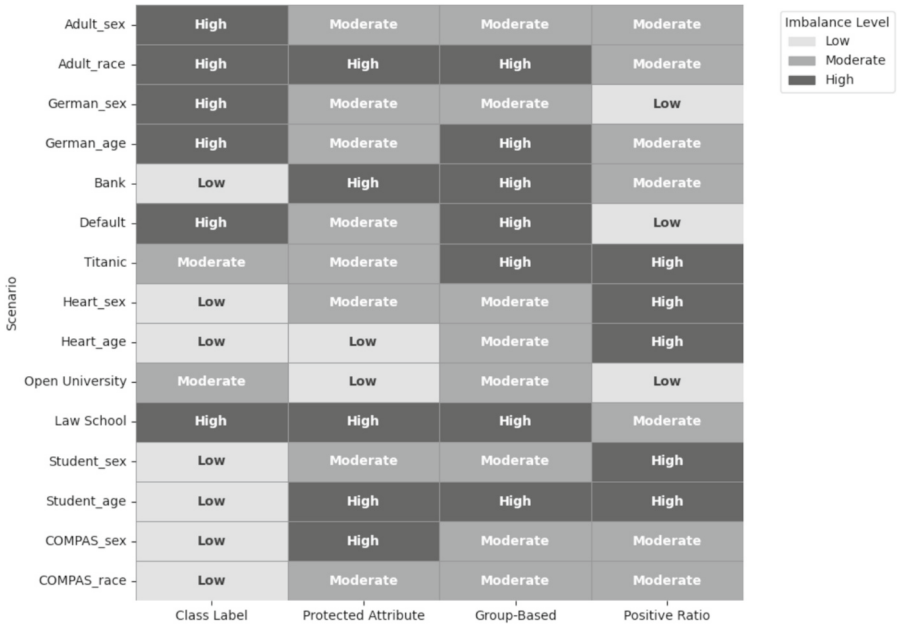

Fig. 1. Class label, protected attribute, group and positive ratio imbalanced level of Scenario

3 Methodology

In this study, we categorize preprocessing sampling methods based on two primary dimensions: group size adjustment strategies and sample selection techniques**.** This dual taxonomy facilitates a comprehensive analysis of how different configurations impact fairness in classification tasks.

3.1 Group Size Adjustment Strategies

In terms of group size adjustment strategies, we identify three principal strategies for adjusting group sizes:

- **Equalized Representation:**

This strategy aims to balance the ratio of positive to negative samples across both privileged and unprivileged groups, in line with the principles of Demographic Parity. The expected size for each subgroup is calculated to ensure that the proportion of positive outcomes is equal across all groups [8, 39]. For the UP group, the expected size is computed in Eq. 4. Similar calculations are applied for the PP, PN, and UN groups by

substituting the corresponding group counts into the formula. This method maintains the overall dataset size while achieving equal positive-to-negative ratios across all groups, thereby promoting fairness without altering the total number of samples.

$$Expected\_Size_{UP} = \frac{|PositiveClass| * |UnprivilegedSamples|}{|TrainingSet|} \tag{4}$$

- **UP -Focused Equalized Representation:**

This method focuses on oversampling the UP group to adjust the dataset in a manner that aims to reduce disparities in model predictions, particularly those measured by the Disparate Impact metric in Eq. 2 [13, 40]. The expected size for the UP group is determined by Eq. 5.

$$Expected\_Size_{UP} = \frac{|PP| * |UN|}{|PN|} - |UP| \tag{5}$$

This approach involves oversampling the UP group, which will increase the total number of positive samples, unprivileged samples, and the overall training set size. For scenarios involving two protected attributes, resulting in eight distinct groups, the expected sizes for groups containing at least one unprivileged attribute will be calculated.

This strategy aims to create a dataset that, when used to train a predictive model, may lead to outcomes with reduced disparate impact, thereby promoting fairness in the model's predictions.

- **Balanced Size Techniques:**

Inspired by the Fair_SMOTE method, this strategy adjusts all subgroup sizes to match the largest group, ensuring uniform representation as shown in Eq. 6 [33, 41]. This ensures a balanced representation across all subgroups [42]. This uniformity aids in mitigating biases arising from unequal group representations.

$$\begin{aligned} Expected\_Size_{PP} &= Expected\_Size_{UP} = Expected\_Size_{PN} \\ &= Expected\_Size_{UN} = \max\{|PP|, |UP|, |PN|, |UN|\} \end{aligned} \tag{6}$$

3.2 Sampling Strategies

The Synthetic Minority Oversampling Technique (SMOTE) is a widely used data augmentation method designed to address class imbalance in datasets [41]. It generates synthetic examples for the minority class by interpolating between existing minority instances and their nearest neighbors in the feature space [43]. This approach helps in creating a more balanced dataset, which can lead to improved model performance on minority classes [43].

In this paper, we applied Fair-SMOTE, a fairness-aware oversampling method developed by Chakraborty et al. [33, 41, 44], to generate new synthetic data points. Unlike standard SMOTE variants, Fair-SMOTE preserves inter-feature associations by extrapolating all variables by the same factor between two nearest neighbors, thereby reducing

distortion in the feature space [41]. It also accounts for data types: boolean, categorical, and numeric features are mutated using dedicated logic [45]. In accordance with the original Fair-SMOTE implementation, we set the mutation amount (f) and crossover frequency (cr) hyperparameters to 0.8, reflecting a strong preference for interpolating new points that remain close to their parent instances [41]. A k-nearest neighbors' algorithm with $k = 3$ was used to identify neighbors for interpolation, ensuring consistent subgroup-level sampling across all datasets. In our work, we explicitly decompose the Fair-SMOTE mechanism into two components:

1. *Parent sample selection and*
2. *Sample generation via extrapolation*

We retain the original Fair-SMOTE generation logic unchanged for the second component. Importantly, throughout this paper, the term "SMOTE" always refers to the synthetic sample generation process implemented using Fair-SMOTE, rather than the standard SMOTE algorithm.

For the first stage, parent selection, we extend the original Fair-SMOTE implementation, which selected parent points randomly, by exploring three alternative strategies aimed at better aligning data augmentation with fairness objectives:

- **Uniform Sampling with SMOTE (US_SM):** Randomly selects samples from underrepresented groups to generate synthetic data until the expected size is reached [8]. All groups' sizes are adjusted based on demographic parity. This strategy mirrors the random parent selection used in the original Fair-SMOTE implementation.
- **Preferential Sampling with SMOTE (PS_SM):** prioritizes samples near the decision boundary for generating synthetic data [39].
- **Weighted Preferential Sampling with SMOTE (WPS_SM):** Assigns sampling weights based on proximity to decision boundaries, offering more nuanced augmentation by focusing on more informative samples [46, 47].

3.3 Preprocessing Techniques

By combining the aforementioned group size adjustment strategies with sampling techniques, we designed and evaluated nine preprocessing methods:

1. **US_SM:** Applies Uniform Sampling with SMOTE to adjust all groups according to demographic parity-based global adjustment.
2. **PS_SM:** Integrates Preferential Sampling with SMOTE, focusing on samples near the decision boundary across all groups to meet demographic parity-based global adjustment.
3. **WPS_SM:** Weighted Preferential Sampling with SMOTE enhances PS_SM by assigning weights to samples based on their distance from the decision boundary. More informative samples have a higher probability of being selected as parents for SMOTE across all groups, in line with demographic parity-based global adjustment.
4. **UP_US_SM:** Applies Uniform Sampling with SMOTE only to the UP group, adjusting its size based on UP-based adjustment approach.
5. **UP_PS_SM:** Applies Preferential Sampling with SMOTE solely to the UP group, targeting samples near the decision boundary according to the UP-based adjustment approach.

6. **UP_WPS_SM:** Applies Weighted Preferential Sampling with SMOTE exclusively to the UP group to meet UP-based adjustment approach.
7. **B_US_SM:** Utilizes Uniform Sampling with SMOTE to equalize all group sizes to the largest group. This technique is functionally equivalent to the Fair-SMOTE method introduced by Chakraborty et al. [41]. To maintain consistency in our naming convention, we refer to it as B_US_SM.
8. **B_PS_SM:** Combines Preferential Sampling with SMOTE, balancing all groups to the largest size as per the Balanced Size Techniques strategy.
9. **B_WPS_SM:** Integrates Weighted Preferential Sampling with SMOTE, ensuring all groups match the size of the largest group.

This structured approach allows for a systematic evaluation of different preprocessing configurations and their impact on fairness and performance metrics. To ensure robust and reliable results, we conducted experiments using 50 random seeds for each combination of dataset and protected attribute. First, a baseline experiment with a standard Logistic Regression, Decision Tree, Gradient Boosting, Random Forest, and Support Vector Classification on each of the conditions (dataset_protected_attribute) was performed to compare and benchmark the results of the debiasing experiments. These five algorithms, with the goal of maximizing accuracy, were employed to compare how pre-processing mitigation techniques impact different models in different datasets with different protected attributes.

All techniques were applied under two scenarios, mitigating bias toward a single protected attribute and mitigating bias toward multiple protected attributes simultaneously. All techniques were evaluated under identical conditions and compared against the current techniques to assess their performance and fairness.

4 Results and Discussion

All the introduced techniques were capable of mitigating bias towards one protected attribute at a time and multiple protected attributes simultaneously. Table 1 presents a demonstration of the results for one of the datasets, the Adult Income dataset, focusing on mitigating bias related to the race protected attribute. Logistic Regression is employed as both the classifier and ranker in this analysis. B_US_SM serves as an established preprocessing sampling technique, while the other methods generate synthetic data based on different methodologies for defining group sizes and sample selection. To evaluate the effectiveness of these techniques, we analyzed the median values of performance and fairness metrics. The results indicate that UP_WPS_SM and PS_SM are competitive in achieving the most favorable outcomes concerning the Demographic Parity fairness metric for this dataset. Regarding Equalized Odds fairness metrics, among the techniques, US_SM outperforms the others. Techniques focusing on the UP group size also perform well concerning these metrics. While B_US_SM and B_PS_SM exhibit strong performance among the balanced group techniques for Equalized Odds fairness metrics, B_WPS_SM adversely affects these metrics. The highest overall accuracy is achieved by techniques focusing on the demographic parity notion of fairness size approach, whereas the highest BAC is attained by techniques employing balanced group sizes. Figure 2 presents a Pareto Frontier plot [48], illustrating the trade-off between DI and

accuracy for different sampling techniques on the Adult Income dataset with the race protected attribute. Techniques like PS_SM and WPS_SM demonstrate significant gains in fairness (DI) but show varied effects on accuracy. In contrast, the UP-based methods maintain higher accuracy but yield more modest improvements in fairness.

Table 1. Results of applying pre-processing techniques to mitigate race-protected attribute in the Adult dataset

Technique	DI	SP	AOD	EO	ACC	BAC
baseline	0.41	-0.09	-0.09	-0.13	0.82	0.68
US_SM	0.73	-0.04	0.01	0.02	***0.82***	0.68
PS_SM	1.21	0.03	0.12	0.19	***0.82***	0.68
WPS_SM	2.46	0.08	0.19	0.33	0.80	0.61
UP_US_SM	0.80	-0.03	0.03	0.06	***0.82***	0.69
UP_PS_SM	0.83	-0.03	0.04	0.07	***0.82***	0.69
UP_WPS_SM	***0.88***	***-0.02***	0.04	0.08	***0.82***	0.69
B_US_SM	0.79	-0.08	***0.00***	0.02	0.77	***0.76***
B_PS_SM	0.76	-0.09	-0.02	***0.01***	0.77	***0.76***
B_WPS_SM	2.23	0.21	0.31	0.40	0.80	0.70

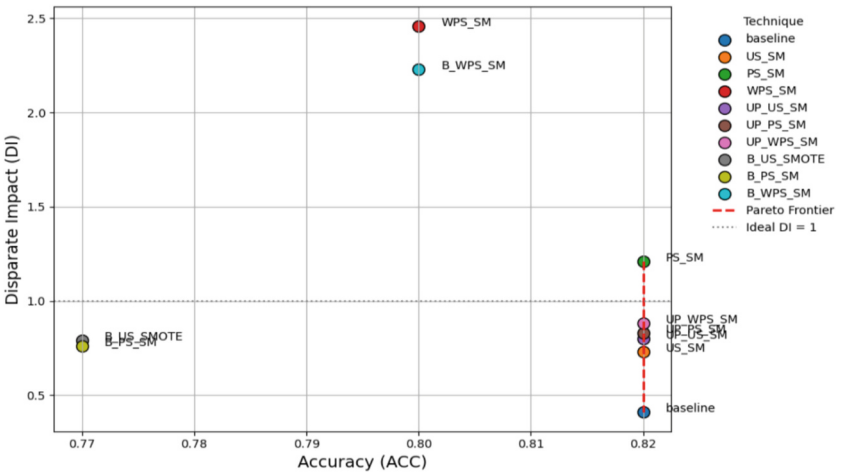

Fig. 2. Pareto Frontier: Tradeoff between Disparate Impact and Accuracy (Adult-race)

Given the extensive nature of the results across ten diverse datasets, presenting all findings in detail would be impractical and could obscure key insights. To succinctly summarize and compare the effectiveness of each technique, we employed the Scott-Knott clustering method, a statistical approach that partitions techniques into distinct

groups based on their performance distributions [41]. This method assigns ranks to each group, where a higher rank signifies superior performance, and techniques within the same group are considered statistically indistinguishable [41]. For a more granular comparison, we reported the number of Wins, Losses, and Ties for each technique. A "Win" indicates that a technique achieved a higher rank compared to another, a "Loss" denotes a lower rank, and a "Tie" reflects statistical parity between techniques. To further quantify these comparisons, we calculated the Win/Loss Ratio (WLR), representing the proportion of wins to losses, and the Tie/Total Ratio (TTR), indicating the ratio of tied outcomes to the total number of comparisons. This analytical framework allowed us to effectively distill the performance of various bias mitigation strategies across multiple datasets, facilitating a robust and comprehensive comparison.

Table 2 shows the Scott-Knott results comparing PS_SM and UP_PS_SM techniques across all datasets in the study. Both apply SMOTE to generate synthetic samples near the decision boundary. PS_SM oversamples UP and PN groups while undersampling UN and PP groups based on the Demographic Parity objective, whereas UP_PS_SM exclusively oversamples the UP group. The results reveal that PS_SM outperforms UP_PS_SM in terms of Demographic Parity fairness metrics and most performance metrics such as accuracy and balanced accuracy. However, UP_PS_SM achieves significantly better results for Equalized Odds fairness metrics.

Table 2. PS_SM vs UP_PS_SM Scott-Knott Result

	1 protected attribute at a time					2 protected attributes at a time				
Metric	Wins	Ties	Losses	WLR	TTR	Wins	Ties	Losses	WLR	TTR
DI	**34**	18	23	1.48	0.24	**24**	16	10	2.4	0.32
SP	**37**	14	24	1.54	0.19	**29**	10	11	2.64	0.2
AOD	10	24	**41**	0.24	0.32	10	18	**22**	0.45	0.36
EO	8	28	**39**	0.21	0.37	4	24	**22**	0.18	0.48
Acc	**25**	28	22	1.14	0.37	16	16	**18**	0.89	0.32
BAC	**33**	24	18	1.83	0.32	**24**	12	14	1.71	0.24

In general, techniques based on UP group size adjustment consistently perform better than those based on balanced group sizes. Among all, B_WPS_SM ranks lowest in fairness metrics, suggesting that balancing all groups equally while focusing only on samples near the decision boundary may not effectively reduce bias. Interestingly, while WPS-based methods underperform when applied within demographic parity-based global size adjustment or balanced group frameworks, the UP_WPS_SM technique shows substantial improvements in all fairness metrics and in accuracy. As illustrated in Table 3, UP_WPS_SM achieves a high number of wins in fairness comparisons against B_WPS_SM, confirming that selective oversampling in the UP group using weighted proximity can effectively mitigate bias without overly disturbing model accuracy.

Table 3. UP_WPS_SM VS B_WPS_SM Scott-Knott Result

	1 protected attribute at a time					2 protected attributes at a time				
Metric	Wins	Ties	Losses	WLR	TTR	Wins	Ties	Losses	WLR	TTR
DI	**47**	12	16	2.94	0.16	**22**	14	14	1.57	0.28
SP	**38**	14	23	1.65	0.19	15	10	**25**	0.6	0.2
AOD	**49**	18	8	6.12	0.24	**24**	16	10	2.4	0.32
EO	**54**	17	4	13.5	0.23	**31**	18	1	31	0.36
Acc	**50**	15	10	5	0.2	**38**	6	6	6.33	0.12
BAC	21	16	**38**	0.55	0.21	16	12	**22**	0.73	0.24

Finally, Table 4 presents a comprehensive ranking of all SMOTE-based bias mitigation techniques, evaluated across all datasets using the Scott-Knott methodology. This analysis considers both fairness metrics (DI and AOD) and performance metrics (accuracy and balanced accuracy) when applying techniques to mitigate bias toward a single protected attribute. The rankings highlight that PS_SM and US_SM excel in enhancing fairness.

Across the evaluated datasets, the performance of SMOTE-based bias mitigation techniques varies, particularly concerning the Demographic Parity fairness metric. The PS_SM technique frequently achieves top performance, ranking first, in 9 out of 15 datasets. The UP_PS_SM method also demonstrates strong performance, often securing the second position. The theoretical foundation supporting our empirical findings lies in the observation that samples located near the decision boundary are more prone to misclassification, particularly for underrepresented or unprivileged groups [39]. Prior research has demonstrated that concentrating synthetic sampling in these boundary regions allows the classifier to gain a more nuanced understanding of ambiguous instances, ultimately reducing classification errors for disadvantaged groups [49]. Consequently, techniques which generate synthetic data in the proximity of the decision boundary, are better equipped to improve Demographic Parity fairness metrics without significantly compromising predictive performance [8].

According to both Fig. 2 and Tabe 4, while the baseline technique performs the worst on both DI and AOD, it achieves the highest accuracy, demonstrating the inherent tradeoff between fairness and accuracy. Similarly, B_US_SM leads in balanced accuracy, another performance metric. However, while PS_SM and US_SM perform best on specific fairness metrics such as DI and AOD, they do not consistently lead in predictive performance metrics. This highlights a core insight of our study: no single technique dominates across all dimensions. In many cases, improvements in fairness come at the cost of reduced predictive performance, underscoring the need for context-dependent decision-making. Practitioners must therefore weigh the relative importance of fairness versus accuracy depending on the societal, ethical, and operational implications of misclassification. Techniques focusing on the UP group, such as UP_WPS_SM and UP_PS_SM, consistently achieve high rankings across both fairness and accuracy. These methods are particularly suitable for scenarios where fairness and accuracy must be

optimized simultaneously. In contrast, techniques such as B_WPS_SM, while enhancing certain fairness metrics, incur a noticeable drop in accuracy, reflecting a steeper trade-off. Interpreted through the lens of cost-sensitive fairness, this implies that the marginal fairness gain may not always justify the associated performance cost [50]. These insights support a multi-objective optimization perspective, where fairness and accuracy are treated not in isolation but as competing objectives to be jointly optimized depending on application-specific constraints [48].

Table 4. Comparative rankings of preprocessing sampling bias mitigation techniques

Rank	DI	AOD	ACC	Bac
1	PS_SM	US_SM	Baseline	B_US_SM
2	UP_WPS_SM	UP_PS_SM	US_SM	B_PS_SM
3	UP _PS_SM	UP_WPS_SM	UP_US_SM	Baseline
4	UP _US_SM	UP_US_SM	UP_WPS_SM	B_WPS_SM
5	US_SM	B_US_SM	PS_SM	US_SM
6	WPS_SM	B_PS_SM	UP_PS_SM	PS_SM
7	B_PS_SM	PS_SM	WPS_SM	UP_US_SM
8	B_US_SM	WPS_SM	B_US_SM	UP_WPS_SM
9	B_WPS_SM	B_WPS_SM	B_PS_SM	UP_PS_SM
10	Baseline	Baseline	B_WPS_SM	WPS_SM

The behavior of each method also varies by dataset. For example, B_WPS_SM excels, particularly in datasets that suffer from both group and positive ratio imbalance, such as the Heart dataset. This method is effective because it focuses on generating more samples relative to their positions. On the other hand, for the Open University dataset, US_SM (randomly sampling) is sufficient, as the imbalance is mild and there is no need to specifically focus on samples near the decision boundary. For the Compass dataset, generating samples only within the UP group is adequate, and this approach results in fair outcomes, making UP_SM_PS perform well for this dataset.

While our proposed techniques yield notable improvements in fairness metrics, several limitations should be acknowledged. First, SMOTE-based methods rely on generating synthetic samples, which may not always reflect the true data distribution, particularly in high-dimensional or non-linear feature spaces, leading to potential overfitting. Second, the effectiveness of these techniques depends on access to protected attribute labels during training, which may not always be available due to legal, ethical, or privacy restrictions. Lastly, computational overhead can be significant when boundary estimation is expensive, especially for large-scale or streaming data environments. These factors highlight the importance of cautious validation and adaptive deployment when applying fairness-aware preprocessing in practice.

5 Conclusion

This study conducted a comprehensive evaluation of SMOTE-based bias mitigation techniques, focusing on their effectiveness in enhancing Demographic Parity and Equalized Odds fairness metrics across ten diverse datasets. The techniques were categorized based on group size determination methods: (1) Demographic Parity-Based Global Adjustment, (2) Demographic Parity focusing solely on the unprivileged positive group, and (3) Balanced group sizes. Additionally, we explored three synthetic sample selection strategies: Uniform Sampling, Preferential Sampling, and Weighted Preferential Sampling. We established a structured framework to assess their impacts. Our analysis revealed that PS_SM and US_SM achieved the best performance among all techniques for Disparate Impact and Average Odds Difference fairness metrics, respectively. Techniques focusing on the UP group, especially those that generate synthetic samples near the decision boundary, offered strong fairness gains with minimal sacrifice in accuracy, validating the hypothesis that boundary-focused augmentation reduces misclassification of disadvantaged groups. These results highlight the trade-off space between fairness and performance, with methods like UP_WPS_SM offering effective middle ground solutions for real-world deployments. While no single method dominates across all settings, our findings emphasize the need for context-specific choices in real-world deployments. Future work should expand to more complex models, including deep learning models. Furthermore, practical deployment considerations, such as the limited availability of protected attribute labels, regulatory compliance, and the risks associated with synthetic data generation, must be central to developing fair, robust, and scalable machine learning systems.

References

1. Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., Galstyan, A.: A survey on bias and fairness in machine learning. ACM Comput. Surv. **54**, (2021). https://doi.org/10.1145/3457607
2. François-Blouin, J.: Responsible AI Symposium – Legal Implications of Bias Mitigation - Lieber Institute West Point. https://lieber.westpoint.edu/legal-implications-bias-mitigation/. Accessed 02 Dec 2023
3. Goethals, S., Calders, T., Martens, D.: Beyond Accuracy-Fairness: stop evaluating bias mitigation methods solely on between-group metrics (2024)
4. Tawakuli, A., Engel, T.: Make your data fair: a survey of data preprocessing techniques that address biases in data towards fair AI. J. Eng. Res. (2024). https://doi.org/10.1016/j.jer.2024.06.016
5. Hort, M., et al.: Bias mitigation for machine learning classifiers: a comprehensive survey. ACM J. Responsible Comput. **1**, 1–52 (2024). https://doi.org/10.1145/3631326
6. Shahrezaei, M.H., Loughran, R., Daid, K.M.: Pre-processing techniques to mitigate against algorithmic bias. In: 2023 31st Irish *Conference on Artificial Intelligence and Cognitive Science* AICS 2023 (2023). https://doi.org/10.1109/AICS60730.2023.10470759
7. Peng, K., Yang, Y., Zhuo, H., Menzies, T.: Whence is a model fair? Fixing fairness bugs via propensity score matching (2025)
8. Kamiran, F., Calders, T.: Data preprocessing techniques for classification without discrimination. Springer (2012). https://doi.org/10.1007/s10115-011-0463-8

9. Celis, L.E., Keswani, V., Vishnoi, N.K.: Data preprocessing to mitigate bias: a maximum entropy based approach. In: 37th *International Conference on Machine Learning*. ICML 2020. PartF16814, pp. 1326–1336 (2020)
10. Yu, Z.: FairBalance: mitigating machine learning bias against multiple protected attributes with data balancing (2021)
11. Xu, D., Wu, Y., Yuan, S., Zhang, L., Wu, X.: Achieving causal fairness through generative adversarial networks. In: Proceedings Twenty-Eighth International Joint Conference on 2019. par.nsf.gov. (2019)
12. Sharma, S., Zhang, Y., Aliaga, J.M.R.O., Bouneffouf, D., Muthusamy, V., Varshney, K.R.: Data augmentation for discrimination prevention and bias disambiguation, pp. 358–364 (2020). https://doi.org/10.1145/3375627.3375865
13. Salazar, T., Santos, M.S., Araujo, H., Abreu, P.H.: FAWOS: fairness-aware oversampling algorithm based on distributions of sensitive attributes. IEEE Access. **9**, 81370–81379 (2021)
14. Friedler, S.A., Choudhary, S., Scheidegger, C., Hamilton, E.P., Venkatasubramanian, S., Roth, D.: A comparative study of fairness-enhancing interventions in machine learning. In: FAT* 2019 - Proceeding of the 2019 Conference Fairness, Accountability, Transparency, pp. 329–338 (2019). https://doi.org/10.1145/3287560.3287589
15. EEOC: Title VII of the Civil Rights Act of 1964 (2013). https://doi.org/10.4135/9781452218533.n690
16. UK ICO: What's new? How do we ensure transparency in AI? Inf. Comm. Off. (2023)
17. Jin, D., et al.: A survey on fairness-aware recommender systems. Inf. Fusion. **100**, 101906 (2023). https://doi.org/10.1016/j.inffus.2023.101906
18. Saxena, N.A.: Perceptions of fairness. In: Proceeding of the 2019 AAAI/ACM Conference on AI, Ethics, and Society, pp. 537–538 (2019). https://doi.org/10.1145/3306618
19. Barocas, S., Hardt, M., Narayanan, A.: Fairness and machine learning (2019)
20. Narayanan, A.: Translation tutorial : 21 fairness definitions and their politics. 21 (2019)
21. Nielsen, A.: Practical fairness: achieving fair and secure data models, 330 (2020)
22. Wang, Y., Singh, L.: Analyzing the impact of missing values and selection bias on fairness. Int. J. Data Sci. Anal. **12**, 101–119 (2021). https://doi.org/10.1007/s41060-021-00259-z
23. Langenberg, A., Ma, S.C., Ermakova, T., Fabian, B.: Formal group fairness and accuracy in automated decision making. Mathematics **11**, 1–25 (2023). https://doi.org/10.3390/math11081771
24. Varshney, K.R.: Trustworthy. Mach. Learn. (2022). https://doi.org/10.1109/MIS.2022.3152946
25. Dwork, C., Hardt, M., Pitassi, T., Reingold, O., Zemel, R.: Fairness through awareness. ITCS **2012**, 214–226 (2012). https://doi.org/10.1145/2090236.2090255
26. Rao, D.: Fairness in AI — Ethical Implications of ML Models (2021)
27. Verma, S., Rubin, J.: Fairness definition explained. ACM **18**, 1–7 (2018). https://doi.org/10.1145/3194770.3194776
28. Hardt, M., Price, E., Srebro, N.N.: Equality of opportunity in supervised learning. Adv. Neural Inf. Process. Syst. **29**, 3323–3331 (2016). https://doi.org/10.48550/arxiv.1610.02413
29. Kleinberg, J., Ludwig, J., Mullainathan, S., Rambachan, A.: Algorithmic fairness. AEA Pap. Proc. **108**, 22–27 (2018). https://doi.org/10.1257/pandp.20181018
30. Grandini, M., Bagli, E., Visani, G.: Metrics for multi-class classification: an overview (2020)
31. Brodersen, K.H., Ong, C.S., Stephan, K.E., Buhmann, J.M.: The balanced accuracy and its posterior distribution. In: Proceedings of the International Conference Pattern Recognition, pp. 3125–3128 (2010). https://doi.org/10.1109/ICPR.2010.764
32. Žliobaite, I., Kamiran, F., Calders, T.: Handling conditional discrimination. In: Proceedings of the IEEE International Conference Data Mining, ICDM., pp. 992–1001 (2011). https://doi.org/10.1109/ICDM.2011.72

33. Yu, Z., Chakraborty, J., Menzies, T.: FairBalance: how to achieve equalized odds with data pre-processing. IEEE Trans. Softw. Eng., 1–15 (2024). https://doi.org/10.1109/TSE.2024.3431445
34. Rosado Gómez, A.A., Calderón Benavides, M.L., Espinosa, O.: Data preprocessing to improve fairness in machine learning models: an application to the reintegration process of demobilized members of armed groups in Colombia. Appl. Soft Comput. **152** (2024). https://doi.org/10.1016/j.asoc.2023.111193
35. Duong, M.K., Conrad, S.: Measuring and mitigating bias for tabular datasets with multiple protected attributes (2024)
36. Turner Lee, N., Resnick, P., Barton, G.: Algorithmic bias detection and mitigation: Best practices and policies to reduce consumer harms (2019)
37. Fabris, A., Messina, S., Silvello, G., Susto, G.A.: Algorithmic fairness datasets: the story so far. Data Min. Knowl. Discov. **36**, 2074–2152 (2022). https://doi.org/10.1007/s10618-022-00854-z
38. Le Quy, T., Roy, A., Iosifidis, V., Zhang, W., Ntoutsi, E.: A survey on datasets for fairness-aware machine learning. Wiley Interdiscip. Rev. Data Min. Knowl. Discov. **12**, 1–59 (2022). https://doi.org/10.1002/widm.1452
39. Kamiran, F., Calders, T.: Classification with no discrimination by preferential sampling. In: Annual Machine Learning Conference. Belgium, Netherlands, pp. 1–6 (2010)
40. Albalak, A., et al.: A survey on data selection for language models, 1–81 (2024)
41. Chakraborty, J., Majumder, S., Menzies, T.: Bias in machine learning software: why? how? what to do? Assoc. Comput. Mach. (2021). https://doi.org/10.1145/3468264.3468537
42. Chakraborty, J., Majumder, S., Yu, Z., Menzies, T.: Fairway: a way to build fair ML software. ESEC/FSE **2020**, 654–665 (2020). https://doi.org/10.1145/3368089.3409697
43. Chawla, N.V., Bowyer, K.W., Hall, L.O., Kegelmeyer, W.P.: SMOTE: synthetic minority over-sampling technique. J. Artif. Intell. Res. **16**, 321–357 (2002). https://doi.org/10.1613/jair.953
44. Peng, K., Chakraborty, J., Menzies, T.: FairMask: better fairness via model-based rebalancing of protected attributes. IEEE Trans. Softw. Eng. **49**, 2426–2439 (2023). https://doi.org/10.1109/TSE.2022.3220713
45. Chakraborty, J.: joymallyac/Fair-SMOTE: GitHub repo for FSE 2021 Paper - Bias in Machine Learning Software: Why? How? What to do?. https://github.com/joymallyac/Fair-SMOTE/tree/master. Accessed 04 Jun 2025
46. Hu, Z., Xu, Y., Gu, J.: Boosting fair classifier generalization through adaptive priority reweighing (2024). https://doi.org/10.1145/3665895
47. Krasanakis, E., Spyromitros-Xioufis, E., Papadopoulos, S., Kompatsiaris, Y.: Adaptive sensitive reweighting to mitigate bias in fairness-aware classification, 853–862 (2018). https://doi.org/10.1145/3178876.3186133
48. Nagpal, R., Shahsavarifar, R., Goyal, V., Gupta, A.: Optimizing fairness and accuracy: a Pareto optimal approach for decision-making. AI Ethics 2024 **52**(5), 1743–1756 (2024). https://doi.org/10.1007/S43681-024-00508-4
49. Gnip, P., Zoričak, M., Kanász, R., Drotár, P., Kanász, R., Drotár, P.: An experimental survey of imbalanced learning algorithms for bankruptcy prediction (2025). https://doi.org/10.1007/s10462-025-11107-y
50. Liu, S., Vicente, L.N., Nunes Vicente, L.: Accuracy and fairness trade-offs in machine learning: a stochastic multi-objective approach. Comput. Manag. Sci. **19**, 513–537 (2020). https://doi.org/10.1007/s10287-022-00425-z