Evaluating the Impact of Situation Testing on SMOTE-Based Sampling Techniques for Bias Mitigation

Maliheh Heidarpour Shahrezaei[1][0000-0002-3644-0468], Kevin Mc Daid[2][0000-0002-0695-9082] and Róisín Loughran[3][0000-0002-0974-7106]

[1] RSRC, Dundalk Institute of Technology, Ireland, Maliheh.heidarpour@dkit.ie,
[2] RSRC, Dundalk Institute of Technology, Ireland, Kevin.mcdaid@dkit.ie
[3] RSRC, Dundalk Institute of Technology, Ireland, Roisin.Loughran@dkit.ie

Abstract. Sampling bias can be mitigated through preprocessing techniques such as oversampling and undersampling in imbalanced datasets. However, they do not address labeling bias, which can reinforce unfair model behavior. To address this, we established a structured evaluation framework that applies Situation Testing as an additional step after six SMOTE-based sampling techniques. Across 10 datasets and multiple classifiers, our findings reveal that the number of biased samples removed by Situation Testing depends on the dataset, classifier, and preprocessing sampling technique applied beforehand. Applying Situation Testing directly to the baseline consistently improved fairness with respect to both Demographic Parity and Equalized Odds, albeit with reduced predictive performance. In contrast, the impact of Situation Testing after sampling varied across strategies, proving to be more effective in group size adjustments based on Equalized Representation than in balanced-size techniques. Overall, the results highlight the trade-off between fairness gains and predictive utility, underscoring the need to align mitigation strategies with dataset characteristics and fairness objectives.

Keywords: Bias, Fairness, Mitigation Techniques, Situation Testing

1 Introduction

Unwanted algorithmic bias refers to unfair and discriminatory outcomes in algorithms that result in disadvantages for certain groups of people [1]. This bias can emerge when algorithms unintentionally favor or disadvantage particular individuals or communities based on race, gender, ethnicity, socioeconomic status, or other characteristics [2]. These characteristics, known as protected attributes, are features of individuals who are legally and ethically safeguarded from discrimination because they represent groups particularly vulnerable to unfair treatment [3, 4]. Protected attributes typically divide a population into privileged and unprivileged groups [5]. The privileged group typically receives favorable treatment or holds a more advantageous position, while the unprivileged group is subjected to unfair treatment in discriminatory decision-making processes [3].

Bias can arise in Machine Learning (ML) models in different ways [6]. The most frequent sources of bias in ML can be grouped into data, algorithm, and user interaction

[1]. When biases are present in the underlying training data, the algorithms trained on them will inevitably incorporate these biases into their predictions. Data bias might be amplified and maintained by algorithms. In addition, even if the data is not biased, algorithms themselves may exhibit biased behavior as a result of specific design decisions [7]. The outputs of these biased algorithms are then fed into actual systems and can influence user decisions, leading to further bias [1, 8]. Therefore, addressing bias early in the model's lifecycle, before it becomes amplified, is critical. One effective approach is through preprocessing techniques, which adjust the training data to mitigate bias [9–11]. Other approaches, such as in-processing techniques, incorporate fairness considerations directly into the training process [12, 13], and post-processing techniques modify model predictions to reduce bias [14, 15].

This study focuses on preprocessing sampling techniques designed to balance training data. These techniques, including oversampling and a combination of oversampling and undersampling, employ methods such as generating synthetic data points to balance underrepresented classes. Once the dataset is balanced, we apply Situation Testing (ST), which involves flipping protected attributes in the data and checking whether predictions change [9]. If predictions differ, the data point is flagged as biased and removed. In this way, ST identifies and eliminates potential labeling bias in the training data. Accordingly, this study addresses the following research question: *Does the application of ST consistently enhance fairness across SMOTE-based preprocessing sampling techniques in ML, without adversely affecting model performance?*

The following section reviews related work on the application of ST. Section 3 outlines the methodology, including sampling and ST techniques. In section 4, we present the fairness and performance metrics used for evaluation, along with key details about the datasets employed in this study. Section 5 provides a detailed analysis of the results, and Section 6 offers the concluding remarks.

2 Related works

ST is a critical method for detecting discrimination, particularly in legal contexts [16]. It involves structured experiments where individuals with similar qualifications but differing in a protected characteristic (e.g., race, gender) are placed in identical situations to detect discriminatory practices. This method has been widely used in European jurisdictions to provide empirical evidence of discrimination. In legal proceedings, it plays a crucial role in shifting the burden of proof to defendants, requiring them to justify their actions with legitimate, non-discriminatory reasons [16]. Beyond its legal applications, ST is also an effective tool for raising public awareness and shaping anti-discrimination policies. Its impact extends to legal professionals, policymakers, and researchers working to combat discrimination and promote equality [16].

In employment discrimination research, ST has been instrumental in exposing biases in hiring decisions [17]. ST involves pairs of individuals, testers, who are matched in qualifications and experience but differ in a specific characteristic. These pairs apply for the same job openings to observe whether employers treat them differently based on the characteristic being tested [17]. Studies conducted in the U.S. have revealed that discriminatory behavior, whether conscious or unconscious, was exhibited by

approximately 20% to 40% of employers [17]. One well-known study demonstrated that résumés with white-sounding names received 50% more callbacks than those with black-sounding names, highlighting racial bias in hiring decisions [18]. These findings have influenced public policy and civil rights enforcement by providing empirical evidence of workplace discrimination [17].

Beyond traditional employment settings, ST has been applied in data-driven environments to detect bias within datasets. One study introduces a framework that integrates ST with causal inference to uncover individual discrimination by isolating disparities linked to protected attributes [19]. Another study applies ST to software systems, systematically testing inputs that differ only in protected attributes to evaluate whether these attributes unjustly influence decision-making processes [20]. By identifying causal relationships that contribute to biased outcomes, this method effectively assesses discrimination in automated decision-making systems [20].

In ML, the "Fairway" approach integrates ST into a two-step bias mitigation framework [21]. The first step applies ST to detect biased data points by altering protected attributes and observing changes in model predictions. Data points that cause prediction shifts are removed from the training dataset to reduce the influence of protected attributes [21]. The second step employs multi-objective optimization to balance fairness and predictive performance during model training. This ensures that bias is mitigated without compromising accuracy, making Fairway a comprehensive tool for fairness improvement [21]. Another ML-based approach, Fair_SMOTE, combines oversampling techniques with ST to enhance fairness [9]. Fair_SMOTE first rebalances data distributions by ensuring equal representation of protected attributes across both positive and negative classes. After this oversampling step, ST is applied to evaluate whether model predictions remain consistent when protected attributes are modified [9]. Building on this line of work, our study isolates the ST component from Fair_SMOTE and systematically applies it across a wider set of state-of-the-art sampling techniques. This allows us to disentangle ST's independent contribution to fairness and performance and evaluate its consistency across diverse preprocessing strategies.

3 Methodology

We applied ST after preprocessing sampling techniques with the expectation that models would be trained not only on balanced datasets but also on datasets with reduced bias related to protected attributes. This combined approach was hypothesized to improve fairness in predictions while limiting trade-offs in dataset size and predictive performance.

Each dataset was partitioned into four canonical subgroups defined by the cross-product of protected attribute and class label: Privileged Positive, Privileged Negative, Unprivileged Positive, and Unprivileged Negative. These subgroups served as the basis for fairness-aware resampling [22].

Preprocessing sampling techniques were organized into two categories according to their subgroup size adjustment strategy. Equalized Representation (ER) balances the ratio of positive to negative samples across both privileged and unprivileged groups, consistent with Demographic Parity [22]. This preserves the overall dataset size while

equalizing outcome distributions. In contrast, the Balanced-Size (B) approach, inspired by Fair_SMOTE [9], adjusts all subgroups to match the size of the largest subgroup, ensuring uniform representation [23].

Standard Fair_SMOTE [9] operates similarly to Uniform Sampling (US) [22], randomly selecting samples to generate synthetic data [23]. We extend this approach by incorporating alternative selection strategies: Preferential Sampling (PS), which prioritizes samples close to the decision boundary where fairness vulnerabilities are most acute [24], and Weighted Preferential Sampling (WPS), which assigns probabilistic weights based on distance to the boundary, giving higher likelihood to fairness-sensitive regions [25].

By combining the two group size adjustment strategies (ER vs. B) with the three selection mechanisms (US, PS, WPS), we obtained six distinct SMOTE-based techniques:

- **ER_US_SM:** Equalized Representation with Uniform Sampling + SMOTE
- **ER_PS_SM:** Equalized Representation with Preferential Sampling + SMOTE
- **ER_WPS_SM:** Equalized Representation with Weighted Preferential Sampling + SMOTE
- **B_US_SM:** Balanced-Size with Uniform Sampling + SMOTE, aligned with the first stage of the Fair_SMOTE methodology [9].
- **B_PS_SM:** Balanced Preferential Sampling + SMOTE
- **B_WPS_SM:** Balanced Weighted Preferential Sampling + SMOTE

The practical steps for integrating ST with preprocessing sampling techniques include:

1. Preprocessing Sampling Techniques:
   a. Apply preprocessing sampling techniques
   b. Evaluate the impact of these techniques using the selected fairness and performance metrics.
2. Apply ST:
   a. Conduct ST by modifying protected attributes and analyzing how these changes affect the model's predictions.
   b. Identify and remove data points whose predictions change solely due to the protected attribute.
   c. Assess the impact of the combined approach (preprocessing sampling + ST) using the same fairness and performance metrics.

A supervised ML model is trained using the preprocessed sampled dataset. The trained model is used to predict outcomes for all data points. For each data point, the value of the protected attribute is flipped to its opposite or an equivalent category. The modified dataset is passed through the same trained model, and new predictions are obtained. If a prediction changes after the flip, it indicates that the model is sensitive to the protected attribute, suggesting it has learned biased patterns from the training data. Such biased data points are removed, and the model is retrained on the reduced dataset.

Combining preprocessing sampling techniques with ST yields several new approaches, denoted by appending "_ST" to the original method names. In this paper, we refer to B_US_SM_ST as Fair_SMOTE, consistent with the terminology used in the original study. Moreover, we applied the ST to the Baseline model, meaning no preprocessing or sampling techniques were applied before the implementation of ST, referred to as Baseline_ST.

4 Evaluation

We evaluated the effectiveness of these methods by examining the bias toward a single protected attribute and multiple (two) protected attributes simultaneously. The first implementation focused on applying preprocessing techniques to mitigate bias toward a single protected attribute. In the second implementation, the dataset was partitioned by the combination of two protected attributes and the class label, resulting in eight subgroups rather than four. Addressing bias in this manner is crucial for ensuring fairness in ML models [26], as focusing on a single attribute may overlook complex interdependencies between different attributes, potentially leading to unintended discriminatory outcomes [27]. For example, mitigating bias solely based on non-white individuals, considering race as the first protected attribute, without addressing gender as a second protected attribute (e.g., female), may result in a model that even increases biases against women. Therefore, adopting comprehensive strategies that address all protected attributes simultaneously are essential to ensure equitable treatment across diverse groups [28]. By considering two protected attributes simultaneously, our techniques enable a more comprehensive bias mitigation strategy that accounts for intersectional biases.

We conducted experiments using five well-known ML algorithms, namely Logistic Regression (LR), Decision Tree (DT), Gradient Boosting (GB), Random Forest (RF), and Support Vector Classification (SVC). For each dataset and protected attribute combination, we repeated the experiments 50 times with randomized sampling. We then applied the Scott-Knott [29] statistical test to compare the techniques.

4.1 Metrics

Fairness in algorithmic decision-making can be assessed through various metrics; broadly, the two main groups are Demographic Parity (DP) and Equalized Odds (EQO) fairness metrics [30]. While both aim to promote fairness, they measure different criteria. DP focuses on balancing outcomes across groups, whereas EQO ensures fairness in predictive performance by considering error rates [31]. This distinction reflects trade-offs between equal treatment and equitable outcomes, making the choice of metric context dependent.

DP fairness metrics assess whether different demographic groups receive similar outcomes, ensuring that no group is disproportionately advantaged or disadvantaged [32]. Statistical Parity (SP), a key metric under Demographic Parity, ensures that the probability of a positive outcome is equal across groups, aiming to prevent allocation harms where resources or opportunities may be unevenly distributed [33]. Disparate Impact (DI), on the other hand, measures the ratio of favorable outcomes between a protected group and the overall population [31].

EQO fairness metrics focus on reducing disparities in error rates between groups by ensuring fair treatment in terms of both true positives and false negatives [31]. In this criteria, Average Odds Difference (AOD) quantifies the disparity in error rates by averaging differences in false positive rates and false negative rates between groups. Equal Opportunity (EO) ensures that true positive rates are equal across demographic groups, thereby preventing bias in granting favorable outcomes [33]. By utilizing both

DP and EQO 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 [34].

Accuracy (Acc) is the most common performance measure which calculates the number of correct predictions divided by the number of total predictions [35]. However, in dealing with an imbalanced dataset, using accuracy alone may not be sufficient. Instead, it is suggested to use Balanced Accuracy (BAC), which calculates the average of sensitivity and specificity [36]. Balanced accuracy provides a more reliable evaluation metric for imbalanced datasets by considering both the true positive rate and true negative rate [36]. Notably, accuracy and balanced accuracy will yield the same value if the dataset is balanced.

4.2 Datasets

There are numerous well-established datasets in the fairness literature that facilitate comparisons with prior work on bias mitigation methods [37]. In this study, we utilize ten widely recognized tabular datasets, summarized in Table 1. This selection enables external validation and supports a rigorous assessment of the proposed techniques. These datasets are characterized by substantial class imbalance, as well as disparities in the distribution of protected attributes, both of which significantly impact model performance and the trade-offs associated with bias mitigation. Notably, five of these datasets contain two protected attributes, providing a unique opportunity to analyze the effects of mitigating bias across multiple protected attributes simultaneously.

Table 1. List of datasets

Dataset	Size	Class Label	Positive Class%	Protected Attributes	Privileged %
Adult Income	(48842,14)	Income	24	sex	67
				race	85
German Credit	(1000,20)	Credit	70	sex	69
				age	85
Bank Marketing	(45211,16)	Term deposit	47	age	14
Default of Credit Card Clients	(30000,23)	Default payment	22	sex	40
Titanic	(891,12)	Survived	38	sex	35
Heart Disease	(303,13)	Heart disease	46	sex	68
				age	64
OULAD	(32593,12)	Final result	68	sex	46
Law School	(20798,12)	Pass exam	95	race	84
Student Performance	(649,33)	Score>=10	54	sex	59
				age	94
COMPAS	(7214,52)	Two-year-recid	55	sex	19
				race	34

5 Results and Analysis

We evaluated five classifiers, using the same model in each case for both performance evaluation and ST-based bias detection. We examined the median percentage of

samples removed from the training data after applying ST. For single protected attributes, techniques that focused on decision-boundary instances, such as ER_PS_SM, ER_WPS_SM, and B_WPS_SM, identified a larger proportion of samples as biased. Consequently, applying ST to these techniques resulted in substantial data reduction. Notably, ER_PS_SM_ST and ER_WPS_SM_ST led to severe data loss, with up to 50% of the training samples removed in the highly imbalanced COMPAS dataset.

It is important to emphasize that the reported percentage reductions for ER-based Group size techniques are calculated relative to the original dataset size, as these preprocessing sampling techniques do not alter the size of the entire training set. In contrast, for balanced techniques, the percentage reductions were computed based on the dataset size after oversampling. Consistent with previous research [7], Fair_SMOTE showed a maximum reduction of 13%, across different classifiers. Table 2 shows the similar pattern when ST was applied after preprocessing bias mitigation techniques aimed at reducing bias toward two protected attributes simultaneously. ER_PS_SM_ST, ER_WPS_SM_ST, and B_WPS_SM_ST techniques, resulted in a significant reduction of training samples, particularly when applied to the COMPAS dataset. In contrast, techniques like Fair_SMOTE, and B_PS_SM_ST led to a more moderate reduction in the dataset, with a maximum of 12% of the training data removed across different datasets.

Table 2. Percentage of removed samples after applying ST following various sampling techniques to mitigate bias toward two protected attributes simultaneously with LR

Dataset		Adult		COMPAS		German		Heart		Student	
Technique		race	sex	race	sex	age	sex	age	sex	age	sex
WRT Original Size	Baseline_ST	12	12	22	22	24	24	12	12	10	10
	ER_US_SM_ST	2	2	20	20	0	0	6	6	9	9
	ER_PS_SM_ST	19	19	32	32	1	1	15	15	23	23
	ER_WPS_SM_ST	27	27	47	47	1	1	20	20	27	27
WRT Over-sampled Size	Fair_SMOTE	2	2	11	11	3	3	5	5	7	7
	B_PS_SM_ST	2	2	12	12	2	2	5	5	9	9
	B_WPS_SM_ST	20	20	19	19	4	4	10	10	7	7

To comprehensively assess the performance of the techniques, we evaluated them on 10 datasets with varying characteristics. Presenting results for all datasets in a single table would require extensive space and may complicate the visual interpretation of the findings. To mitigate this issue, we adopted a methodology inspired by Chakraborty et al. [7, 37] using the Scott-Knott test to compare result distributions across all datasets. This statistical test was applied to all methods to identify performance groupings based on statistically significant differences. The Scott-Knott procedure ranks the methods, with higher ranks indicating superior performance. If two distributions are statistically indistinguishable, they are assigned to the same rank. The terms Wins, Losses, and Ties are defined as follows: a Win occurs when one technique achieves a higher rank for a particular metric compared to another; a Loss denotes a lower rank; and a Tie indicates no significant difference in performance between the two techniques. By employing this approach, we were able to succinctly summarize performance across multiple

datasets and conduct a robust comparison between the new and established methods, as suggested in previous studies [7, 37].

Table 3 presents the Scott-Knott statistical test results comparing the Baseline model with its counterpart after applying ST, referred to as Baseline_ST. WLR (Win/Loss Ratio) represents the proportion of wins to losses for each technique across all evaluation metrics. TTR (Tie/Total Ratio) indicates the ratio of tied outcomes to the total number of comparisons made. The results demonstrate that incorporating ST alone significantly improves fairness in terms of both DP and EQO fairness metrics. However, this improvement comes with a trade-off, as applying ST leads to a reduction in overall model performance. This suggests that while ST effectively enhances fairness, it may also impact the predictive capability of the model, highlighting the balance between fairness and accuracy in bias mitigation strategies.

Table 3. Baseline_ST VS Baseline 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	**36**	36	3	12	0.48	**29**	20	1	29	0.4
SP	**41**	30	4	10.25	0.4	**31**	14	5	6.2	0.28
AOD	**40**	31	4	10	0.41	**31**	14	5	6.2	0.28
EO	**37**	34	4	9.25	0.45	**30**	16	4	7.5	0.32
Acc	4	**51**	20	0.2	0.68	0	26	**24**	0	0.52
BAC	2	**38**	35	0.06	0.51	0	20	**30**	0	0.4

Finally, according to the Scott Knott statistical test, we ranked all techniques, determining which techniques performed best for each evaluation metric. Tables 4 and 5 summarize the performance of various bias mitigation techniques concerning fairness and performance metrics, ranking each method based on its effectiveness across all evaluated criteria. According to Table 4, all techniques improved fairness in terms of DP fairness metrics when compared to the Baseline. Surprisingly, applying ST to the original training set (Baseline_ST) achieved better performance than B_WPS_SM. Among sampling approaches, techniques that incorporated ER-based group size adjustments consistently outperformed both Fair_SMOTE and balanced-size methods with respect to the DP fairness metric. Moreover, sampling methods targeting decision-boundary instances (ER_PS_SM and ER_WPS_SM) are more effective in improving DI than random oversampling approaches such as ER_US_SM. In contrast, for the EO fairness metric, this trend is reversed, with random oversampling methods outperforming decision-boundary-based techniques. The effectiveness of ST, however, varied depending on the bias-mitigation setting:

- Single Protected Attribute: For ER-based group size techniques, applying ST improved DI and EO when mitigating bias for a single protected attribute.
- Two Protected Attributes: When addressing bias across two protected attributes simultaneously, ST did not provide further DI improvements for ER-based group size techniques, though it did enhance EO.

For balanced-size techniques, ST generally did not lead to additional fairness gains in either the single- or multi-attribute scenarios. The exception was B_WPS_SM_ST, which outperformed its preprocessing counterparts for both DI and EO. Notably, this

method was more effective when mitigating bias for two protected attributes than for a single attribute.

Table 4. Ranking of each technique based on fairness metrics

	Single protected attribute		Two protected attributes simultaneously	
Rank	DI	EO	DI	EO
1	ER_PS_SM_ST	ER_US_SM_ST	ER_PS_SM	ER_US_SM_ST
2	ER_PS_SM	ER_US_SM	ER_WPS_SM	ER_US_SM
3	ER_US_SM_ST	ER_PS_SM_ST	ER_WPS_SM_ST	ER_PS_SM_ST
4	ER_WPS_SM_ST	ER_WPS_SM_ST	ER_PS_SM_ST	ER_WPS_SM_ST
5	ER_US_SM	B_US_SM	ER_US_SM	Baseline_ST
6	ER_WPS_SM	Baseline_ST	B_WPS_SM	ER_PS_SM
7	B_US_SM	B_PS_SM	B_WPS_SM_ST	B_US_SM
8	B_PS_SM	Fair_SMOTE	ER_US_SM_ST	Fair_SMOTE
9	B_WPS_SM_ST	B_PS_SM_ST	B_US_SM	B_PS_SM
10	Fair_SMOTE	ER_PS_SM	B_PS_SM	B_PS_SM_ST
11	B_PS_SM_ST	B_WPS_SM_ST	Fair_SMOTE	ER_WPS_SM
12	Baseline_ST	ER_WPS_SM	B_PS_SM_ST	B_WPS_SM_ST
13	B_WPS_SM	Baseline	Baseline_ST	B_WPS_SM
14	Baseline	B_WPS_SM	Baseline	Baseline

Table 5 summarizes the ranking of bias-mitigation techniques in terms of predictive performance metrics. Overall, ER-group-size-based techniques achieved higher accuracy than methods that adjusted group sizes to match the largest group, across both single and multi-attribute bias simulations. Among these, ER_US_SM demonstrated

Table 5. Ranking of each technique based on performance metrics

	Single protected attribute		Two protected attributes simultaneously	
Rank	ACC	BAC	ACC	BAC
1	Baseline	B_US_SM	Baseline	B_PS_SM_ST
2	ER_US_SM	Fair_SMOTE	ER_US_SM	Fair_SMOTE
3	ER_US_SM_ST	B_PS_SM	ER_US_SM_ST	B_PS_SM
4	Baseline_ST	B_PS_SM_ST	Baseline_ST	B_US_SM
5	ER_PS_SM_ST	B_WPS_SM_ST	ER_PS_ SM_ST	Baseline
6	ER_PS_SM	Baseline	ER_PS_SM	ER_US_SM
7	ER_WPS_SM_ST	B_WPS_SM	ER_WPS_ SM_ST	ER_US_SM_ST
8	ER_WPS_SM	ER_US_SM	B_PS_SM_ST	B_WPS_SM_ST
9	B_US_SM	ER_PS_SM_ST	Fair_SMOTE	ER_PS_SM_ST
10	B_PS_SM_ST	ER_PS_SM	ER_WPS_SM	B_WPS_SM
11	B_WPS_SM_ST	ER_US_SM_ST	B_PS_SM	ER_PS_SM
12	Fair_SMOTE	Baseline_ST	B_US_SM	Baseline_ST
13	B_PS_SM	ER_WPS_SM	B_WPS_SM_ST	ER_WPS_SM_ST
14	B_WPS_SM	ER_WPS_SM_ST	B_WPS_SM	ER_WPS_SM

the best overall accuracy. Incorporating ST into sampling techniques that used directed-selection further improved accuracy. In contrast, results for balanced accuracy followed a different pattern. ER-based adjustment methods generally performed worse than Baseline, whereas techniques that matched group sizes to the largest group typically showed improvements, except for B_WPS_SM, which did not yield substantial gains. Notably, applying ST to directed sampling techniques within the ER group-size category enhanced balanced accuracy.

In response to our research question, we find that ST does not consistently enhance fairness across all sampling techniques without affecting performance. While it reliably improves fairness on the baseline, its effectiveness after sampling varies, and fairness gains are often offset by reduced accuracy and balanced accuracy.

6 Conclusion

This study systematically evaluated the impact of ST on preprocessing sampling-based bias mitigation techniques, examining both fairness and performance. Several important patterns emerged. First, the number of biased samples removed by ST varies depending on the dataset, classifier, and preprocessing sampling technique applied beforehand. Decision-boundary approaches such as ER_PS_SM_ST and ER_WPS_SM_ST caused substantial data loss, particularly in highly imbalanced datasets like COMPAS. Second, the effectiveness of ST depended on the underlying sampling design, especially how group sizes were defined, and which samples were selected for synthetic data generation. Third, applying ST directly to the baseline consistently improved fairness with respect to both Demographic Parity and Equalized Odds, though these gains came at the cost of predictive utility, as evidenced by reductions in both accuracy and balanced accuracy. This underscores a central trade-off between fairness enhancement and predictive utility.

Overall, the results emphasize that ST is not a universal solution but can provide substantial fairness gains when applied directly to the baseline. Future work should investigate the drivers of data removal across technique–dataset–classifier settings and investigate fairness-aware oversampling strategies designed to counteract label bias.

Acknowledgments. This research is funded through the HEA's Technological University Transfor-mation Fund, co-funded by the Dundalk Institute of Technology

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/, last accessed 2023/12/02.
3. Goethals, S., Calders, T., Martens, D.: Beyond Accuracy-Fairness: Stop evaluating bias

mitigation methods solely on between-group metrics. CoRR. (2024). https://doi.org/10.48550/ARXIV.2401.13391.
4. Fang, B., Jiang, M., Cheng, P.-Y., Shen, J., Fang, Y.: Achieving Outcome Fairness in Machine Learning Models for Social Decision Problems. (2020).
5. Das, A.: Detection and Mitigation of Bias in Machine Learning Software and Datasets, (2023).
6. Suresh, H., Guttag, J. V.: A Framework for Understanding Sources of Harm throughout the Machine Learning Life Cycle. ACM Int. Conf. Proceeding Ser. (2019). https://doi.org/10.1145/3465416.3483305.
7. Pagano, T.P., Loureiro, R.B., Araujo, M.M., Lisboa, F.V.N., Peixoto, R.M., Guimaraes, G.A. de S., Santos, L.L. dos, Cruz, G.O.R., de Oliveira, E.L.S., Cruz, M., Winkler, I., Nascimento, E.G.S.: Bias and unfairness in machine learning models: a systematic literature review. 1–19 (2022).
8. Ferrara, E.: The Butterfly Effect in artificial intelligence systems : Implications for AI bias and fairness. Mach. Learn. with Appl. 15, 100525 (2024). https://doi.org/10.1016/j.mlwa.2024.100525.
9. Chakraborty, J., Majumder, S., Menzies, T.: Bias in machine learning software: Why? how? what to do? Association for Computing Machinery (2021). https://doi.org/10.1145/3468264.3468537.
10. Calmon, F.P., Wei, D., Vinzamuri, B., Ramamurthy, K.N., Varshney, K.R.: Optimized pre-processing for discrimination prevention. Adv. Neural Inf. Process. Syst. 2017-Decem, 3993–4002 (2017).
11. Heidarpour Shahrezaei, M., Loughran, R., Mc Daid, K.: Pre-processing Techniques to Mitigate Against Algorithmic Bias. 2023 31st Irish Conf. Artif. Intell. Cogn. Sci. AICS 2023. (2023). https://doi.org/10.1109/AICS60730.2023.10470759.
12. Zhang, L., Wu, Y., Wu, X.: Achieving non-discrimination in prediction. IJCAI Int. Jt. Conf. Artif. Intell. 2018-July, 3097–3103 (2018). https://doi.org/10.24963/ijcai.2018/430.
13. Grgi´cgrgi´c-Hlača, N., Zafar, M.B., Gummadi, K.P., Weller, A.: The Case for Process Fairness in Learning: Feature Selection for Fair Decision Making. (2016).
14. Kamiran, F., Mansha, S., Karim, A., Zhang, X.: Exploiting reject option in classification for social discrimination control. Inf. Sci. (Ny). 425, 18–33 (2018). https://doi.org/10.1016/J.INS.2017.09.064.
15. Calders, T., Verwer, S.: Three naive Bayes approaches for discrimination-free classification. Data Min. Knowl. Discov. 21, 277–292 (2010). https://doi.org/10.1007/s10618-010-0190-x.
16. Rorive, I., Lappalainen, P.: Proving Discrimination Cases - the Role of Situation Testing. (2009).
17. Bendick, M., Consultant, J.R.: Situation Testing for Employment Discrimination in the United States of America Perspective internationale. (2007).
18. Bertrand, M., Mullainathan, S.: Discrimination in the Job Market in the United States | The Abdul Latif Jameel Poverty Action Lab, https://www.povertyactionlab.org/evaluation/discrimination-job-market-united-states?utm_source=chatgpt.com, last accessed 2025/02/27.
19. Zhang, L., Wu, Y., Wu, X.: Situation testing-based discrimination discovery: A causal inference approach. IJCAI Int. Jt. Conf. Artif. Intell. 2016-Janua, 2718–2724 (2016).
20. Galhotra, S., Brun, Y., Meliou, A.: Fairness testing: Testing software for discrimination.

Proc. ACM SIGSOFT Symp. Found. Softw. Eng. Part F130154, 498–510 (2017). https://doi.org/10.1145/3106237.3106277.
21. 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.
22. Kamiran, F., Calders, T.: Data preprocessing techniques for classification without discrimination. Springer (2012). https://doi.org/10.1007/s10115-011-0463-8.
23. Heidarpour Shahrezaei, M., Mc Daid, K., Loughran, R.: Mitigating Algorithmic Bias through Sampling : The Role of Group Size and Sample Selection. 1–15.
24. Kamiran, F., Calders, T.: Classification with No Discrimination by Preferential Sampling. Informal Proc. 19th Annu. Mach. Learn. Conf. Belgium Netherlands. 1–6 (2010).
25. Hu, Z., Xu, Y., Gu, J.: Boosting Fair Classifier Generalization through Adaptive Priority Reweighing. (2024). https://doi.org/10.1145/3665895.
26. Duong, M.K., Conrad, S.: Measuring and Mitigating Bias for Tabular Datasets with Multiple Protected Attributes. (2024).
27. Turner Lee, N., Resnick, P., Barton, G.: Algorithmic bias detection and mitigation: Best practices and policies to reduce consumer harms (2019).
28. Rabonato, R.T., Berton, · Lilian: A systematic review of fairness in machine learning. AI Ethics 2024. 1–12 (2024). https://doi.org/10.1007/S43681-024-00577-5.
29. Faria, osé C., Jelihovschi, Ê.G., Allaman, I.B.: The ScottKnott Clustering Algorithm. (2025). https://doi.org/10.1590/1984.
30. 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.
31. Sweeney, L.: Discrimination in online ad delivery. Commun. ACM. 56, 44–54 (2013). https://doi.org/10.1145/2447976.2447990.
32. Chouldechova, A.: Fair Prediction with Disparate Impact: A Study of Bias in Recidivism Prediction Instruments. Big data. 5, 153–163 (2017). https://doi.org/10.1089/BIG.2016.0047.
33. Chouldechova, A., John Heinz, H., G'sell, M.: Fairer and more accurate, but for whom?
34. Fairlearn contributors: Common fairness metrics — Fairlearn 0.13.0.dev0 documentation, https://fairlearn.org/main/user_guide/assessment/common_fairness_metrics.html?utm_source=chatgpt.com, last accessed 2025/02/25.
35. Grandini, M., Bagli, E., Visani, G.: Metrics for Multi-Class Classification: an Overview. (2020).
36. Brodersen, K.H., Ong, C.S., Stephan, K.E., Buhmann, J.M.: The balanced accuracy and its posterior distribution. Proc. - Int. Conf. Pattern Recognit. 3125–3128 (2010). https://doi.org/10.1109/ICPR.2010.764.
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.