The Unseen Hand: Manipulating Model Fairness and SHAP with Targeted Identity Re-Association Attacks

Sannaan Khan*
mkhan.msse24sines@student.nust.edu.pk

Muhammad U. S. Khan
usman.shahid@sines.nust.edu.pk

National University of Sciences and Technology (NUST)

Abstract

As machine learning models grow more influential and opaque, algorithmic fairness and explainability are critical for ensuring accountability. However, we demonstrate that these auditing mechanisms are themselves vulnerable to subtle manipulation, camouflaging the influence of protected features. While prior work on data-agnostic attacks has exposed this vulnerability, they leave behind detectable artifacts that compromise their stealth. We introduce Targeted Identity Re-Association (TIRA) attacks, a novel family of attacks that iteratively and probabilistically manipulate a model's outputs without requiring access to the model's internals or feature representations. We formalize two algorithms: Probabilistic Micro-Shuffling (PMiS), which applies localized adjacent swaps, and Probabilistic Rank-Shift Micro-Perturbation (PRSMP), which introduces small, randomized rank shifts. We empirically demonstrate that TIRA attacks are highly effective at pushing fairness metrics towards ideal values. Crucially, TIRA attacks successfully confound SHAP-based explanations, leaving effectively zero residual attribution for protected features, a major improvement over prior work.

1 Introduction

The increasing integration of black box machine learning models into critical societal domains underscores the pressing need for algorithmic fairness and trustworthy explainability [5, 1, 3]. Given the stakes involved, the recent quest for transparent and explainable artificial intelligence (XAI) has led to the development of quantitative fairness [2] and explainability techniques like SHAP [9]. However, the wide adoption of such fairness assessments and post-hoc explanations gives birth to a new insidious vulnerability: are these explanations themselves robust to adversarial forces?

Earlier research has explored adversarial attacks against XAI methods. These require access to the underlying data distribution, which limits their applicability in real-world scenarios [13]. To overcome this barrier, recent work proposed 'fooling' strategies which often involve overt perturbations that, while effective and theoretically grounded, induce 'fairness score drops' and leave detectable SHAP attribution values that would trigger suspicion [16]. This leaves a critical need for subtle, data-agnostic, post-hoc manipulation techniques that effectively alter the model's perceived fairness and challenge the robustness of XAI explanations without leaving footprints of manipulations.

This work addresses this gap by introducing a novel family of Targeted Identity Re-Association (TIRA) attacks. Unlike previous work that employs more overt or deterministic shuffling, TIRA attacks are meticulously crafted to induce granular, iterative, and localized changes. We concretize this family through two distinct algorithms, Probabilistic Micro-Shuffling (PMiS) attack and Probabilistic Rank-Shift Micro-Perturbation (PRSMP) attack, to mimic natural score variation. The algorithms

*Corresponding Author.

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: MusIML.

expand the design space of shuffling attacks and show that subtle variation can further erode the reliability of fairness audits and explanation tools like SHAP.

We posit that the fairness assessment itself is brittle if a model's fairness perception can be altered by subtle, non-intrusive changes to its output rankings. Our comprehensive empirical investigations across diverse machine learning models and real-world datasets evaluate the efficacy of TIRA attacks in influencing AIF360 metrics and their impact on SHAP's attribution capabilities. Importantly, we highlight the fine-grained control offered by the intricate interplay of the tunable parameters, which showcase a scalable influence from subtle change to more pronounced alterations in fairness. Furthermore, our comparative analysis against existing work reveals that our probabilistic micro-shuffling strategies can achieve comparable or even better shifts in fairness, with greater stealth.

This work makes the following contributions: 1) We formalize and introduce Targeted Identity Re-Association (TIRA) attacks as novel, post-hoc, data-agnostic, and probabilistic methods for subtly manipulating model outputs. 2) We provide empirical validation of the TIRA attacks' ability to significantly alter AIF360 fairness metrics across various model-dataset configurations, guiding them towards desired thresholds. 3) We illustrate the intricate and insidious impact of TIRA attacks on the SHAP's ability to accurately attribute feature importance, particularly protected features. 4) We benchmark the performance of TIRA attacks against existing work and demonstrate that the TIRA attacks offer finer granularity and more control compared to existing, more overt attacks.

2 Methodology

This section delineates the details of our proposed Targeted Identity Re-Association (TIRA) attacks and the adversarial framework deployed.

2.1 Threat Model and Framework

This investigation operates under the assumption that the adversary has black-box access to the model's outputs. This is a realistic scenario as the adversary can be a model distributor or a model broker, or an actor operating on a logging or reporting layer. An adversary with the ability to perform proxy attribute inference is an advanced scenario. The adversary's capabilities are limited to: 1) The adversary can query the black-box model and obtain raw prediction scores for a given input instance. 2) The adversary is aware of the protected feature for each input instance, as this information is vital for targeting and shuffling the scores. 3) Importantly, the adversary has no access to the model's architecture, training data, or decision logic. 4) The manipulation is performed only on the ordered list of scores and the corresponding identities.

The adversarial objective is two-fold: 1) To shift and obscure the values of fairness metrics. 2) To fool SHAP so that the protected feature appears less influenced than it is.

The core framework for TIRA attacks proceeds as follows: **Score Sorting:** For given input data points ($\mathbf{x}$) and their protected features, the target model gives prediction scores ($\mathbf{s}$). The scores are then sorted in descending order, forming a ranked list. **Identity Re-Association:** Our attacks probabilistically and iteratively modify the ranked list, without changing the scores themselves. In each iteration, they apply a localized perturbation to identity-score associations. This means that a score remains in its position, but the identity associated with it is altered. **Output Re-Assembly:** After a predefined number of iterations, the reconstructed set of shuffled scores is used to calculate the fairness metrics and Shapley values.

2.2 Targeted Identity Re-Association Attacks

The proposed algorithms, Probabilistic Micro-Shuffling (PMiS) and Probabilistic Rank-Shift Micro-Perturbation (PRSMP), introduce controlled and probabilistic swaps. The attack algorithms reduce the apparent disparities while keeping the score distribution unchanged.

3 Results and Analysis

The shift in fairness metrics demonstrates the efficacy of the TIRA attacks, and the SHAP value attribution evaluates their subtlety.

Algorithm 1 Probabilistic Micro-Shuffling (PMiS) Attack

Require: Model f, dataset X_{test}, protected attribute A, swap probability p_{swap}, iterations T
Ensure: Shuffled scores S' re-associated with their original indices

$S \leftarrow f(X_{test})$
$L \leftarrow$ list of tuples (s_i, a_i, idx_i) for $i = 1 \dots N$
Sort L by score s_i in descending order
for $t = 1$ to T **do**
    for $i = 0$ to $N - 2$ **do**
        if $L[i].a = 0$ **and** $L[i+1].a = 1$ **then**
            if Rand(0,1) $\leq p_{swap}$ **then**
                Swap $(L[i].a, L[i].idx)$ with $(L[i+1].a, L[i+1].idx)$ ▷ Scores remain fixed, only identities are re-associated.
            end if
        end if
    end for
end for
return S' ▷ By re-ordering original scores based on the indices in L.

Algorithm 2 Probabilistic Rank-Shift Micro-Perturbation (PRSMP) Attack

Require: Model f, dataset X_{test}, protected attribute A, shift probability p_{shift}, max shift rank k_{max}, iterations T
Ensure: Shuffled scores S' re-associated with their original indices

$S \leftarrow f(X_{test})$
$L \leftarrow$ list of tuples (s_i, a_i, idx_i) for $i = 1 \dots N$
Sort L by score s_i in descending order
for $t = 1$ to T **do**
    for $i = 0$ to $N - 1$ **do**
        if $L[i].a = 0$ **and** Rand(0,1) $\leq p_{shift}$ **then**
            $k \leftarrow$ Rand$(1, k_{max})$
            $J \leftarrow \{j \in [i+1, \min(N-1, i+k)] \mid L[j].a = 1\}$
            if J is not empty **then**
                $j_{target} \leftarrow$ Rand-Choice(J)
                Swap $(L[i].a, L[i].idx)$ with $(L[j_{target}].a, L[j_{target}].idx)$
            end if
        end if
    end for
end for
return S' ▷ By re-ordering original scores based on the indices in L.

3.1 Fairness Metrics Manipulation

We quantitatively compare the results of the original, unattacked models against the outputs subjected to our TIRA attacks, as well as the benchmark attack. We consistently observed a significant shift in the perceived fairness, pushing the values towards ideal and fair thresholds. The key finding is that these probabilistic iterative attacks constantly achieved these results with notable precision.

Table 1 provides a summary of the results for the diabetes dataset using a logistic regression model. Both PMiS and PRSMP have a more pronounced impact on the metrics, compared to DomSwap and MixSwap attacks. This shows that iterative probabilistic, micro-level strategies provide fine-grained control and usually can more precisely tune the perceived fairness of a model's outputs to meet a desired threshold.

Table 2 summarizes the results for the credit dataset using neural networks, which showcase their generalizability across the datasets and models.

Note: **p** refers to the swap/shift probability (p_{swap} or p_{shift}), **I** refers to the number of iterations (T), and **r** refers to the maximum shift rank (k_{max}).

Table 1: Fairness Metrics Values on Bangladeshi Diabetes Dataset (LR Model)

Metric	Baseline	DomSwap	MixSwap	PMiS			PRSMP	
				(p=0.50, I=5)	(p=0.25, I=5)	(p=0.25, I=10)	(p=0.25, r=5, I=10)	(p=0.10, r=5, I=10)
Equal Opportunity	0.09	0.07	-0.06	**0.01**	**-0.07**	-0.13	-0.10	**0.03**
Demographic Parity	0.47	0.29	0.15	**0.25**	**0.04**	**-0.01**	**0.34**	**0.42**
Equal Odds	0.03	-0.07	0.26	-0.10	-0.23	-0.26	-0.06	**0.00**
Disparate Impact	2.40	1.75	1.26	**1.62**	**1.08**	**0.99**	**1.90**	**2.22**
Theil Index	0.00	0.01	0.02	0.01	0.04	0.05	0.01	**0.00**

Table 2: Fairness Metrics Values on German Credit Dataset (NN Model)

Metric	Baseline	DomSwap	MixSwap	PMiS			PRSMP		
				(p=0.25, I=5)	(p=0.25, I=10)	(p=0.50, I=5)	(p=0.25, r=5, I=10)	(p=0.33, r=5, I=10)	(p=0.25, r=5, I=5)
Equal Opportunity	-0.08	0.07	0.12	**-0.05**	**-0.07**	**-0.08**	-0.12	-0.13	**-0.07**
Demographic Parity	-0.06	0.10	0.07	**-0.05**	-0.12	-0.07	-0.10	-0.14	-0.08
Equal Odds	-0.02	0.15	0.07	**-0.01**	-0.11	**-0.02**	-0.05	-0.10	-0.05
Disparate Impact	0.90	1.14	1.09	**0.94**	0.84	**0.92**	0.87	0.82	**0.91**
Theil Index	0.00	2.21	0.00	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**

3.2 SHAP Attribution Analysis

Another core objective of the TIRA attacks is to fool SHAP by obfuscating the model's reliance on the protected features, without leaving a detectable footprint. Figure 1 show that after applying TIRA attacks, the SHAP value for the protected feature is effectively zero.

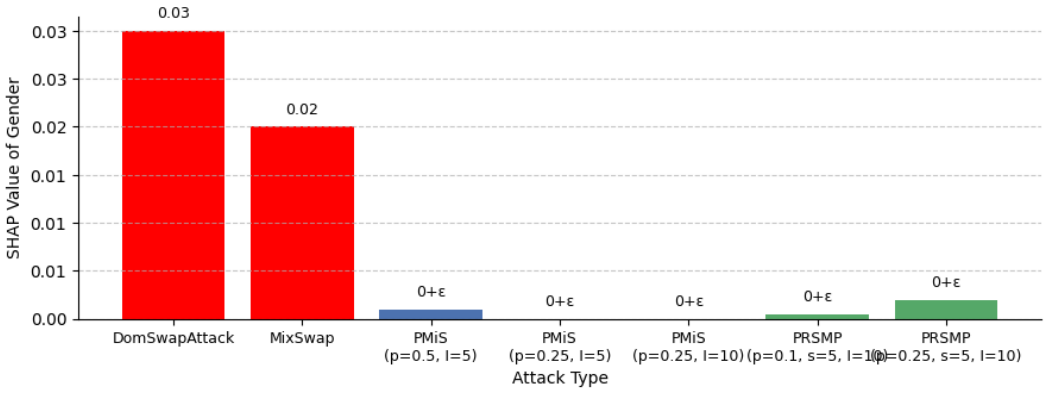

Figure 1: SHAP Values of the Protected Feature post-Attack for Bangladeshi Diabetes Dataset (LR Model)

4 Conclusion

The TIRA attacks show that the absence of a strong attribution signal does not necessarily equate to fairness or lack of adversarial perturbation. Crucially, our findings demonstrate a dual-pronged threat not achieved by prior work. TIRA can fool both AIF360 audits by pushing metrics to near-ideal values and XAI audits by reducing SHAP attribution to effectively zero. Our work challenges the assumption that trustworthy AI can be achieved through post-hoc explanation methods. The enhanced stealth of the TIRA attacks shows the need to develop integrity-based evaluation methodologies.

References

[1] A. B. Arrieta, N. Díaz-Rodríguez, J. Del Ser, A. Bennetot, S. Tabik, A. Barbado, S. García, S. Gil-López, D. Molina, R. Benjamins, R. Chatila, and F. Herrera. Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion*, 2020. URL https://arxiv.org/abs/1910.10045.

[2] R. K. E. Bellamy, K. Dey, M. Hind, S. C. Hoffman, S. Houde, K. Kannan, P. Lohia, J. Martino, S. Mehta, A. Mojsilovic, S. Nagar, K. N. Ramamurthy, J. Richards, D. Saha, P. Sattigeri, M. Singh, K. R. Varshney, and Y. Zhang. AI fairness 360: An extensible toolkit for detecting and mitigating algorithmic bias. *IBM Journal of Research and Development*, 2019. URL https://arxiv.org/abs/1810.01943.

[3] A. Das and P. Rad. Opportunities and challenges in explainable artificial intelligence (XAI): A survey. *arXiv preprint*, 2020. URL https://arxiv.org/abs/2006.11371.

[4] B. Dimanov, U. Bhatt, M. Jamnik, and A. Weller. You shouldn't trust me: Learning models which conceal unfairness from multiple explanation methods. In *ECAI*, 2020. URL https://ebooks.iospress.nl/pdf/doi/10.3233/FAIA200380.

[5] S. A. Friedler, C. Scheidegger, S. Venkatasubramanian, S. Choudhary, E. P. Hamilton, and D. Roth. A comparative study of fairness-enhancing interventions in machine learning. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, 2019. URL https://arxiv.org/abs/1802.04422.

[6] H. Hofmann. Statlog (german credit data). UCI Machine Learning Repository, 1994. DOI: https://doi.org/10.24432/C5NC77.

[7] M. Islam, R. Ferdousi, S. Rahman, and H. Y. Bushra. Likelihood prediction of diabetes at early stage using data mining techniques. In *Computer Vision and Machine Intelligence in Medical Image Analysis: International Symposium, ISCMM 2019*, 2020. URL https://link.springer.com/chapter/10.1007/978-981-15-2428-2_10.

[8] G. Laberge, U. Aïvodji, S. Hara, M. Marchand, and F. Khomh. Fool SHAP with stealthily biased sampling. In *ICLR*, 2023. URL https://arxiv.org/abs/2205.15419.

[9] S. Lundberg and S. I. Lee. A unified approach to interpreting model predictions. In *NeurIPS*, 2017. URL https://arxiv.org/abs/1705.07874.

[10] M. T. Ribeiro, S. Singh, and C. Guestrin. "why should i trust you?": Explaining the predictions of any classifier. In *KDD*, 2016. URL https://arxiv.org/abs/1602.04938.

[11] C. Rudin. Stop explaining black box machine learning models for high-stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 2019. URL https://pmc.ncbi.nlm.nih.gov/articles/PMC9122117/pdf/nihms-1058031.pdf.

[12] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra. Grad-CAM: Visual explanations from deep networks via gradient-based localization. *IJCV*, 2019. URL https://arxiv.org/abs/1610.02391.

[13] D. Slack, S. Hilgard, E. Jia, S. Singh, and H. Lakkaraju. Fooling LIME and SHAP: Adversarial attacks on post hoc explanation methods. In *AAAI/ACM Conference on Artificial Intelligence, Ethics, and Society (AIES)*, 2020. URL https://arxiv.org/abs/1911.02508.

[14] T. Speicher, H. Heidari, N. Grgic-Hlaca, K. P. Gummadi, A. Singla, A. Weller, and M. B. Zafar. A unified approach to quantifying algorithmic unfairness: Measuring individual & group unfairness via inequality indices. In *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining Proceedings*, 2018. URL https://arxiv.org/abs/1807.00787.

[15] M. Sundararajan, A. Taly, and Q. Yan. Axiomatic attribution for deep networks. In *ICML*, 2017. URL https://arxiv.org/abs/1703.01365.

[16] J. Yuan and A. Dasgupta. Fooling SHAP with output shuffling attacks. In *AAAI*, 2024. URL https://arxiv.org/abs/2408.06509.

A Related Work

Explainability aims to render opaque deep learning models understandable and transparent [11]. Rooted in co-operative game theory, SHAP stands as a cornerstone of post-hoc explainability. SHAP offers a theoretically grounded framework for attributing the contribution of each feature to each model output [9]. Beyond SHAP, other well-known post-hoc explainability methods include LIME [10], Integrated Gradients [15], and Grad-CAM [12]. Crucially, these methods identify the potential biases from the protected features [1, 3].

Complementing XAI, the field of algorithmic fairness is committed to identifying, quantifying, and mitigating biases in machine learning models. Several fairness metrics, such as demographic parity, equal opportunity difference, equal odds difference, and disparate impact, have been formulated to assess the disparities in model outcomes across demographic groups [14]. Frameworks like the AI Fairness (AIF360) toolkit [2] provide open-source and standardized implementations for assessing these fairness metrics. Although quantitative metrics offer a lens into fairness, their reliability hinges on the robustness of the model's outputs and the methods used to understand them.

Early efforts, such as scaffolding attacks, demonstrated that a classifier could be constructed to deceive LIME and SHAP to hide the reliance on protected features [13]. However, these methods often necessitate access to training data or the ability to retrain the models, which present significant limitations for model distributions or external auditors operating with black box access. This highlighted a critical need for more practical, data-agnostic attacks. Other methods use stealthy bias sampling [8] or learning models that hide unfairness from multiple explanation methods [4].

A more directly relevant line of inquiry is output sampling attacks, which shuffle the model's ranked outputs or prediction scores without modifying the input data distribution or the internal structure of the model [16]. This work made a vital contribution by proposing a family of attacks that theoretically proved that Shapley values cannot intrinsically detect shuffling attacks due to their order-agnostic nature in expectation calculations. However, their own empirical studies revealed that practical SHAP estimation algorithms could detect these attacks with varying degrees of effectiveness. Furthermore, these attacks can collapse into one another (i.e., 'Swapping' can become equivalent to 'Dominance'). What was learned from their contribution is that even though shuffling attacks are potent, their insufficiency in some contexts lies in their more overt patterns that might still be detectable by a vigilant auditor or other consistency checks.

We also observe two indicators of brittle control with [16]: (i) different shuffling variants can produce identical post-attack fairness values, and (ii) the attack sometimes fails to change fairness at all. This observed lack of robust, fine-grained control also motivates our work.

Our work advances the current state of shuffling attacks by introducing an enhanced paradigm of subtlety and control. Unlike the existing attacks, TIRA attacks are fundamentally probabilistic at a granular level. Their design makes the adversarial perturbations distributed and cumulative, which is a key differentiator and crucial for situations where detectability is the primary concern of the attacker. Our work explores the impact of attack-specific parameters, giving insights into how the degree of manipulation can be controlled, a level of explicit tunability not documented in prior literature.

B Extended Methodology

B.1 Targeted Identity Re-Association Attacks

B.1.1 Probabilistic Micro-Shuffling (PMiS) Attack

The PMiS attack uses the sorted list of individuals and their prediction scores in descending order. For a binary protected feature, the algorithm examines the adjacent pair of individuals in each iteration. The algorithm flips the disadvantaged individual with the advantaged individual with a predefined probability $p_{swap} \in [0, 1]$. This process is repeated for a predefined number of iterations. The subtlety of PMiS attacks is the result of its strict locality, probabilistic execution, and iterative accumulation. This leads to cumulative and gradual drift in the distribution of the protected attribute relative to score, which makes it challenging for SHAP or other auditing methods to catch the presence of an adversarial pattern.

B.1.2 Probabilistic Rank-Shift Micro-Perturbation (PRSMP) Attack

Like the PMiS attack, the PRSMP attack algorithm operates on a sorted list. PRSMP attack offers broader locality by introducing an additional layer of randomness with variable shifts. The attack targets the disadvantaged individual with a probability $p_{shift} \in [0, 1]$, and nudges its identity within a predefined small window, k_{max}. PRSMP also relies on multiple iterations, as a result of which these small probabilistic shifts accumulate into a gradual, but significant rearrangement of identities.

B.2 Experimental Setup

To rigorously evaluate the subtlety and efficacy of the TIRA attacks, we conducted experiments across diverse model architectures and real-world datasets to ensure reproducibility.

B.2.1 Datasets

We utilize two publicly available datasets, which are widely used in fairness assessments. Both datasets underwent standard preprocessing. Categorical features are one-hot encoded, and target variables are mapped to a binary format. **Bangladeshi Diabetes Dataset:** This dataset, which focuses on predicting diabetes risk, contains 520 patient records with clinical and demographic information [7]. The protected feature is binarized Gender, where male is the advantaged group and female is the disadvantaged group. **German Credit Dataset:** This dataset comprises financial and demographic features for credit assessment. The protected feature in this 1000 loan applicant dataset is Gender, with females designated as the disadvantaged group, while males as the advantaged group [6].

B.2.2 Models

We evaluate the TIRA attacks on two representative machine learning models to assess the generalizability of the algorithms. Both models are trained on 80% of the respective dataset, while the remaining 20% is reserved for testing and evaluation. **Logistic Regression:** Logistic regression is chosen as a baseline for its inherent interpretability, allowing for a clear understanding and evaluation of the efficacy of the attacks. **Neural Networks:** As a non-linear black-box model consisting of multiple dense layers with ReLU activations, and sigmoid activation for binary classification.

B.3 Fairness Metrics

We quantitatively assess the impact of TIRA attacks on the perceived fairness by employing five widely used fairness metrics from the AIF360 toolkit. The model's prediction scores are binarized for all the metrics using a fixed threshold. The ideal value of all the metrics is 0, except Disparate Impact. **Demographic Parity Difference:** A fairness metric used to measure the difference in the proportion of favorable outcomes between the advantaged and disadvantaged groups. **Equal Opportunity Difference:** Measures the difference in the true positive rates (recall) between the advantaged and disadvantaged groups. **Disparate Impact:** This metric measures the ratio of favorable outcomes between the advantaged and disadvantaged groups. Here, the ideal value is 1. **Odds Difference:** A fairness metric used to measure the average difference in the true positive rates and false positive rates between the advantaged and disadvantaged groups. **Theil Index:** Between-group generalized entropy error, a measure of inequality within an allocation, where an ideal value indicates more equality.

B.4 SHAP Attribution Analysis

We employ SHAP to assess the effectiveness of TIRA attacks on post-hoc interpretability to highlight how they confound SHAP. To get exact values, we used the permutation-based explainer. This also ensures cross-modal comparability and avoids inductive bias. SHAP was applied to both the original model's predictions, as a baseline representing the inherent degree of fairness, particularly for the protected feature, and the manipulated outputs to visualize and quantify the value of the protected feature.

B.5 Comparative Benchmarking

To further contextualize the performance of TIRA attacks, we demonstrate the distinct advantages of TIRA attacks by benchmarking their performance against DomSwap and MixSwap attacks, two representative state-of-the-art output shuffling attacks. Our comparisons shed light on the trade-offs between more overt shuffling and our probabilistic micro-level strategies in terms of inferred subtlety, their effectiveness in shifting fairness metrics, and SHAP's value attribution.

C Additional Results

C.1 Fairness Metrics Manipulation

To further validate our work, we provide a summary of additional results in Tables 3 and 4. To see the joint tradeoff between fairness manipulation and stealth, we plot the Demographic Parity against Disparate Impact. In Figure 2, the points nearer $(x \approx 1, y \approx 0)$ represent attacks that are stealthy and make the model look fair. In particular, PMiS often achieves stronger reductions in Demographic Parity. PRSMP, in contrast, usually tends to preserve the ratios, thereby enhancing stealth.

Table 3: Fairness Metrics Values on Bangladeshi Diabetes Dataset (NN Model)

Metric	**Baseline**	**DomSwap**	**MixSwap**	**PMiS**			**PRSMP**	
				(p=0.25, I=10)	(p=0.25, I=5)	(p=0.5, I=5)	(p=0.1, r=5, I=10)	(p=0.25, r=5, I=10)
Equal Opportunity	0.00	-0.03	0.01	-0.11	-0.03	-0.11	-0.09	-0.15
Demographic Parity	0.54	0.49	0.45	**0.28**	**0.45**	**0.24**	**0.42**	**0.34**
Equal Odds	0.00	-0.05	-0.05	-0.16	-0.05	-0.18	-0.07	-0.12
Disparate Impact	2.23	2.10	1.97	**1.53**	**1.97**	**1.44**	**1.91**	**1.68**
Theil Index	0.00	0.00	0.01	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**

Table 4: Fairness Metrics Values on German Credit Dataset (LR Model)

Metric	**Baseline**	**PMiS**			**PRSMP**			
		(p=0.25, I=25)	(p=0.25, I=10)	(p=0.33, I=25)	(p=0.25, r=5, I=10)	(p=0.33, r=5, I=10)	(p=0.25, r=10, I=10)	(p=0.25, r=5, I=5)
Equal Opportunity	0.13	**0.08**	**0.11**	**0.04**	**0.06**	**0.01**	**0.02**	**0.06**
Demographic Parity	0.09	**0.05**	**0.08**	**0.01**	**0.02**	**-0.01**	**-0.03**	**0.03**
Equal Odds	0.12	**0.08**	**0.10**	**0.04**	**-0.05**	**0.01**	**0.02**	**0.05**
Disparate Impact	1.13	**1.08**	**1.11**	**1.02**	**1.04**	**0.98**	**0.95**	**1.04**
Theil Index	0.00	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**

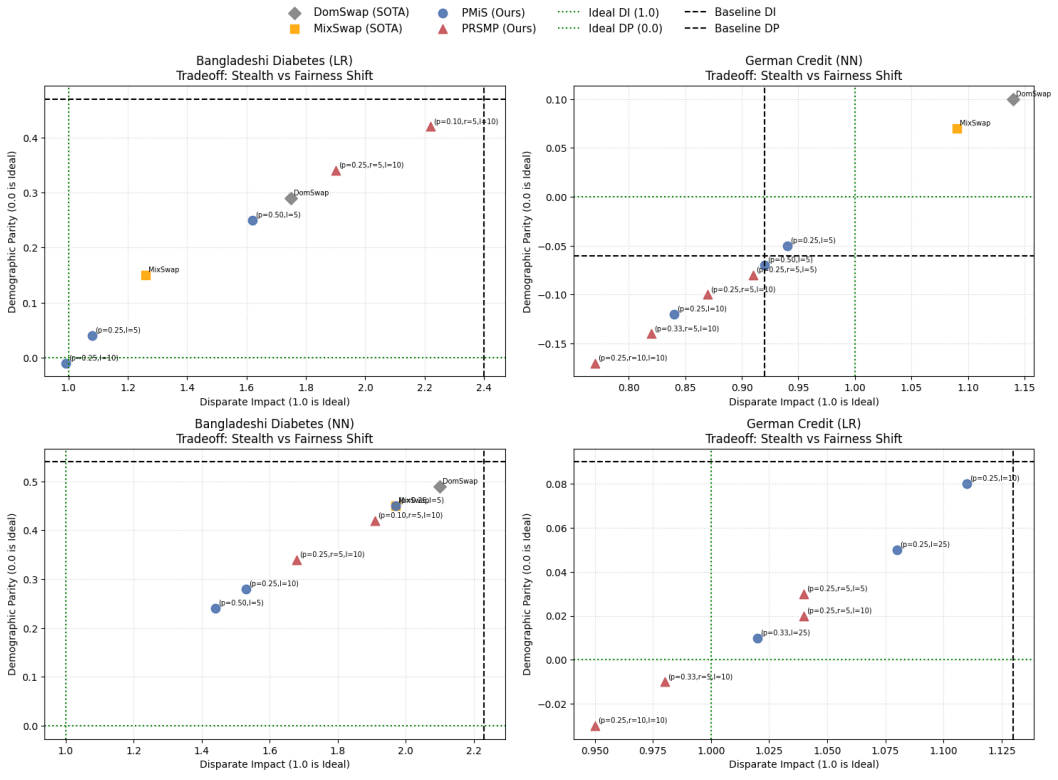

Figure 2: Tradeoff Curves between Disparate Impact (x-axis) and Demographic Parity (y-axis) across Four Dataset-Model Combinations

C.2 SHAP Attribution Analysis

Additional SHAP analysis are summarized in Figure 3, providing further support for our work.

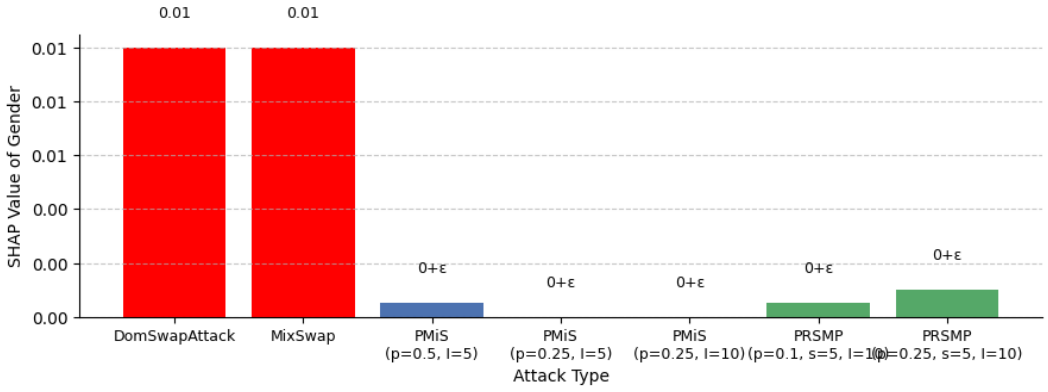

Figure 3: SHAP Values of the Protected Feature post-Attack for Bangladeshi Diabetes Dataset (NN Model)

D Discussion

Our work introduces a new family of shuffling attacks. The enhanced strength of our attacks lies in their micro-level, probabilistic, and iterative nature. Our findings demonstrate that it is possible to significantly alter the perceived fairness and undermine SHAP-based explanations by creating effective adversarial attacks that are stealthy.

D.1 The Nuance of Parametric Control

A key contribution of our work is that TIRA attacks are not blunt instruments, but rather a highly tunable class of attacks. This parameter-driven control demonstrates that an adversary can precisely calibrate the attacks' subtlety and intensity. The ability to 'dial-in' a specific level of perceived fairness while confounding SHAP highlights a previously undocumented attack surface. We systematically explored the correlation between these parameters and the outputs, revealing a scalable influence on the attack's efficacy and subtlety from almost imperceptible changes to more pronounced changes.

Probabilistic control: The p_{swap} (PMiS) and p_{shift} (PRSMP) parameters empower the adversary to control the frequency of perturbations. Increasing the probability leads to a more pronounced shift in fairness metrics. Low probability value ensures that the manipulations are distributive and cumulative.

Locality: The k_{max} (PRSMP) defines the locality of shifts, acting as a knob for controlling the locality of the attack. This parameter allows the adversary to perform perturbation either within a small window to maintain stealth or expend the window for more aggressive shifts.

Cumulative Effect: The number of iterations is directly proportional to the strength of the attack. It allows for the temporal dimension of the attack. An attack with low probability over an extended period ensures that the attack is undetectable, but the impact is significant.

D.2 Limitations and Future Work

While our work presents a robust framework for TIRA attacks that expose the adversarial vulnerabilities in fairness tools and post-hoc explainability, several avenues for future research are still unexplored.

Expanding Attack Vectors: Our study is effective on binary and continuous protected features. Future work should extend the applicability of such attacks to more complex datasets and explainability tools.

Theoretical Guarantees: Our work empirically demonstrates the TIRA attacks' efficacy, but a deeper theoretical investigation is required. Proving the asymptotic properties of such attacks as well as formalizing their relationship to SHAP's estimation algorithms would provide a much-needed theoretical foundation.

Developing Robust Defenses: The most pressing next step is the development of robust defenses. Future work should focus on designing auditing frameworks that can detect inconsistencies between a model's outputs and its explanations.