FairSHAP: Preprocessing for Fairness Through Attribution-Based Data Augmentation

Lin Zhu[*]
Technical University of Denmark

Yijun Bian[*]
University of Copenhagen

Lei You[†]
Technical University of Denmark

Abstract

Ensuring fairness in machine learning models is critical, particularly in high-stakes domains where biased decisions can lead to serious societal consequences. However, existing preprocessing approaches generally lack transparent mechanisms for identifying which features are responsible for unfairness. This obscures the rationale behind data modifications. We introduce FairSHAP, a novel preprocessing framework that leverages Shapley value attribution to improve both individual and group fairness. FairSHAP identifies fairness-critical features in the training data using an interpretable measure of feature importance, and systematically modifies them through instance-level matching across sensitive groups. Our method effectively reduces discriminative risk (DR) with an instance-wise guarantee up to an interaction residual term, which is bounded under local matching, while simultaneously bounding the upper limit of demographic parity (DP), which in practice leads to its reduction. Experiments on multiple tabular datasets show that we achieve state-of-the-art or comparable performance across DR, DP, and equality of opportunity (EO) with minimal modifications, thereby preserving data fidelity. As a model-agnostic and transparent method, FairSHAP integrates seamlessly into existing machine learning pipelines and provides actionable insights into the sources of bias. Our code is available on `https://github.com/ZhuMuMu0216/FairSHAP`.

[*]Equal contributions. Proceedings of the 29th International Conference on Artificial Intelligence and Statistics (AISTATS) 2026, Tangier, Morocco. PMLR: Volume 300. Copyright 2026 by the author(s). [†]Correspondence: `leiyo@dtu.dk`

1 INTRODUCTION

Machine learning (ML) has become one of the most prominent and rapidly evolving fields in artificial intelligence (AI), with researchers continuously developing methods to improve model accuracy and reliability. Yet in sensitive domains such as healthcare, finance, and law, prioritizing accuracy alone may not serve society's best interests; fairness and interpretability are equally critical. Well-documented cases, including racial bias in criminal risk assessment tools and gender bias in automated hiring systems [ProPublica, Angwin et al., 2016, Dastin, 2018], highlight how ML models can perpetuate inequities and erode trust. These examples underscore the need for responsible AI approaches that balance predictive performance with fairness and transparency.

To address bias in ML models, mitigation methods are often categorized by their intervention stage—preprocessing, in-processing, and post-processing [Mehrabi et al., 2021, Binns, 2018, Caton and Haas, 2024]. **Preprocessing** modifies training data before model learning [Hort et al., 2024]. Typical strategies include perturbation of features or distributions [Weerts et al., 2023, Wang et al., 2019, Calmon et al., 2017, Feldman et al., 2015], reweighting of group instances [Calders and Žliobaitė, 2013, Kamiran and Calders, 2012, Arnaiz-Rodriguez and Oliver, 2023], fair representation learning through invariant embeddings [Zemel et al., 2013, Louizos et al., 2015], and resampling to balance group distributions [Shekhar et al., 2021, Ustun et al., 2019, Bastani et al., 2019]. While widely applicable, these methods risk overfitting, data loss, or reduced fidelity when perturbations are excessive. **In-processing** integrates fairness as constraints into training. Approaches include adversarial debiasing [Grari et al., 2023, Zhang et al., 2018, Madras et al., 2018], constrained optimization [Zafar et al., 2017, Agarwal et al., 2018], fairness-aware regularization [Kamishima et al., 2012, Donini et al., 2018, Zhao et al., 2022], reweighted gradient updates [Wan et al., 2023], and counterfactual fairness [Kusner et al., 2017, Chen and Zhu, 2025]. **Post-processing** adjusts model outputs after train-

ing. Common methods include threshold tuning for equalized odds [Hardt et al., 2016], probability calibration [Dwork et al., 2018], and boundary modification in uncertain regions [Petersen et al., 2021]. Among these, preprocessing methods are often model-agnostic and widely applicable, but it must carefully balance fairness gains against preserving data integrity and predictive fidelity [Mehrabi et al., 2021, Weerts et al., 2023].

Despite extensive work on fairness in ML, two key gaps remain. First, fairness and interpretability are often studied in isolation, even though interpretability is essential for uncovering hidden biases and enabling transparent interventions; without their integration, explanations risk being misleading or even "fairwashed*" [Doshi-Velez and Kim, 2017, Poursabzi-Sangdeh et al., 2021, Lakkaraju et al., 2017, Aïvodji et al., 2019]. Second, most methods focus narrowly on group fairness metrics such as demographic parity or equalized odds [Mehrabi et al., 2021], thereby overlooking intersectional biases [Kearns et al., 2018], the incompatibility of fairness definitions [Kleinberg, 2018], and the importance of individual-level fairness [Dwork et al., 2012].

Building on these limitations, recent work shows that optimizing only for group fairness can lead to individual arbitrariness [Long et al., 2023]. Prior research further suggests that individual and group fairness, rather than being contradictory, are different manifestations of the same underlying principles of consistency and egalitarianism [Binns, 2020]. Under certain conditions, ensuring fairness for individuals can also promote fairness across groups. We therefore argue that strengthening individual fairness provides a principled path toward achieving both individual- and group-level fairness. *Our approach aims to improve both individual and group fairness.*

Motivated by this insight, we propose an innovative approach that explicitly emphasizes improving individual fairness, hypothesizing that addressing individual-level biases can concurrently enhance group-level fairness. *Our approach positions fairness enhancement as an interpretability-driven task*, where transparency in identifying bias at the individual level facilitates targeted and principled mitigation. To operationalize this concept, we utilize Shapley values, a widely recognized interpretability metric that quantifies feature importance in individual predictions. Shapley values allow us to pinpoint precisely which features and data points contribute most significantly to fairness disparities. Through targeted adjustments informed by Shapley values, our method ensures data integrity and interpretability while effectively reducing bias. Our main contributions are as follows:

- We propose FairSHAP, a novel fairness enhancement method leveraging Shapley value analysis and instances matching. It systematically identifies fairness-sensitive data points using individual fairness measures and Shapley values, selectively modifying these points to enhance fairness.
- FairSHAP consistently achieves superior fairness improvements with minimal data perturbation, surpassing existing benchmark methods, without requiring extra transformations on the test set.
- FairSHAP effectively reduces both individual and group fairness measures (DR, DP, EO) across diverse datasets, with negligible accuracy loss and occasional improvements in predictive performance.
- As a preprocessing technique, FairSHAP integrates seamlessly into existing ML workflows, providing fully reproducible and transparent implementations.

By aligning fairness enhancement closely with interpretability, our proposed approach paves the way for more accountable, transparent, and equitable ML systems.

2 PRELIMINARIES

We summarize some necessarily preliminary knowledge here before elaborate on our methodology.

2.1 Shapley value and relevant methods

The Shapley value is a foundational concept in cooperative game theory, originally introduced to fairly distribute a collective payoff among players based on their individual contributions. Given its axiomatic fairness properties, the Shapley value has been widely adopted in various domains, including ML explainability, resource allocation, and feature attribution [Sundararajan and Najmi, 2020].

For a set of players (or elements) $\mathcal{N} = \{1, 2, \ldots, n\}$ and a characteristic function $v : 2^{\mathcal{N}} \to \mathbb{R}$ that assigns a value to each coalition (or combination) $S \subseteq \mathcal{N}$, the Shapley value of player k is defined as:

$$\phi_k(v) = \sum_{S \subseteq \mathcal{N} \setminus \{k\}} \frac{|S|!(n-|S|-1)!}{n!} \big(v(S \cup \{k\}) - v(S)\big) . \tag{1}$$

This formula ensures a fair allocation by averaging the marginal contribution of each player over all possible coalition formations. The key properties of the Shapley value include *efficiency*, *symmetry*, *dummy player*, and *additivity*, making it a unique solution satisfying these fairness criteria.

Reference data In SHAP (SHapley Additive exPla-

*The term "fairwashing" was popularized by Aïvodji et al. (2019) to describe the risk of rationalizing biased decisions through deceptive explanations.

nations) library [Lundberg and Lee, 2017], reference data serves as a baseline distribution used to simulate the absence of features when computing Shapley values. By replacing features of the instance to be explained with values sampled from the reference data, SHAP estimates the marginal contribution of each feature. The selection of reference data significantly influences the interpretation results.

Baseline Shapley This method is proposed to address the issue that feature values cannot be empty when computing Shapley values using ML [Lundberg and Lee, 2017, Merrick and Taly, 2020, You et al., 2024]. It relies on a specific reference data point $\mathbf{r}_j$ for any instance $\mathbf{x}_i$, where $\mathbf{x}_i$ denotes the i-th instance, and the set function is defined accordingly:

$$v_{\mathrm{B}}^{(i,j)}(S) = f(\mathbf{x}_{i,S}; \mathbf{r}_{j,\mathcal{F}\setminus S}) - f(\mathbf{r}_j)\,, \tag{2}$$

where a feature's absence is modeled using its value in the reference baseline data point $\mathbf{r}$, and $\mathcal{F}$ denotes the set of all features.

Random baseline Shapley This method is proposed to address the issue of feature absence in Shapley value computation by using a random sampling approach from a reference distribution $\mathcal{D}$ [Lundberg and Lee, 2017], which is a variant of Baseline Shapley and is applied in the famous SHAP library. Instead of relying on a single reference data point, it uses a reference distribution to model missing feature values. The set function is defined as follows:

$$v_{\mathrm{RB}}^{(i)}(S) = \mathbb{E}_{\mathbf{x}'\sim\mathcal{D}}\left[f(\mathbf{x}_{i,S}; \mathbf{x}'_{\mathcal{F}\setminus S})\right] - \mathbb{E}_{\mathbf{x}'\sim\mathcal{D}}\left[f(\mathbf{x}')\right]\,. \tag{3}$$

2.2 Discriminative Risk (DR), an individual fairness measure

Since Shapley value can only be applied to individual instances for feature attribution based on the output results, we consider combining it with individual fairness measures rather than group fairness measures. This approach is more suitable for analyzing single instances and enhances the focus on and evaluation of individual fairness, while also providing a means to further mitigate potential biases inherent in group fairness measures.

This approach allows us to leverage both the original data and background data to select feature values that can effectively reduce biases in the training set. In previous work, Bian and Zhang [2023] proposed a critical individual fairness measure named *discriminative risk (DR)*, and we show it (denoted as L_{DR}) as follows:

$$L_{\mathrm{DR}}(f, \mathbf{x}) = \left|\, f(\check{\mathbf{x}}, A=0) - f(\check{\mathbf{x}}, A=1) \,\right| \le \varepsilon\,. \tag{4}$$

Here, let $\mathcal{F}$ denote the set of all features. An instance is written as $\mathbf{x} = (\check{\mathbf{x}}, A)$, where $\check{\mathbf{x}} \in \mathcal{X}_{\mathcal{F}\setminus\{A\}}$ represents the non-sensitive features and $A \in \{0,1\}$ is the sensitive attribute, with 0 **corresponding to the unprivileged group(s) and** 1 **to the privileged group**. A higher $L_{\mathrm{DR}}(f, \mathbf{x})$ indicates that the model's predictions are overly sensitive to changes in the sensitive attribute, potentially reflecting bias.

For definitions of other group fairness measures, such as demographic parity (DP), equality of opportunity (EO) and predictive quality parity (PQP), please refer to Appendix B. The group fairness measures discussed in this paper quantify the disparities between groups. A higher value indicates a greater degree of unfairness; therefore, our objective is to minimize these measures.

3 THE PROPOSED FAIRSHAP AND ITS THEORETICAL FOUNDATION

Bias in machine learning models typically stems from two sources: data bias and algorithmic bias. Instead of addressing them separately, we propose tailoring preprocessing strategies to each algorithm on the same dataset, thereby jointly mitigating both sources of bias. This approach is novel in that it adapts preprocessing to algorithm-specific characteristics while remaining independent of model internals—model-agnostic at the architectural level, but algorithm-adaptive in design. In doing so, it retains broad applicability while offering a more principled way to integrate data and algorithmic considerations in fairness enhancement.

3.1 Framework of FairSHAP

To address this problem, we propose FairSHAP, a preprocessing method that enhances individual fairness by identifying fairness-critical positions in the training set using Shapley values, followed by targeted data augmentation. The method aims to reduce the individual fairness measure DR as a primary objective, while also leading to reductions in group fairness measures such as DP and EO. Unlike generic data preprocessing approaches that operate without any model information, our method consults the model once at the beginning to obtain prior knowledge about fairness-critical features. Importantly, this single interaction does not modify the model itself; it only guides the preprocessing step to produce a minimally modified dataset. After this preprocessing stage, the model will be subsequently trained on adjusted new dataset, ensuring that FairSHAP remains a data-level intervention rather than a model-level modification. Therefore, we regard FairSHAP as belonging to the preprocessing approaches.†

†Feature engineering is standard preprocessing in traditional ML, even when guided by model feedback. Thus, using model information at the preprocessing stage does not contradict the notion of preprocessing.

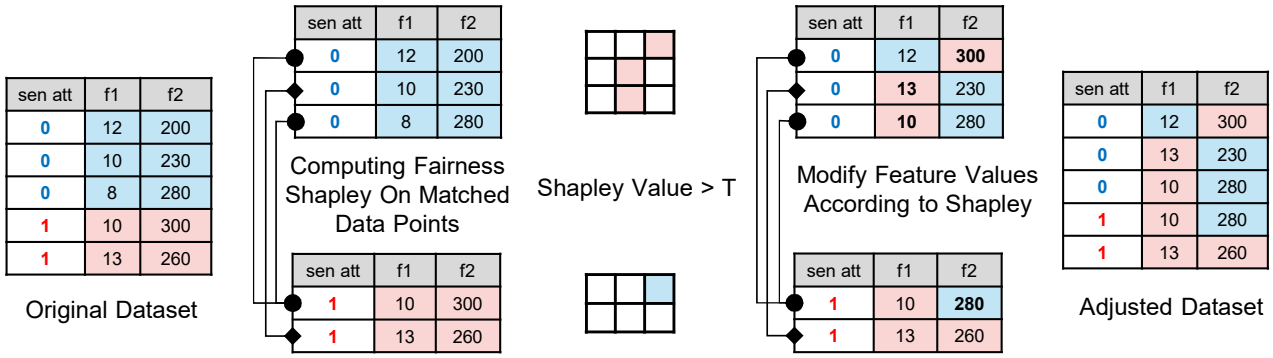

Figure 1: Overall framework of FairSHAP. (Left) Training data are first split by sensitive attribute and aligned via nearest-neighbor matching to produce paired instances. (Right) For each target group, feature values whose Shapley value exceeds a threshold are adjusted to reduce DR, and the modified instances from both groups are recombined into an augmented, fairness-improved training set.

FairSHAP consists of two interconnected components, which is presented in Figure 1. The left-hand side of the illustration corresponds to the first component, while the right-hand side depicts the second component. These components work together to identify and mitigate bias in the training data. Below, we describe each component in detail:

(1) On the left side of the figure, the process begins by dividing the training set into two groups (privileged group and unprivileged group) based on the sensitive attribute (e.g., gender, race, etc.). To ensure a precise mapping relationship between the two groups, a nearest neighbor matching procedure is applied. This step aligns instances from one group with similar instances in the other group, creating matched pairs that facilitate a more equitable analysis. This alignment is visually emphasized in the left portion of the figure, highlighting the pairing process.

(2) On the right side of the figure, the focus shifts to leveraging these matched pairs for bias mitigation. Specifically, one group is selected as the target group, while the other group serves as the reference data for computing Shapley values for instances within the target group. These Shapley values are combined with DR to identify specific feature values that contribute to fairness disparities. Once identified, these feature values are modified to reduce DR. The process is repeated with the roles of the two groups reversed, ensuring that each group can effectively learn useful feature values from the other. This bidirectional learning approach helps to capture a more balanced and comprehensive representation of the data, ultimately improving fairness and reducing disparities. Finally, the modified instances from both groups are combined to form a new augmented training set, which is designed to reduce biases and improve fairness.

Algorithm 1 Overall framework: Enhancing fairness via matching and Shapley values

Input: Model f, dataset $\mathcal{D} \in \mathbb{R}^{(n+m)\times d}$ with sensitive attribute $A \in \{0,1\}$, threshold T matching method $\mathcal{M}_{\text{method}}$, where method $\in$ {NearestNeighbor, OptimalTransport}

Output: $\mathcal{D}_{\text{new}}$

1: Split $\mathcal{D}$ into two subgroups: $\mathcal{G} \in \mathbb{R}^{n\times d}$ (e.g., $A=0$) and $\widetilde{\mathcal{G}} \in \mathbb{R}^{m\times d}$ (e.g., $A=1$)

2: $\mathcal{G}' \leftarrow$ `FairSHAP`(target $=\mathcal{G}$, non-target $=\widetilde{\mathcal{G}}$, model $=f$, T, $\mathcal{M}_{\text{method}}$) // see Algorithm 2

3: $\widetilde{\mathcal{G}}' \leftarrow$ `FairSHAP`(target $=\widetilde{\mathcal{G}}$, non-target $=\mathcal{G}$, model $=f$, T, $\mathcal{M}_{\text{method}}$) // see Algorithm 2

4: $\mathcal{D}_{\text{new}} \leftarrow \text{Concat}(\mathcal{G}', \widetilde{\mathcal{G}}') \in \mathbb{R}^{(n+m)\times d}$

5: **return** $\mathcal{D}_{\text{new}}$

Overall framework: The detailed overall process is described in the following algorithm 1. Given a dataset $\mathcal{D} \in \mathbb{R}^{(n+m)\times d}$, we partition it into two disjoint groups (privileged group and unprivileged group), $\mathcal{G} \in \mathbb{R}^{n\times d}$ and $\widetilde{\mathcal{G}} \in \mathbb{R}^{m\times d}$, based on the sensitive attribute $A \in \{0,1\}$, where 0 and 1 represent the unprivileged and privileged groups, respectively. Note that the unprivileged group(s) in the original dataset may possibly include multiple values. Subsequently, we treat $\mathcal{G}$ and $\widetilde{\mathcal{G}}$ as the target group in turn and input them into FairSHAP (see algorithm 2), obtaining the modified groups $\mathcal{G}'$ and $\widetilde{\mathcal{G}}'$, which are then concatenated to form $\mathcal{D}_{\text{new}}$ for training a new model. In this process, we also set the threshold T that determines which Shapley values trigger modifications. We set T=0.05 as a default threshold balancing fairness improvement and data fidelity. Empirically, modification counts concentrate within $\phi \in [0.01, 0.05]$ and show diminishing returns beyond this range, while too small/large thresholds respectively introduce many near-zero edits or miss

Algorithm 2 `FairSHAP`

Input: Target group $\mathcal{G} \in \mathbb{R}^{n\times d}$ and non-target group $\widetilde{\mathcal{G}} \in \mathbb{R}^{m\times d}$, model f, threshold T, matching method $\mathcal{M}_{\text{method}}$
Output: modified dataset $\mathcal{G}'$

```
Use M_method(G, G̃) to obtain joint probability P(g, g̃) ∈ R^{n×m}
Use Eqs. (1) and (5) to obtain Shapley value matrix φ ∈ R^{n×d}
Initialize reference data B ← 0_{n×d}
for i = 1 to n do
    j* ← arg max_{1≤j≤m} P_{i,j}
    B_{i,:} ← G̃_{j*,:}
Let G' ← G(G' ∈ R^{n×d})
for i = 1 to n do
    for k = 1 to d do
        if φ_{i,k} ≥ T then
            G'_{i,k} ← B_{i,k}
return G'
```

fairness-critical features (Appendix G, Fig. 4). We observe stable behavior for $T \in [0.03, 0.07]$ and thus use T=0.05 for all experiments.

Next, as shown in Algorithm 2, we denote by $\mathbf{g}$ and $\tilde{\mathbf{g}}$ the random variables drawn from $\mathcal{G}$ and $\widetilde{\mathcal{G}}$, respectively. For each $\mathbf{g} \in \mathcal{G}$, we identify its closest counterpart $\tilde{\mathbf{g}} \in \widetilde{\mathcal{G}}$, and normalize this matching so that it forms a joint distribution $\mathcal{P} \in \mathbb{R}^{n\times m}$ over the two groups. Importantly, when nearest-neighbor matching is conducted, we not only align instances across sensitive groups (e.g., male versus female), but also enforce label consistency, such that positive-labeled instances in one group are matched exclusively with positive-labeled instances in the other, and analogously for negative-labeled cases. Based on the matched pairs, we then compute the Shapley values ϕ_k, which quantify the average marginal contribution of the k-th feature in an instance to the model prediction relative to a baseline. Features whose Shapley values exceed a threshold T are regarded as fairness-critical and replaced with values from their matched references. This procedure is applied symmetrically to both groups, thereby ensuring balanced fairness improvements.

Finally, the modified groups $\mathcal{G}'$ and $\widetilde{\mathcal{G}}'$ are combined to create a new dataset $\mathcal{D}_{\text{new}}$, which is used for retraining the model with improved fairness. Note that

$$v^{(i)}(S) = \mathbb{E}_{\tilde{\mathbf{g}}\sim\mathcal{P}(\tilde{\mathbf{g}}|\mathbf{g}_i)}\left[L_{\text{DR}}(\mathbf{g}_{i,S}; \tilde{\mathbf{g}}_{\mathcal{F}\setminus S})\right] - \mathbb{E}_{\tilde{\mathbf{g}}\sim\mathcal{P}(\tilde{\mathbf{g}}|\mathbf{g}_i)}\left[L_{\text{DR}}(\tilde{\mathbf{g}}_{\mathcal{F}})\right], \tag{5a}$$

$$\text{s.t.} \quad \mathcal{P}(\mathbf{g}, \tilde{\mathbf{g}}) = \mathcal{M}_{\text{method}}(\mathcal{G}, \widetilde{\mathcal{G}}), \tag{5b}$$

where L_{DR} is the individual fairness measure, with DR being used as individual fairness in this paper, S represents a subset of features considered in the fairness evaluation, and $\mathcal{F}$ denotes the full set of features, with $\mathcal{F} \setminus S$ indicating the complement of S. The matching method can be instantiated as nearest neighbor, optimal transport, or any other appropriate similarity-based matching approach.

3.2 Theoretical foundation

We provide the theoretical guarantees of FairSHAP in reducing DR. The key idea is that features with large Shapley contributions are fairness-critical; replacing them with values from reference data can improve individual fairness.

Theorem 3.1 (Instance-wise DR reduction)**.** *Fix a threshold $T > 0$ and define the set of fairness-critical features as*

$$S_T(\mathbf{x}) = \{\, k \in [d] \;\mid\; \phi_k(\mathbf{x}) \geq T \,\},$$

where $[d]$ denotes $\{1, 2, ..., d\}$ for brevity. Let $\mathbf{x}_{\text{after}}$ denote the feature vector obtained from $\mathbf{x}$ after FairSHAP replaces every feature $k \in S_T(\mathbf{x})$ by its value in the matched reference $\mathbf{r}$. Then the DR satisfies

$$\begin{aligned} L_{\text{DR}}(f, \mathbf{x}_{\text{after}}) &= L_{\text{DR}}(f, \mathbf{x}_{\text{before}}) - \sum_{k\in\mathcal{S}_T(\mathbf{x})} \phi_k(\mathbf{x}) + \mathcal{I}(\mathbf{x}) \\ &\leq L_{\text{DR}}(f, \mathbf{x}_{\text{before}}) - T\,|\mathcal{S}_T(\mathbf{x})| + \mathcal{I}(\mathbf{x}). \end{aligned} \tag{6}$$

where $\mathcal{I}(\mathbf{x})$ denotes the interaction residual term.

This theorem guarantees that each replaced feature has a Shapley-assigned contribution of at least T toward reducing the instance-wise DR, up to the interaction residual $\mathcal{I}(\mathbf{x})$. The detailed proof is provided in appendix C.2. To rigorously justify that this interaction residual is negligible in practice, we establish the following proposition. It demonstrates that our use of Nearest Neighbor or Optimal Transport for local matching ensures $\mathcal{I}(\mathbf{x})$ decays as a second-order term, thereby validating the dominance of the Shapley attribution:

Proposition 3.2 (Bound on Interaction Residual)**.** *Assume the classifier f is twice differentiable with a bounded Hessian. Since FairSHAP uses Nearest Neighbor or Optimal Transport to select a reference $\mathbf{r}$ from the local neighborhood of $\mathbf{x}$, let $\epsilon = \|\mathbf{x} - \mathbf{r}\|$ be the matching distance. Then, the interaction residual is bounded by a second-order term:*

$$|\mathcal{I}(\mathbf{x})| \leq C \cdot \epsilon^2,$$

where C is a constant depending on the curvature of f. Consequently, as $\epsilon \to 0$, the residual $\mathcal{I}(\mathbf{x})$ vanishes faster than the Shapley attribution $\phi_k(\mathbf{x})$ (which is $O(\epsilon)$), theoretically justifying the approximation in Theorem C.2.

Proposition C.3 relies on the premise that the matching distance $\epsilon = \|\mathbf{x} - \mathbf{r}\|$ remains sufficiently small to

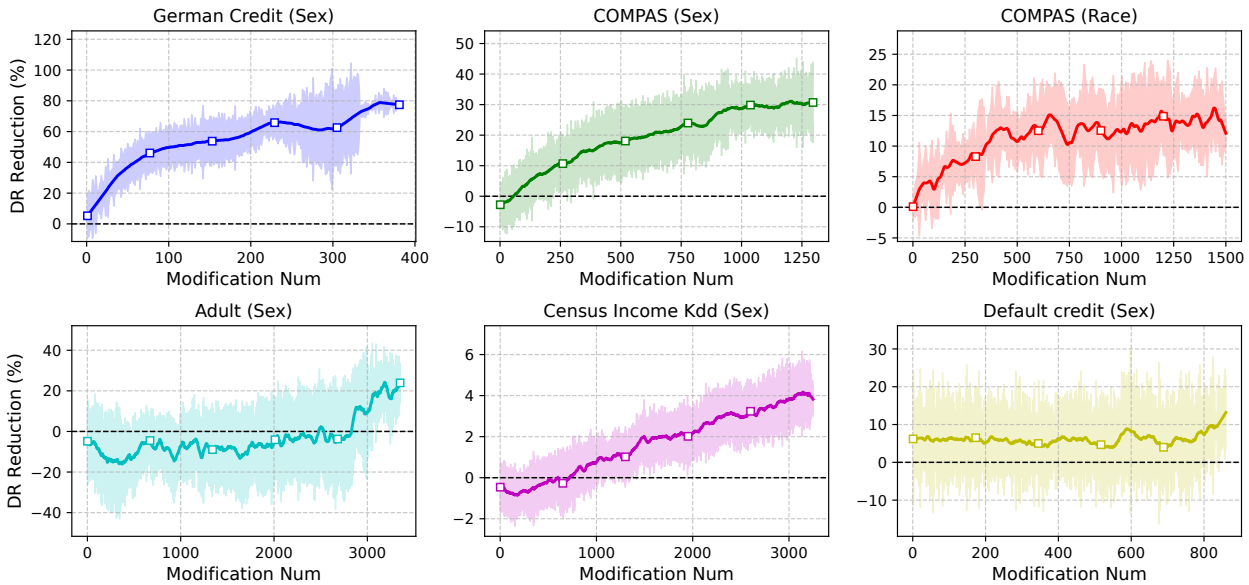

Figure 2: Percentage reduction in the DR across different datasets. The x-axis denotes the number of modifications applied (up to the maximum required under a fairness threshold $T = 0.05$), while the y-axis indicates the relative in DR, expressed as a percentage of the original value.

suppress higher-order terms. We empirically validate this assumption in Appendix E. Our analysis of the matching distance distribution confirms that the proposed Nearest Neighbor Matching strategy consistently selects reference samples within the local neighborhood of $\mathbf{x}$, thereby ensuring the validity of the approximation in practice.

Complexity (two-sided). Let n and m be the sample sizes of the two sensitive groups, and d the feature dimension. With nearest neighbor matching, building the dense $n \times m$ cross-group cost matrix costs $\mathcal{O}(nmd)$ time and $\mathcal{O}(nm)$ space. We estimate Shapley values using M permutations and R reference draws per set-function evaluation; let C_f be the cost of one forward pass of f. The total cost is $\mathcal{O}\big(nmd + (n{+}m)\, MdRC_f\big)$ time and $\mathcal{O}\big(nm + (n{+}m)d\big)$ space; see Appendix F for details.

Let $\mathcal{X}$ denote the feature space. Each instance is written as $\mathbf{x} = (\check{\mathbf{x}}, A)$, where $\check{\mathbf{x}} \in \mathcal{X}_{\mathcal{F}\setminus\{A\}}$ denotes the non-sensitive features. We denote by $D_a(\check{\mathbf{x}})$ the conditional distribution of non-sensitive features given $A = a \in \{0, 1\}$.

Theorem 3.3 (Identical-distribution case)**.** *If $D_0 = D_1$, then for any classifier f,*

$$L_{\mathrm{DP}}(f) \;\leq\; L_{\mathrm{DR}}(f)\,.$$

Theorem 3.4 (General case)**.** *For any classifier f, regardless of D_0 and D_1,*

$$L_{\mathrm{DP}}^{score}(f) \;\leq\; \min\{\, L_{\mathrm{DR}}(f; D_0),\, L_{\mathrm{DR}}(f; D_1)\,\} + \mathrm{TV}(D_0, D_1)\,.$$

where $\mathrm{TV}(D_0, D_1)$ is the total variation distance.

Theorems D.1 and D.3 establish that any uniform reduction in individual fairness measure L_{DR} induced by FairSHAP also yields a corresponding decrease in L_{DP}. Detailed proofs are deferred to Appendix D.

Thus, our theoretical analysis not only justifies the design of FairSHAP but also highlights its ability to simultaneously improve both individual and group fairness guarantees in a principled and interpretable manner.

4 EXPERIMENTS

We conducted a comprehensive experimental evaluation to assess the proposed method's performance across five critical fairness measures: accuracy, DR, DP, EO, and PQP.

Our experiments incorporates 5 typical tabular datasets spanning criminal justice, finance, and social services domains: COMPAS [ProPublica, Angwin et al., 2016], ADULT [Becker and Kohavi, 1996], German Credit [UCI Machine Learning Repository, 2020], Census Income KDD [UCI Machine Learning Repository, 2000], and Default Credit [Yeh, 2009].

The comparative analysis included a baseline model without bias mitigation mechanisms, and three benchmark preprocessing methods: Remove Sensitive Attribute (RSA), Disparate Impact Remover (DIR) [Feldman et al., 2015] and Correlation Remover (CR) [Weerts et al., 2023]. In the following, we refer to them simply as RSA, DIR and CR, respectively.

Table 1: Compare FairSHAP with other fairness mitigation methods across different datasets. Here, "s.a." denotes the sensitive attribute used in fairness evaluation, and "Kdd" is an abbreviation for the Census Income KDD dataset. Note that TrainingAR: Training Set Adjustment Rate; TestAN: Test Set Adjustment Necessity. Data Fidelity is measured using the Wasserstein Distance to quantify the difference between the original and adjusted data distributions. Best and second-best results for fairness and efficiency metrics (DR, DP, EO, PQP, Data Fidelity, TrainingAR) are **bold** and underlined, respectively.

Dataset (s.a.)	Methods	Accuracy	DR	DP	EO	PQP	Data Fidelity	TrainingAR	TestAN
German (sex)	Baseline	0.6650±0.0257	0.0785±0.0211	0.0512±0.0346	0.1287±0.0590	0.1341±0.0486	—	—	No
	RSA	0.6590±0.0287	—	0.0522±0.0246	0.1524±0.1007	0.2189±0.0815	—	—	Yes
	CR	0.6680±0.0238	**0.0028±0.0029**	0.0844±0.0557	0.1559±0.0609	**0.0723±0.0330**	0.0183±0.0211	0.9615	Yes
	DIR	**0.6720±0.0337**	0.0966±0.0112	0.0946±0.0373	0.1737±0.0729	0.1529±0.0634	0.0155±0.0440	0.0774	Yes
	FairSHAP	0.6630±0.0275	0.0243±0.0112	**0.0301±0.0347**	**0.1126±0.0783**	0.1852±0.1074	**0.0049±0.0085**	**0.0156**	**No**
COMPAS (sex)	Baseline	**0.6698±0.0051**	0.0883±0.0064	0.1548±0.0241	0.1243±0.0510	0.0492±0.0084	—	—	No
	RSA	0.6676±0.0067	—	**0.1075±0.0160**	**0.0842±0.0475**	0.0635±0.0266	—	—	Yes
	CR	0.6679±0.0045	**0.0082±0.0070**	0.1407±0.0248	0.1291±0.0317	0.0714±0.0517	0.0189±0.0193	0.9174	Yes
	DIR	0.6644±0.0098	0.1150±0.0091	0.1155±0.0239	0.0952±0.0359	0.0747±0.0370	0.0387±0.0640	0.0650	Yes
	FairSHAP	0.6609±0.0106	0.0629±0.0091	0.1326±0.0407	0.0985±0.0603	**0.0452±0.0383**	**0.0025±0.0048**	**0.0113**	**No**
COMPAS (race)	Baseline	**0.6689±0.0108**	0.0995±0.0076	0.1436±0.0209	0.1438±0.0233	0.0522±0.0406	—	—	No
	RSA	0.6623±0.0078	—	0.2069±0.0128	0.2024±0.0349	0.0575±0.0291	—	—	Yes
	CR	0.6611±0.0112	**0.0418±0.0092**	0.1502±0.0341	0.1621±0.0530	0.0592±0.0367	0.0250±0.0222	0.8920	Yes
	DIR	0.6149±0.0286	0.1185±0.0181	0.1359±0.1241	**0.1117±0.0945**	**0.0399±0.0338**	0.0512±0.0736	0.0701	Yes
	FairSHAP	0.6627±0.0069	0.0842±0.0049	**0.1344±0.0332**	0.1568±0.0343	0.0508±0.0469	**0.0040±0.0055**	**0.0126**	**No**
Adult (sex)	Baseline	**0.8722±0.0033**	0.0315±0.0037	0.1805±0.0066	0.0735±0.0275	0.0275±0.0321	—	—	No
	RSA	0.8721±0.0024	—	0.1770±0.0066	0.0678±0.0277	**0.0272±0.0252**	—	—	Yes
	CR	0.8706±0.0029	**0.0000±0.0000**	0.1824±0.0055	0.0955±0.0243	0.0278±0.0173	0.0167±0.0391	0.9887	Yes
	DIR	0.8550±0.0067	0.0499±0.0076	0.1607±0.0157	0.0772±0.0624	0.0360±0.0253	0.0046±0.0417	0.0081	Yes
	FairSHAP	0.8692±0.0046	0.0273±0.0047	**0.1558±0.0130**	**0.0393±0.0254**	0.0474±0.0319	**0.0010±0.0073**	**0.0012**	**No**
Adult (race)	Baseline	**0.8721±0.0033**	0.0398±0.0025	0.1034±0.0110	0.0808±0.0326	0.0302±0.0265	—	—	No
	RSA	0.8710±0.0044	—	0.0971±0.0043	0.0864±0.0347	0.0427±0.0142	—	—	Yes
	CR	0.8713±0.0033	**0.0000±0.0000**	0.1008±0.0115	0.0983±0.0389	0.0480±0.0235	0.0300±0.0450	0.9620	Yes
	DIR	0.8320±0.0173	0.0740±0.0209	**0.0703±0.0515**	0.0871±0.0355	0.0482±0.0730	0.0252±0.0480	0.0089	Yes
	FairSHAP	0.8720±0.0023	0.0284±0.0017	0.0851±0.0155	**0.0287±0.0277**	**0.0259±0.0318**	**0.0030±0.0084**	**0.0014**	**No**
Kdd (sex)	Baseline	0.9377±0.0026	0.0767±0.0011	0.0012±0.0008	0.0012±0.0008	0.0796±0.0054	—	—	No
	RSA	0.9380±0.0026	—	**0.0008±0.0004**	0.0008±0.0004	0.0796±0.0055	—	—	Yes
	CR	0.9377±0.0025	**0.0000±0.0000**	0.0013±0.0004	0.0013±0.0003	0.0796±0.0054	0.0007±0.0013	0.9790	Yes
	DIR	0.9377±0.0028	0.0766±0.0014	0.0012±0.0010	0.0013±0.0011	0.0797±0.0055	0.0001±0.0007	0.0013	Yes
	FairSHAP	**0.9381±0.0029**	0.0732±0.0020	0.0008±0.0007	**0.0006±0.0008**	**0.0794±0.0060**	**0.0000±0.0000**	**0.0003**	**No**
DefaultCredit (sex)	Baseline	0.8141±0.0059	0.0226±0.0019	0.0335±0.0072	0.0348±0.0211	0.0269±0.0127	—	—	No
	RSA	0.8144±0.0059	—	**0.0259±0.0063**	0.0244±0.0122	**0.0154±0.0172**	—	—	Yes
	CR	0.8144±0.0068	**0.0003±0.0002**	0.0339±0.0077	0.0419±0.0203	0.0363±0.0203	0.0051±0.0064	0.9844	Yes
	DIR	0.8138±0.0051	0.0224±0.0027	0.0308±0.0091	0.0334±0.0244	0.0255±0.0214	0.0023±0.0056	0.0741	Yes
	FairSHAP	**0.8145±0.0050**	0.0214±0.0026	0.0306±0.0054	**0.0216±0.0164**	0.0289±0.0075	**0.0001±0.0003**	**0.0004**	**No**

Overall setup We used the aforementioned datasets, where the sensitive attribute (s.a.) was preprocessed according to its type. When the s.a. is *sex*, it is treated as binary (since the sex attribute in all datasets is binary), and we subsequently refer to it simply as s.a. When the s.a. is *race*, it naturally involves multiple classes, corresponding to the multi-class case. In all scenarios, for the purpose of fairness analysis, we further group the data into a privileged group and an unprivileged group. In addition, categorical features were one-hot encoded and numerical features were standardized. For model training, we employed the XGBoost Classifier with default hyperparameter settings throughout the project [Chen and Guestrin, 2016]. All experiments were conducted using 5-fold cross-validation and conducted on computer equipped with i5-13500H CPU.

4.1 Main results

The impact of FairSHAP on DR is shown in Figure 2. Each curve represents the average over 5 folds, with shaded areas indicating standard deviation. The dashed horizontal line at $y = 0$ corresponds to the original DR of the unmodified model. This figure compares the relative DR reduction of a model retrained on a FairSHAP-modified training set and evaluated on the test set, against the original model evaluated on the same test set. Higher percentages indicate a greater reduction in DR compared to the original model, and thus reflect improved fairness.

We set the threshold $T = 0.05$, meaning that we modify the feature values corresponding to Shapley values greater than 0.05, and the number of modifications under this threshold varies across datasets (see Appendix G for a detailed justification of this choice). As the number of modifications increases, we observe a consistent decrease in DR across all datasets. Under the threshold of $T = 0.05$, FairSHAP is able to reduce DR by up to 80% for the German Credit dataset and by nearly 40% for the COMPAS dataset. For the other two datasets, FairSHAP reduces the DR by around 20%. However, for the Census Income dataset, which contains

Table 2: Compare FairSHAP with ablation studies across different datasets. The sensitive attribute is sex in all cases. Best and second-best results for fairness measures (DR, DP, EO, PQP) are **bold**.

Dataset	Methods	Accuracy	DR	DP	EO	PQP
German	Baseline	0.6650±0.0257	0.0785±0.0211	0.0512±0.0346	0.1287±0.0590	**0.1341±0.0486**
	Ablation study 1	**0.6690±0.0237**	0.0709±0.0239	0.0446±0.0106	0.0972±0.1091	0.1463±0.1133
	Ablation study 2	0.6470±0.0452	0.0640±0.0176	0.0708±0.0510	0.1619±0.0730	0.1475±0.1384
	FairSHAP	0.6630±0.0275	**0.0243±0.0112**	**0.0301±0.0347**	**0.1126±0.0783**	0.1852±0.1074
COMPAS	Baseline	0.6698±0.0051	0.0883±0.0064	0.1548±0.0241	0.1243±0.0510	0.0492±0.0084
	Ablation study 1	**0.6713±0.0090**	0.0903±0.0110	0.1519±0.0114	0.1315±0.0379	0.0490±0.0465
	Ablation study 2	0.6699±0.0076	0.0721±0.0059	0.1537±0.0177	0.1125±0.0431	**0.0318±0.0121**
	FairSHAP	0.6609±0.0106	**0.0629±0.0091**	**0.1326±0.0407**	**0.0985±0.0603**	0.0452±0.0383
Adult	Baseline	0.8722±0.0033	0.0315±0.0037	0.1805±0.0066	0.0735±0.0275	**0.0275±0.0321**
	Ablation study 1	0.8717±0.0031	0.0332±0.0020	0.1842±0.0048	0.0932±0.0186	0.0344±0.0228
	Ablation study 2	**0.8722±0.0016**	0.0278±0.0018	0.1816±0.0084	0.0761±0.0325	0.0442±0.0266
	FairSHAP	0.8692±0.0046	**0.0273±0.0047**	**0.1558±0.0130**	**0.0393±0.0254**	0.0474±0.0319
CensusIncome	Baseline	0.9377±0.0026	0.0767±0.0011	0.0012±0.0008	0.0012±0.0008	0.0796±0.0054
	Ablation study 1	0.9377±0.0027	0.0765±0.0019	0.0010±0.0006	0.0011±0.0007	0.0797±0.0055
	Ablation study 2	0.9378±0.0024	0.0744±0.0018	0.0010±0.0003	0.0010±0.0003	0.0796±0.0056
	FairSHAP	**0.9381±0.0029**	**0.0732±0.0020**	**0.0008±0.0007**	**0.0006±0.0008**	**0.0794±0.0060**
DefaultCredit	Baseline	0.8141±0.0059	0.0226±0.0019	0.0335±0.0072	0.0348±0.0211	**0.0269±0.0127**
	Ablation study 1	0.8149±0.0061	0.0221±0.0027	0.0327±0.0067	0.0431±0.0111	0.0369±0.0166
	Ablation study 2	**0.8150±0.0051**	0.0217±0.0031	0.0331±0.0070	0.0259±0.0135	0.0325±0.0137
	FairSHAP	0.8145±0.0050	**0.0214±0.0026**	**0.0306±0.0054**	**0.0216±0.0164**	0.0289±0.0075

a large number of features, the DR reduction is only about 5%. This reduction in DR is also accompanied by corresponding decreases in DP and EO, indicating that FairSHAP effectively enhances overall model fairness.

We also evaluated the performance of FairSHAP-modified data on other group fairness measures; detailed results can be found in Appendix A. Moreover, in datasets such as COMPAS, Census Income KDD, and Default Credit, the accuracy improves after applying modifications, demonstrating that FairSHAP can simultaneously enhance both fairness and predictive performance. In the remaining datasets, the accuracy experiences only a marginal decrease, indicating a balanced trade-off between fairness and model efficacy.

Furthermore, we conducted comparative experiments with other benchmark methods on the same datasets. The results, presented in Table 1, show that FairSHAP improves all three fairness metrics (DR, DP, EO) with minimal modifications. In some datasets, it reduces all fairness measures simultaneously. These improvements generally come with only a slight decrease in model accuracy, and in three of the datasets, they even lead to an accuracy increase. This comparative analysis confirms the reliability and effectiveness of FairSHAP in enhancing model fairness.

Although CorrelationRemover performs well in terms of DR, its mechanism relies on projection to identify the influence of the sensitive attribute on the non-sensitive attribute, subsequently adjusting the non-sensitive attribute to achieve a low DR value. However, this approach has several drawbacks: *(1)* It modifies a large proportion of the data, exceeding 95%, which significantly compromises data integrity. *(2)* Its performance on DP and EO metrics is substantially worse than that of FairSHAP. *(3)* It requires applying the same projection matrix to modify the test set.

What's more, while DisparateImpactRemover can reduce certain values in terms of DP and EO, *(1)* the extent of improvement across almost all metrics is generally inferior to that achieved by FairSHAP. *(2)* It requires more modifications to the dataset compared to FairSHAP. *(3)* It also necessitates applying the same transformations to the test set, which is not required by FairSHAP.

4.2 Ablation study

To evaluate the importance of each component in our method, we conducted below ablation studies:

- **Ablation study 1:** Instead of using the aforementioned $\mathcal{G}$ and $\tilde{\mathcal{G}}$, randomly sample a single data point from dataset $\mathcal{D}$ and replace its value in a specific column with a randomly selected value from the same column in other rows. The number of modifications performed should be approximately equal to that of FairSHAP.
- **Ablation study 2:** Using $\mathcal{G}$ and $\tilde{\mathcal{G}}$, match them via nearest neighbor, but instead of employing Shapley values to compute feature importance, directly replace values in $\mathcal{G}$ by randomly sampling from its reference data and replace values in $\tilde{\mathcal{G}}$ by randomly sampling from its corresponding reference data. The number of modifications performed should be approximately equal to that of FairSHAP.

The results, summarized in Table 2, show that the metrics in Ablation study 1 fluctuate slightly around the baseline, with overall performance slightly worse than the baseline. This deviation is due to the random modification of a small number of data points. In contrast, the metrics in Ablation study 2 show a slight improvement over the baseline, confirming that using matching and modifying values based on background data has a positive effect. This also highlights the importance of matching in FairSHAP. Finally, FairSHAP achieves overall improvements in DR, DP, and EO, with some datasets even showing increased accuracy. These two ablation studies demonstrate that FairSHAP, based on matching and Shapley values, is effective for fairness-aware data augmentation.

However, when dealing with large datasets, the gap between FairSHAP and the Ablation study becomes smaller. We believe this is because one-hot encoding, used during preprocessing, generates a large number of features. As SHAP's kernel method relies on a random baseline to reduce computational complexity, identifying important features and making adjustments becomes more challenging.

5 CONCLUSION

In this paper, we introduced FairSHAP, an innovative preprocessing framework that leverages Shapley value attribution to enhance fairness in ML models. FairSHAP effectively identifies and systematically adjusts fairness-critical instances using interpretable feature importance measures and instance-level matching across sensitive groups. Our approach significantly reduces fairness metrics such as DR, DP and EO across various datasets. Crucially, FairSHAP achieves these improvements while perturbing the data as minimally as possible, largely preserving data fidelity and, in certain cases, even enhancing predictive accuracy. As a transparent and model-agnostic method, FairSHAP integrates into existing machine learning pipelines, providing actionable insights into bias sources and fostering accountable and equitable AI systems.

Future work In our future work, we target: *(1)* Finding a new matching method relationship in addition to nearest-neighbor matching; *(2)* Accelerating FairSHAP by integrating more efficient Shapley value estimation methods, in order to overcome computational complexity limitations (see Appendix F and Appendix I).

References

Alekh Agarwal, Alina Beygelzimer, Miroslav Dudík, John Langford, and Hanna Wallach. A reductions approach to fair classification. In *International conference on machine learning*, pages 60–69. PMLR, 2018.

Ulrich Aïvodji, Hiromi Arai, Olivier Fortineau, Sébastien Gambs, Satoshi Hara, and Alain Tapp. Fairwashing: the risk of rationalization. In *International Conference on Machine Learning*, pages 161–170. PMLR, 2019.

Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. How we analyzed the compas recidivism algorithm. https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm, 2016.

Adrian Arnaiz-Rodriguez and Nuria Oliver. Towards algorithmic fairness by means of instance-level data re-weighting based on shapley values. *arXiv preprint arXiv:2303.01928*, 2023.

Osbert Bastani, Xin Zhang, and Armando Solar-Lezama. Probabilistic verification of fairness properties via concentration. *Proceedings of the ACM on Programming Languages*, 3(OOPSLA):1–27, 2019.

Barry Becker and Ronny Kohavi. Adult. UCI Machine Learning Repository, 1996.

Yijun Bian and Kun Zhang. Increasing fairness via combination with learning guarantees. *arXiv preprint arXiv:2301.10813*, 2023.

Reuben Binns. Fairness in machine learning: Lessons from political philosophy. In *Conference on fairness, accountability and transparency*, pages 149–159. PMLR, 2018.

Reuben Binns. On the apparent conflict between individual and group fairness. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pages 514–524, 2020.

Toon Calders and Indrė Žliobaitė. Why unbiased computational processes can lead to discriminative decision procedures. In *Discrimination and Privacy in the Information Society: Data mining and profiling in large databases*, pages 43–57. Springer, 2013.

Flavio Calmon, Dennis Wei, Bhanukiran Vinzamuri, Karthikeyan Natesan Ramamurthy, and Kush R Varshney. Optimized pre-processing for discrimination prevention. *Advances in neural information processing systems*, 30, 2017.

Simon Caton and Christian Haas. Fairness in machine learning: A survey. *ACM Comput. Surv.*, 56(7), April 2024. ISSN 0360-0300.

Shuyi Chen and Shixiang Zhu. Counterfactual fairness through transforming data orthogonal to bias. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 2*, pages 239–249, 2025.

Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pages 785–794, 2016.

Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Machine learning*, 20(3):273–297, 1995.

Jeffrey Dastin. Amazon scraps secret ai recruiting tool that showed bias against women. *Reuters*, 2018. URL https://www.reuters.com/article/us-amazon-com-jobs-automation-insight-idUSKCN1MK08G. Accessed: 2025-03-21.

Michele Donini, Luca Oneto, Shai Ben-David, John S Shawe-Taylor, and Massimiliano Pontil. Empirical risk minimization under fairness constraints. *Advances in neural information processing systems*, 31, 2018.

Finale Doshi-Velez and Been Kim. Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608*, 2017.

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

Cynthia Dwork, Nicole Immorlica, Adam Tauman Kalai, and Max Leiserson. Decoupled classifiers for group-fair and efficient machine learning. In *Conference on fairness, accountability and transparency*, pages 119–133. PMLR, 2018.

Michael Feldman, Sorelle A Friedler, John Moeller, Carlos Scheidegger, and Suresh Venkatasubramanian. Certifying and removing disparate impact. In *proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining*, pages 259–268, 2015.

Vincent Grari, Sylvain Lamprier, and Marcin Detyniecki. Adversarial learning for counterfactual fairness. *Machine Learning*, 112(3):741–763, 2023.

Moritz Hardt, Eric Price, and Nati Srebro. Equality of opportunity in supervised learning. *Advances in neural information processing systems*, 29, 2016.

Noah Hollmann, Samuel Müller, Lennart Purucker, Arjun Krishnakumar, Max Körfer, Shi Bin Hoo, Robin Tibor Schirrmeister, and Frank Hutter. Accurate predictions on small data with a tabular foundation model. *Nature*, 01 2025.

Max Hort, Zhenpeng Chen, Jie M Zhang, Mark Harman, and Federica Sarro. Bias mitigation for machine learning classifiers: A comprehensive survey. *ACM Journal on Responsible Computing*, 1(2):1–52, 2024.

Faisal Kamiran and Toon Calders. Data preprocessing techniques for classification without discrimination. *Knowledge and information systems*, 33(1):1–33, 2012.

Toshihiro Kamishima, Shotaro Akaho, Hideki Asoh, and Jun Sakuma. Fairness-aware classifier with prejudice remover regularizer. In *Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2012, Bristol, UK, September 24-28, 2012. Proceedings, Part II 23*, pages 35–50. Springer, 2012.

Michael Kearns, Seth Neel, Aaron Roth, and Zhiwei Steven Wu. Preventing fairness gerrymandering: Auditing and learning for subgroup fairness. In *International conference on machine learning*, pages 2564–2572. PMLR, 2018.

Jon Kleinberg. Inherent trade-offs in algorithmic fairness. In *Abstracts of the 2018 ACM International Conference on Measurement and Modeling of Computer Systems*, pages 40–40, 2018.

Matt J Kusner, Joshua Loftus, Chris Russell, and Ricardo Silva. Counterfactual fairness. *Advances in neural information processing systems*, 30, 2017.

Himabindu Lakkaraju, Ece Kamar, Rich Caruana, and Eric Horvitz. Identifying unknown unknowns in the open world: Representations and policies for guided exploration. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 31, 2017.

Carol Long, Hsiang Hsu, Wael Alghamdi, and Flavio Calmon. Individual arbitrariness and group fairness. *Advances in Neural Information Processing Systems*, 36:68602–68624, 2023.

Christos Louizos, Kevin Swersky, Yujia Li, Max Welling, and Richard Zemel. The variational fair autoencoder. 2015.

Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30, 2017.

David Madras, Elliot Creager, Toniann Pitassi, and Richard Zemel. Learning adversarially fair and transferable representations. In *International Conference on Machine Learning*, pages 3384–3393. PMLR, 2018.

Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. A survey on bias and fairness in machine learning. *ACM computing surveys (CSUR)*, 54(6):1–35, 2021.

Luke Merrick and Ankur Taly. The explanation game: Explaining machine learning models using shapley values. In *Machine Learning and Knowledge Extraction: 4th IFIP TC 5, TC 12, WG 8.4, WG 8.9, WG 12.9 International Cross-Domain Conference, CD-MAKE 2020, Dublin, Ireland, August 25–28, 2020, Proceedings 4*, pages 17–38. Springer, 2020.

Felix Petersen, Debarghya Mukherjee, Yuekai Sun, and Mikhail Yurochkin. Post-processing for individual fairness. *Advances in Neural Information Processing Systems*, 34:25944–25955, 2021.

Forough Poursabzi-Sangdeh, Daniel G Goldstein, Jake M Hofman, Jennifer Wortman Wortman Vaughan, and Hanna Wallach. Manipulating and measuring model interpretability. In *Proceedings of the 2021 CHI conference on human factors in computing systems*, pages 1–52, 2021.

ProPublica. propublica/compas-analysis. https://github.com/propublica/compas-analysis. Accessed: 23 March 2025.

Khushnaseeb Roshan and Aasim Zafar. Using kernel shap xai method to optimize the network anomaly detection model. In *2022 9th International Conference on Computing for Sustainable Global Development (INDIACom)*, pages 74–80. IEEE, 2022.

Shubhanshu Shekhar, Greg Fields, Mohammad Ghavamzadeh, and Tara Javidi. Adaptive sampling for minimax fair classification. *Advances in Neural Information Processing Systems*, 34:24535–24544, 2021.

Erik Strumbelj and Igor Kononenko. An efficient explanation of individual classifications using game theory. *The Journal of Machine Learning Research*, 11:1–18, 2010.

Mukund Sundararajan and Amir Najmi. The many shapley values for model explanation. In *International conference on machine learning*, pages 9269–9278. PMLR, 2020.

UCI Machine Learning Repository. Census-Income (KDD). https://archive.ics.uci.edu/dataset/117/census-income-kdd, 2000.

UCI Machine Learning Repository. South german credit. UCI Machine Learning Repository, 2020.

Berk Ustun, Yang Liu, and David Parkes. Fairness without harm: Decoupled classifiers with preference guarantees. In *International Conference on Machine Learning*, pages 6373–6382. PMLR, 2019.

Mingyang Wan, Daochen Zha, Ninghao Liu, and Na Zou. In-processing modeling techniques for machine learning fairness: A survey. *ACM Transactions on Knowledge Discovery from Data*, 17(3):1–27, 2023.

Hao Wang, Berk Ustun, and Flavio Calmon. Repairing without retraining: Avoiding disparate impact with counterfactual distributions. In *International Conference on Machine Learning*, pages 6618–6627. PMLR, 2019.

Hilde Weerts, Miroslav Dudík, Richard Edgar, Adrin Jalali, Roman Lutz, and Michael Madaio. Fairlearn: Assessing and Improving Fairness of AI Systems. *Journal of Machine Learning Research*, 24, 2023.

I-Cheng Yeh. Default of Credit Card Clients. UCI Machine Learning Repository, 2009.

Lei You, Yijun Bian, and Lele Cao. Refining counterfactual explanations with joint-distribution-informed shapley towards actionable minimality. *arXiv preprint arXiv:2410.05419*, 2024.

Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rogriguez, and Krishna P Gummadi. Fairness constraints: Mechanisms for fair classification. In *Artificial intelligence and statistics*, pages 962–970. PMLR, 2017.

Rich Zemel, Yu Wu, Kevin Swersky, Toni Pitassi, and Cynthia Dwork. Learning fair representations. In Sanjoy Dasgupta and David McAllester, editors, *Proceedings of the 30th International Conference on Machine Learning*, volume 28 of *Proceedings of Machine Learning Research*, pages 325–333, Atlanta, Georgia, USA, Jun 2013. PMLR.

Brian Hu Zhang, Blake Lemoine, and Margaret Mitchell. Mitigating unwanted biases with adversarial learning. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, pages 335–340, 2018.

Tianxiang Zhao, Enyan Dai, Kai Shu, and Suhang Wang. Towards fair classifiers without sensitive attributes: Exploring biases in related features. In *Proceedings of the Fifteenth ACM International Conference on Web Search and Data Mining*, pages 1433–1442, 2022.

Checklist

1. For all models and algorithms presented, check if you include:
   (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model. [Yes]
   (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. [Yes]

(c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries. [Yes]

2. For any theoretical claim, check if you include:
   (a) Statements of the full set of assumptions of all theoretical results. [Yes]
   (b) Complete proofs of all theoretical results. [Yes]
   (c) Clear explanations of any assumptions. [Yes]

3. For all figures and tables that present empirical results, check if you include:
   (a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL). [Yes]
   (b) All the training details (e.g., data splits, hyperparameters, how they were chosen). [Yes]
   (c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times). [Yes]
   (d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). [Yes]

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:
   (a) Citations of the creator If your work uses existing assets. [Yes]
   (b) The license information of the assets, if applicable. [Yes]
   (c) New assets either in the supplemental material or as a URL, if applicable. [Yes]
   (d) Information about consent from data providers/curators. [Not Applicable]
   (e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. [Not Applicable]

5. If you used crowdsourcing or conducted research with human subjects, check if you include:
   (a) The full text of instructions given to participants and screenshots. [Not Applicable]
   (b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. [Not Applicable]
   (c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. [Not Applicable]

Appendices

A Fairness metrics across datasets

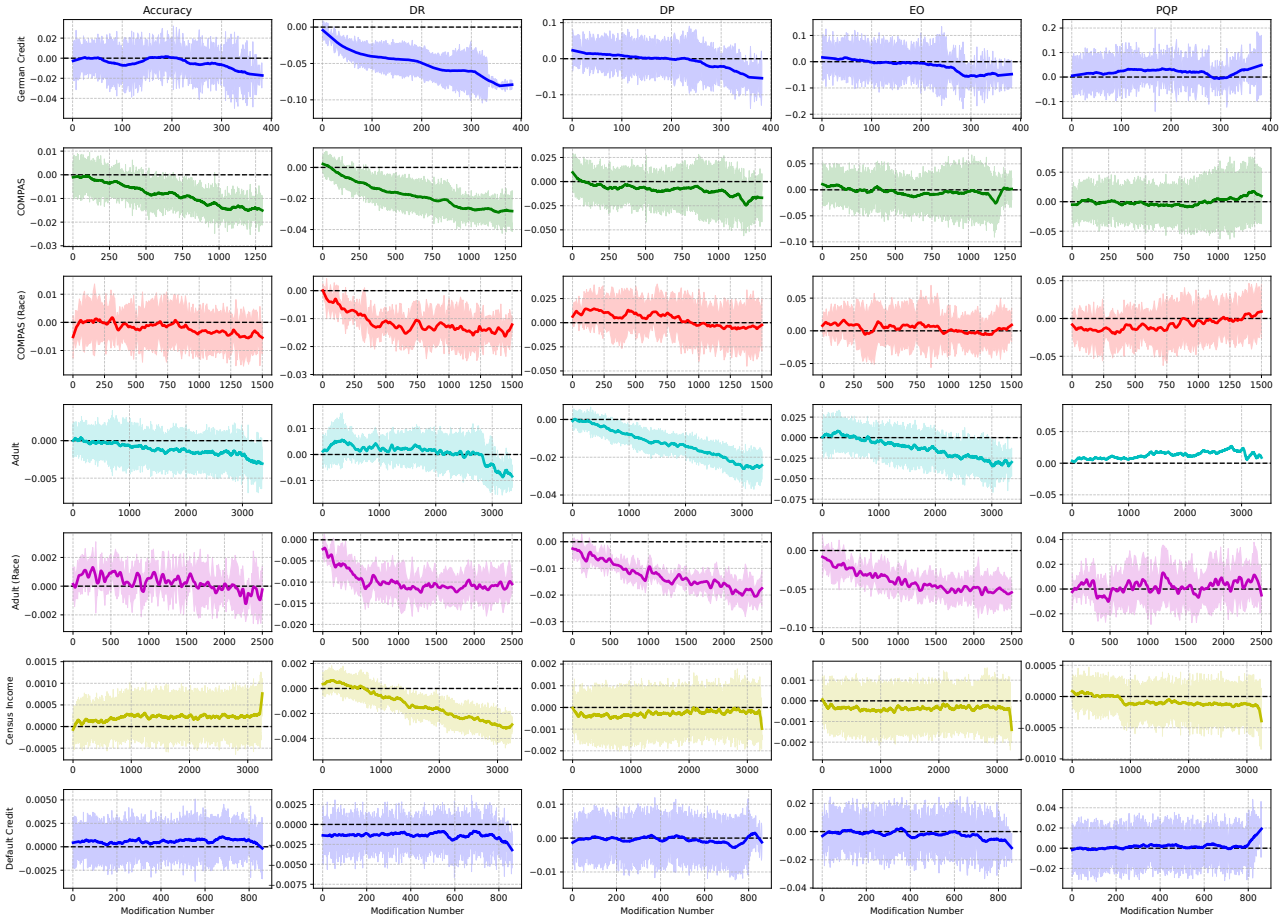

Figure 3: Improvement in fairness metrics (Accuracy, DR, DP, EO, and PQP) as the number of modifications increases using FairSHAP on various datasets. Solid lines represent the mean, and shaded areas indicate the standard deviation from 5-fold cross-validation. The x-axis shows the number of modifications taken, up to the maximum required under a given threshold T, where $T = 0.05$.

B Definition of group fairness measures

Notation and setup. Let $\mathcal{X}$ denote the feature space, and each instance be written as $\mathbf{x} = (\check{\mathbf{x}}, A)$, where $\check{\mathbf{x}} \in \mathcal{X}_{\mathcal{F}\setminus\{A\}}$ are the non-sensitive features and $A \in \{0, 1\}$ is the sensitive attribute (0 and 1 representing the unprivileged and privileged groups, respectively). We denote by $\mathbb{P}_{\mathcal{D}}(\cdot)$ the probability with respect to the underlying data distribution $\mathcal{D}$ over $(\mathcal{X}, \mathcal{Y})$.

Demographic parity (DP). DP requires that the probability of a positive prediction be independent of the sensitive attribute:

$$L_{\mathrm{DP}}(f) = \left|\mathbb{P}_{\mathcal{D}}\big(f(\mathbf{x}) = 1 \mid A = 0\big) - \mathbb{P}_{\mathcal{D}}\big(f(\mathbf{x}) = 1 \mid A = 1\big)\right| \leq \varepsilon\,. \tag{7}$$

A model violating this condition exhibits *disparate impact*, since the chance of receiving a favorable outcome differs between groups.

Equality of opportunity (EO). EO requires that, among truly qualified individuals ($y = 1$), the chance of a positive prediction does not depend on the sensitive attribute:

$$L_{\mathrm{EO}}(f) = \left|\mathbb{P}_{\mathcal{D}}\big(f(\mathbf{x}) = 1 \mid A = 0, y = 1\big) - \mathbb{P}_{\mathcal{D}}\big(f(\mathbf{x}) = 1 \mid A = 1, y = 1\big)\right| \leq \varepsilon\,. \tag{8}$$

Failure to satisfy EO indicates *disparate treatment*, disadvantaging some groups despite equal qualifications.

Predictive quality parity (PQP). PQP requires that the reliability of positive predictions be consistent across groups:

$$L_{\mathrm{PQP}}(f) = \left|\mathbb{P}_{\mathcal{D}}\big(y=1 \mid f(\mathbf{x})=1, A=0\big) - \mathbb{P}_{\mathcal{D}}\big(y=1 \mid f(\mathbf{x})=1, A=1\big)\right| \le \varepsilon\,. \tag{9}$$

Violation of PQP implies that the correctness of positive predictions differs across groups, which is undesirable in high-stakes domains such as healthcare or finance.

C DR Reduction Achieved by FairSHAP

We now establish the theoretical foundation of FairSHAP by analyzing its impact on the individual fairness measure, DR. Our analysis begins by showing that L_{DR} admits a natural decomposition into feature-wise Shapley contributions, which allows us to identify fairness-critical features. Building on this, we derive a provable lower bound on the reduction in L_{DR} that can be achieved by selectively modifying such features.

Notation and setup. Let $\mathcal{X}$ denote the feature space, $\mathcal{F}$ denotes full set of features, and each instance be written as $\mathbf{x} = (\check{\mathbf{x}}, A)$, where $\check{\mathbf{x}} \in \mathcal{X}_{\mathcal{F}\setminus\{A\}}$ are the non-sensitive features and $A \in \{0,1\}$ is the sensitive attribute (0 and 1 representing the unprivileged and privileged groups, respectively). We denote by $\mathbb{P}_{\mathcal{D}}(\cdot)$ the probability with respect to the underlying data distribution $\mathcal{D}$ over $(\mathcal{X}, \mathcal{Y})$.

Lemma C.1 (Shapley decomposition of individual DR)**.** *Let $\mathbf{x} = (\check{\mathbf{x}}, A)$ be an instance, where $\check{\mathbf{x}} \in \mathcal{X}_{\mathcal{F}\setminus\{A\}}$ denotes the non-sensitive features and $A \in \{0,1\}$ is the sensitive attribute.*

The individual discriminative risk is defined as

$$L_{\mathrm{DR}}(f, \mathbf{x}) = \big|f(\check{\mathbf{x}}, A=0) - f(\check{\mathbf{x}}, A=1)\big|\,.$$

Fix a reference instance $\mathbf{r}$ drawn from the reference set $\mathcal{B}$ and consider the coalition game

$$v_{\mathrm{DR}}^{(\mathbf{x})}(S) = L_{\mathrm{DR}}\big(\mathbf{x}_S;\, \mathbf{r}_{\mathcal{F}\setminus S}\big)\,, \qquad S \subseteq [d]\,.$$

Let $\phi_k(\mathbf{x})$ denote the Shapley value of feature k in this game and define the baseline term as

$$\phi_0(\mathbf{x}) := v_{\mathrm{DR}}^{(\mathbf{x})}(\emptyset)\,.$$

Then, for the grand coalition $S = [d]$, we have

$$L_{\mathrm{DR}}(f, \mathbf{x}) = \phi_0(\mathbf{x}) + \sum_{k=1}^{d} \phi_k(\mathbf{x})\,.$$

Proof. We recall that the Shapley value ϕ_j for feature j satisfies the *efficiency* axiom, i.e.,

$$\textstyle\sum_{j=1}^{d} \phi_j(\mathbf{x}) = v_{\mathrm{DR}}^{(\mathbf{x})}([d]) - v_{\mathrm{DR}}^{(\mathbf{x})}(\emptyset)\,,$$

which states that the sum of the contributions of all features equals the total value of the grand coalition relative to the empty coalition.

In our setting, for the grand coalition $S = [d]$, we have

$$v_{\mathrm{DR}}^{(\mathbf{x})}([d]) = L_{\mathrm{DR}}(f, \mathbf{x})\,,$$

and for the empty coalition $S = \emptyset$, all features are taken from the reference instance, hence

$$v_{\mathrm{DR}}^{(\mathbf{x})}(\emptyset) = L_{\mathrm{DR}}(f, \mathbf{r})\,.$$

Rearranging the efficiency identity and using the definition $\phi_0(\mathbf{x}) := v_{\mathrm{DR}}^{(\mathbf{x})}(\emptyset)$ yields

$$L_{\mathrm{DR}}(f, \mathbf{x}) = v_{\mathrm{DR}}^{(\mathbf{x})}([d]) = v_{\mathrm{DR}}^{(\mathbf{x})}(\emptyset) + \sum_{j=1}^{d} \phi_j(\mathbf{x}) = \phi_0(\mathbf{x}) + \sum_{j=1}^{d} \phi_j(\mathbf{x})\,.$$

□

Lemma C.1 provides a rigorous way to decompose L_{DR} into additive contributions of individual features. Intuitively, features with large Shapley values exert disproportionate influence on model predictions, and are thus most likely

to drive individual-level unfairness. We therefore define these as fairness-critical features. By neutralizing their effect through FairSHAP, we can guarantee a measurable reduction in L_{DR}.

Theorem C.2 (Instance-wise DR reduction achieved by FairSHAP)**.** *Fix a threshold $T > 0$ and define the set of fairness-critical features as*

$$S_T(\mathbf{x}) = \{\, k \in [d] \;\mid\; \phi_k(\mathbf{x}) \geq T \,\}\,.$$

Let $\mathbf{x}_{\mathrm{after}}$ denote the feature vector obtained from $\mathbf{x}$ after FairSHAP replaces every feature $k \in S_T(\mathbf{x})$ by its value in the matched reference $\mathbf{r}$. Define the interaction residual term as

$$\mathcal{I}(\mathbf{x}) \;:=\; L_{\mathrm{DR}}(f, \mathbf{x}_{\mathrm{after}}) - \left(L_{\mathrm{DR}}(f, \mathbf{x}_{\mathrm{before}}) - \sum_{k \in S_T(\mathbf{x})} \phi_k(\mathbf{x}) \right).$$

Then

$$L_{\mathrm{DR}}(f, \mathbf{x}_{\mathrm{after}}) = L_{\mathrm{DR}}(f, \mathbf{x}_{\mathrm{before}}) - \sum_{k \in S_T(\mathbf{x})} \phi_k(\mathbf{x}) + \mathcal{I}(\mathbf{x}) \;\;\leq\;\; L_{\mathrm{DR}}(f, \mathbf{x}_{\mathrm{before}}) - T\,|S_T(\mathbf{x})| + \mathcal{I}(\mathbf{x})\,,$$

where $\mathcal{I}(\mathbf{x})$ denotes the interaction residual term. In particular, if $\mathcal{I}(\mathbf{x}) \leq T|S_T(\mathbf{x})|$, then $L_{\mathrm{DR}}(f, \mathbf{x}_{\mathrm{after}}) \leq L_{\mathrm{DR}}(f, \mathbf{x}_{\mathrm{before}})$, i.e., the DR does not increase.

Proof. By Lemma C.1,

$$L_{\mathrm{DR}}(f, \mathbf{x}_{\mathrm{before}}) = \phi_0(\mathbf{x}) + \sum_{k \in S_T(\mathbf{x})} \phi_k(\mathbf{x}) + \sum_{k \notin S_T(\mathbf{x})} \phi_k(\mathbf{x})\,.$$

By the definition of $\mathcal{I}(\mathbf{x})$ above (which accounts for non-additive interactions when replacing the subset $S_T(\mathbf{x})$),

$$L_{\mathrm{DR}}(f, \mathbf{x}_{\mathrm{after}}) = L_{\mathrm{DR}}(f, \mathbf{x}_{\mathrm{before}}) - \sum_{k \in S_T(\mathbf{x})} \phi_k(\mathbf{x}) + \mathcal{I}(\mathbf{x})\,.$$

Since $\phi_k(x) \geq T$ for all $k \in S_T(x)$, by the definition of $S_T(x)$ we have

$$\sum_{k \in S_T(x)} \phi_k(x) \;\geq\; T\,|S_T(x)|\,.$$

Substituting completes the proof. □

Theorem C.2 provides a per-instance bound on the post-modification DR *up to the residual* $\mathcal{I}(\mathbf{x})$. Therefore, a strict improvement guarantee requires controlling $\mathcal{I}(\mathbf{x})$, e.g., via a separate bound in Proposition C.3. More precisely, the theorem implies

$$L_{\mathrm{DR}}(f, \mathbf{x}_{\mathrm{after}}) \;\leq\; L_{\mathrm{DR}}(f, \mathbf{x}_{\mathrm{before}}) \;-\; T|S_T(\mathbf{x})| \;+\; \mathcal{I}(\mathbf{x})\,,$$

so a *non-increase* (and hence an improvement) is guaranteed whenever $\mathcal{I}(\mathbf{x}) \leq T|S_T(\mathbf{x})|$, in which case the reduction is at least $T|S_T(\mathbf{x})| - \mathcal{I}(\mathbf{x})$. Accordingly, we refrain from claiming a uniform per-feature reduction of at least T without an explicit bound on $\mathcal{I}(\mathbf{x})$. This establishes a principled lower bound on fairness improvement (assuming $|\mathcal{I}(\mathbf{x})|$ is small via local matching): the more fairness-critical features are neutralized, the greater the guaranteed reduction in individual-level unfairness.

Proposition C.3 (Bound on Interaction Residual)**.** *Let $L(\mathbf{x}) := L_{\mathrm{DR}}(f, \mathbf{x})$ denote the instance-wise DR objective used in Theorem C.2. Since the absolute-value DR may be non-differentiable when $f(\check{\mathbf{x}}, 0) = f(\check{\mathbf{x}}, 1)$, we assume either (i) a smooth surrogate L_δ is used (e.g., $L_\delta(\mathbf{x}) = \sqrt{(f(\check{\mathbf{x}}, 0) - f(\check{\mathbf{x}}, 1))^2 + \delta}$), or (ii) we restrict to a neighborhood where the sign of $f(\check{\mathbf{x}}, 0) - f(\check{\mathbf{x}}, 1)$ is constant, so that L is twice differentiable locally.*

Since FairSHAP uses NN or Optimal Transport to select a reference $\mathbf{r}$ from the local neighborhood of $\mathbf{x}_{\mathrm{before}}$, let $\epsilon := \|\mathbf{x}_{\mathrm{before}} - \mathbf{r}\|$ be the matching distance. Assume L is twice differentiable on the closed ball $B(\mathbf{x}_{\mathrm{before}}, \epsilon) := \{\mathbf{z} : \|\mathbf{z} - \mathbf{x}_{\mathrm{before}}\| \leq \epsilon\}$ and its Hessian is bounded there:

$$\sup_{\mathbf{z} \in B(\mathbf{x}_{\mathrm{before}}, \epsilon)} \left\| H_L(\mathbf{z}) \right\|_{\mathrm{op}} \leq M\,,$$

so that ∇L is M-Lipschitz on $B(\mathbf{x}_{\mathrm{before}}, \epsilon)$.

If $\mathbf{x}_{\mathrm{after}}$ is obtained by replacing the subset $S_T(\mathbf{x}_{\mathrm{before}})$ with reference coordinates, define $\Delta := \mathbf{x}_{\mathrm{after}} - \mathbf{x}_{\mathrm{before}}$. Then $\|\Delta\| \leq \epsilon$. Then, the interaction residual is bounded by a second-order term:

$$|\mathcal{I}(\mathbf{x}_{\mathrm{before}})| \leq \left(\frac{M}{2} + M\sqrt{|S_T(\mathbf{x}_{\mathrm{before}})|} \right) \cdot \epsilon^2\,,$$

hence in particular $|\mathcal{I}(\mathbf{x}_{\text{before}})| \le C\epsilon^2$ *for* $C := \frac{M}{2} + M\sqrt{d}$ *when the dimension* d *is fixed. Consequently, as* $\epsilon \to 0$*, the residual* $\mathcal{I}(\mathbf{x}_{\text{before}})$ *is second-order. Moreover, if* $\|\nabla L(\mathbf{x}_{\text{before}})\|_\infty$ *is locally bounded, then each first-order term scales as* $O(\epsilon)$*, so the residual vanishes faster than the leading contributions, justifying the approximation in Theorem C.2.*

Proof. Let $\Delta := \mathbf{x}_{\text{after}} - \mathbf{x}_{\text{before}}$. By the multivariate second-order Taylor theorem with Lagrange remainder, there exists $t^\star \in (0,1)$ such that

$$L(\mathbf{x}_{\text{after}}) = L(\mathbf{x}_{\text{before}}) + \nabla L(\mathbf{x}_{\text{before}})^\top \Delta + \frac{1}{2}\,\Delta^\top H_L(\mathbf{x}_{\text{before}} + t^\star\Delta)\,\Delta\,.$$

Rearranging gives

$$L(\mathbf{x}_{\text{before}}) - L(\mathbf{x}_{\text{after}}) = -\nabla L(\mathbf{x}_{\text{before}})^\top \Delta - \frac{1}{2}\,\Delta^\top H_L(\mathbf{x}_{\text{before}} + t^\star\Delta)\,\Delta\,.$$

Taking absolute values and using $\|H_L(\cdot)\|_{\text{op}} \le M$ on $B(\mathbf{x}_{\text{before}}, \epsilon)$ yields

$$\Big|\,\big(L(\mathbf{x}_{\text{before}}) - L(\mathbf{x}_{\text{after}})\big) - \big(-\nabla L(\mathbf{x}_{\text{before}})^\top \Delta\big)\,\Big| \le \frac{M}{2}\,\|\Delta\|^2\,.$$

Next we relate the linear term to the Shapley sum under the baseline-replacement game $v^{(\mathbf{x}_{\text{before}})}(U) = L\big((\mathbf{x}_{\text{before}})_U;\, \mathbf{r}_{[d]\setminus U}\big)$. Define the hybrid point $\mathbf{z}_U$ by $(\mathbf{z}_U)_i = (\mathbf{x}_{\text{before}})_i$ if $i \in U$ and $(\mathbf{z}_U)_i = r_i$ otherwise. Then

$$\|\mathbf{z}_U - \mathbf{x}_{\text{before}}\|^2 = \sum_{i \notin U} (r_i - (\mathbf{x}_{\text{before}})_i)^2 \le \sum_{i=1}^{d} (r_i - (\mathbf{x}_{\text{before}})_i)^2 = \epsilon^2\,,$$

so $\mathbf{z}_U \in B(\mathbf{x}_{\text{before}}, \epsilon)$ for all $U \subseteq [d]$. Since $B(\mathbf{x}_{\text{before}}, \epsilon)$ is convex, the segment between any two hybrid points (hence the mean-value point $\boldsymbol{\xi}_{S,k}$ below) also lies in $B(\mathbf{x}_{\text{before}}, \epsilon)$.

For any $k \in S_T(\mathbf{x}_{\text{before}})$, the Shapley value is a weighted average of marginal increments $m_{S,k} := v^{(\mathbf{x}_{\text{before}})}(S \cup \{k\}) - v^{(\mathbf{x}_{\text{before}})}(S)$ over $S \subseteq [d] \setminus \{k\}$. Fix such an S and define $g(t) := L(\mathbf{z}_S + t((\mathbf{x}_{\text{before}})_k - r_k)\mathbf{e}_k)$. By the mean value theorem, $m_{S,k} = g(1) - g(0) = \partial_k L(\boldsymbol{\xi}_{S,k})\,((\mathbf{x}_{\text{before}})_k - r_k)$ for some $\boldsymbol{\xi}_{S,k}$ on the segment between the two hybrids. Since $\boldsymbol{\xi}_{S,k} \in B(\mathbf{x}_{\text{before}}, \epsilon)$ and ∇L is M-Lipschitz on this ball, we have

$$|\partial_k L(\boldsymbol{\xi}_{S,k}) - \partial_k L(\mathbf{x}_{\text{before}})| \le \|\nabla L(\boldsymbol{\xi}_{S,k}) - \nabla L(\mathbf{x}_{\text{before}})\| \le M\|\boldsymbol{\xi}_{S,k} - \mathbf{x}_{\text{before}}\| \le M\epsilon\,.$$

Therefore,

$$|m_{S,k} - \partial_k L(\mathbf{x}_{\text{before}})\big((\mathbf{x}_{\text{before}})_k - r_k\big)| \le M\epsilon\big|(\mathbf{x}_{\text{before}})_k - r_k\big|\,.$$

Averaging over S (i.e., over the Shapley weights) yields

$$|\phi_k(\mathbf{x}_{\text{before}}) - \partial_k L(\mathbf{x}_{\text{before}})\big((\mathbf{x}_{\text{before}})_k - r_k\big)| \le M\epsilon\big|(\mathbf{x}_{\text{before}})_k - r_k\big|\,.$$

Summing over $k \in S_T(\mathbf{x}_{\text{before}})$ and using Cauchy–Schwarz gives

$$\begin{aligned}\Bigg|\sum_{k \in S_T(\mathbf{x}_{\text{before}})} \phi_k(\mathbf{x}_{\text{before}}) - \sum_{k \in S_T(\mathbf{x}_{\text{before}})} \partial_k L(\mathbf{x}_{\text{before}})\big((\mathbf{x}_{\text{before}})_k - r_k\big)\Bigg| &\le M\epsilon \sum_{k \in S_T(\mathbf{x}_{\text{before}})} \big|(\mathbf{x}_{\text{before}})_k - r_k\big| \\ &\le M\epsilon\sqrt{|S_T(\mathbf{x}_{\text{before}})|}\,\|\Delta\| \\ &\le M\sqrt{|S_T(\mathbf{x}_{\text{before}})|}\,\epsilon^2\,.\end{aligned}$$

Finally, note that $-\nabla L(\mathbf{x}_{\text{before}})^\top \Delta = \sum_{k \in S_T(\mathbf{x}_{\text{before}})} \partial_k L(\mathbf{x}_{\text{before}})\big((\mathbf{x}_{\text{before}})_k - r_k\big)$, and recall the residual definition from Theorem C.2:

$$\mathcal{I}(\mathbf{x}_{\text{before}}) := L(\mathbf{x}_{\text{after}}) - \left(L(\mathbf{x}_{\text{before}}) - \sum_{k \in S_T(\mathbf{x}_{\text{before}})} \phi_k(\mathbf{x}_{\text{before}})\right).$$

Combining the two bounds yields

$$|\mathcal{I}(\mathbf{x}_{\text{before}})| \le \frac{M}{2}\|\Delta\|^2 + M\sqrt{|S_T(\mathbf{x}_{\text{before}})|}\,\epsilon^2 \le \left(\frac{M}{2} + M\sqrt{|S_T(\mathbf{x}_{\text{before}})|}\right)\epsilon^2\,,$$

and using $\|\Delta\| \le \epsilon$ completes the proof. □

D Bounding demographic parity by discriminative risk

Beyond individual fairness, we show that reducing individuallevel DR also controls group level statistical disparities such as DP. Intuitively, features that are fairnesscritical at the individual level tend to correlate with (or encode) the sensitive attribute, so neutralizing their effect reduces systematic differences between demographic groups. This section formalizes that link.

Notation and setup. Let $\mathcal{X}$ denote the feature space, $\mathcal{F}$ denotes full set of features, and each instance be written as $\mathbf{x} = (\check{\mathbf{x}}, A)$, where $\check{\mathbf{x}} \in \mathcal{X}_{\mathcal{F}\setminus\{A\}}$ are the non-sensitive features and $A \in \{0,1\}$ is the sensitive attribute (0 and 1 representing the unprivileged and privileged groups, respectively). For a classifier $f : \mathcal{X}_{\mathcal{F}\setminus\{A\}} \times \{0,1\} \to [0,1]$, we write $f(\check{\mathbf{x}}, A = a) = 1$, if the instance is predicted positive.

Discriminative risk (DR) For a single instance $\mathbf{x} = (\check{\mathbf{x}}, A)$,

$$L_{\mathrm{DR}}(f, \mathbf{x}) = \big|\, f(\check{\mathbf{x}}, A = 0) - f(\check{\mathbf{x}}, A = 1) \,\big| \,.$$

For any distribution D over $\check{\mathbf{x}}$ (e.g., the empirical mixture of the two groups), the population DR is

$$L_{\mathrm{DR}}(f; D) = \mathbb{E}_{\check{\mathbf{x}}\sim D}\big[L_{\mathrm{DR}}(f, \mathbf{x})\big] = \mathbb{E}_{\check{\mathbf{x}}\sim D}\big|\, f(\check{\mathbf{x}}, A = 0) - f(\check{\mathbf{x}}, A = 1) \,\big| \,. \tag{10}$$

Demographic parity (DP) As introduced earlier in equation (7), the classical definition of demographic parity is based on the difference in positive prediction rates between the two sensitive groups. For clarity, we rewrite it here and denote it with a superscript "decision":

$$L_{\mathrm{DP}}^{\mathrm{decision}}(f) = \big|\mathbb{P}_{\mathcal{D}}\big(f(\mathbf{x}) = 1 \mid A = 0\big) - \mathbb{P}_{\mathcal{D}}\big(f(\mathbf{x}) = 1 \mid A = 1\big)\big| \,. \tag{11}$$

In practice, however, modern models typically output continuous scores (e.g., probabilities in $[0,1]$), and the decision-based formulation depends on the choice of a threshold to binarize predictions. This threshold dependence makes the measure unstable and less suitable for optimization. To address this, we adopt a score-based relaxation of demographic parity, which compares the average prediction scores between the two groups. Let D_0 denote the conditional distribution of $\check{\mathbf{x}}$ given $A = 0$, and D_1 denote the conditional distribution of $\check{\mathbf{x}}$ given $A = 1$. The score-based DP gap is defined as

$$L_{\mathrm{DP}}^{\mathrm{score}}(f) = \big|\, \mathbb{E}_{\check{\mathbf{x}}\sim D_0}[f(\check{\mathbf{x}}, A = 0)] - \mathbb{E}_{\check{\mathbf{x}}\sim D_1}[f(\check{\mathbf{x}}, A = 1)] \,\big| \,. \tag{12}$$

D.1 Identical conditional distributions

Theorem D.1 (Identical-distribution case)**.** *If $D_0 = D_1 = D$, then for any classifier f,*

$$L_{\mathrm{DP}}^{score}(f) \;\le\; L_{\mathrm{DR}}(f; D) \,.$$

Proof. Under $D_0 = D_1 = D$,

$$\begin{aligned} L_{\mathrm{DP}}^{\mathrm{score}}(f) &= \big|\, \mathbb{E}_{\check{\mathbf{x}}\sim D}[f(\check{\mathbf{x}}, A = 0)] - \mathbb{E}_{\check{\mathbf{x}}\sim D}[f(\check{\mathbf{x}}, A = 1)] \,\big| \\ &= \big|\, \mathbb{E}_{\check{\mathbf{x}}\sim D}\big[f(\check{\mathbf{x}}, A = 0) - f(\check{\mathbf{x}}, A = 1)\big] \,\big| \\ &\le\; \mathbb{E}_{\check{\mathbf{x}}\sim D}\big|f(\check{\mathbf{x}}, A = 0) - f(\check{\mathbf{x}}, A = 1)\big| \\ &= L_{\mathrm{DR}}(f; D) \,, \end{aligned}$$

where we used Jensen's inequality. □

D.2 General case: distribution shift between groups

We recall that D_0 denotes the conditional distribution of $\check{\mathbf{x}}$ given $A = 0$, and D_1 denotes the conditional distribution of $\check{\mathbf{x}}$ given $A = 1$. We assume throughout this section that

$$f(\check{\mathbf{x}}, A) \in [0,1] \,, \quad \text{for } A \in \{0,1\} \,.$$

Definition D.2 (Total variation distance)**.** Let D_0, D_1 be two probability distributions over the non-sensitive feature space $\check{\mathbf{x}} \in \mathcal{X}_{\mathcal{F}\setminus\{A\}}$. The *total-variation (TV) distance* between them is defined as

$$\mathrm{TV}(D_0, D_1) \;=\; \sup_{B \subseteq \mathcal{X}_{\mathcal{F}\setminus\{A\}}} \big|\, D_0(B) - D_1(B) \,\big| \,, \tag{13}$$

where the supremum sup is taken over all measurable subsets of the non-sensitive feature space. Equivalently, for any measurable $h : \mathcal{X}_{\mathcal{F}\setminus\{A\}} \to [0,1]$,

$$\left|\, \mathbb{E}_{\check{\mathbf{x}}\sim D_0}[h(\check{\mathbf{x}})] - \mathbb{E}_{\check{\mathbf{x}}\sim D_1}[h(\check{\mathbf{x}})] \,\right| \;\le\; \mathrm{TV}(D_0, D_1)\,. \tag{14}$$

Theorem D.3 (General case: distribution shift)**.** *For any classifier f with outputs in* $[0,1]$*,*

$$L_{\mathrm{DP}}^{score}(f) \;\le\; \min\{\, L_{\mathrm{DR}}(f; D_0),\, L_{\mathrm{DR}}(f; D_1)\,\} \;+\; \mathrm{TV}(D_0, D_1)\,. \tag{15}$$

Proof. Starting from the definition of the score-based DP gap:

$$L_{\mathrm{DP}}^{\mathrm{score}}(f) = \left|\, \mathbb{E}_{\check{\mathbf{x}}\sim D_0}[f(\check{\mathbf{x}}, 0)] - \mathbb{E}_{\check{\mathbf{x}}\sim D_1}[f(\check{\mathbf{x}}, 1)] \,\right|.$$

Introduce a coupling $\pi(\check{\mathbf{x}}_0, \check{\mathbf{x}}_1)$ with marginals D_0 and D_1. Then

$$L_{\mathrm{DP}}^{\mathrm{score}}(f) = \left|\, \mathbb{E}_{\pi}\big[f(\check{\mathbf{x}}_0, 0) - f(\check{\mathbf{x}}_1, 1)\big] \,\right|.$$

Add and subtract $f(\check{\mathbf{x}}_0, 1)$ inside the expectation:

$$f(\check{\mathbf{x}}_0, 0) - f(\check{\mathbf{x}}_1, 1) = \big(f(\check{\mathbf{x}}_0, 0) - f(\check{\mathbf{x}}_0, 1)\big) + \big(f(\check{\mathbf{x}}_0, 1) - f(\check{\mathbf{x}}_1, 1)\big)\,.$$

By the triangle inequality,

$$L_{\mathrm{DP}}^{\mathrm{score}}(f) \;\le\; \mathbb{E}_{\pi}\big|f(\check{\mathbf{x}}_0, 0) - f(\check{\mathbf{x}}_0, 1)\big| + \big|\mathbb{E}_{\pi}[f(\check{\mathbf{x}}_0, 1) - f(\check{\mathbf{x}}_1, 1)]\big|\,.$$

The first term gives $L_{\mathrm{DR}}(f; D_0)$, and by symmetry the same argument with $f(\check{\mathbf{x}}_1, 0)$ gives $L_{\mathrm{DR}}(f; D_1)$. Thus we can bound by the smaller of the two:

$$\min\{\, L_{\mathrm{DR}}(f; D_0),\, L_{\mathrm{DR}}(f; D_1)\,\}\,.$$

For the second term, define $h(\check{\mathbf{x}}) = f(\check{\mathbf{x}}, 1) \in [0,1]$. Then

$$\mathbb{E}_{\pi}\big[f(\check{\mathbf{x}}_0, 1) - f(\check{\mathbf{x}}_1, 1)\big] = \mathbb{E}_{\check{\mathbf{x}}\sim D_0}[h(\check{\mathbf{x}})] - \mathbb{E}_{\check{\mathbf{x}}\sim D_1}[h(\check{\mathbf{x}})]\,,$$

which is bounded by $\mathrm{TV}(D_0, D_1)$ via (14).

Therefore,

$$L_{\mathrm{DP}}^{\mathrm{score}}(f) \;\le\; \min\{\, L_{\mathrm{DR}}(f; D_0),\, L_{\mathrm{DR}}(f; D_1)\,\} + \mathrm{TV}(D_0, D_1)\,.$$

□

Implications.

- If $D_0 = D_1$, then $\mathrm{TV}(D_0, D_1) = 0$, and the DP gap is bounded directly by the population discriminative risk, i.e., $L_{\mathrm{DR}}(f; D)$; equivalently, since $D_0 = D_1 = D$, one can also write the bound in terms of either group-specific risk $L_{\mathrm{DR}}(f; D_0)$ or $L_{\mathrm{DR}}(f; D_1)$.
- If $D_0 \neq D_1$, the theorem shows that reducing the discriminative risk of either group still tightens an upper bound on the DP gap. However, an irreducible term $\mathrm{TV}(D_0, D_1)$ remains due to inherent distribution shift rather than model bias.

E Empirical Validation of Matching Distance

Proposition C.3 relies on the premise that the matching distance $\epsilon = \|\mathbf{x} - \mathbf{r}\|$ remains sufficiently small, so that the higher-order interaction residual $\mathcal{I}(\mathbf{x})$ is suppressed relative to the first-order Shapley terms. In addition, since our method relies on distance-based matching, we validate that the selected matched pairs are indeed closer than random pairings even in high-dimensional and potentially sparse feature spaces, where Euclidean distances can be less reliable.

For each query instance $\mathbf{x}$, we select a reference $\mathbf{r}$ using our matching strategy (Nearest Neighbor matching with $k = 1$; we use the same feature space and preprocessing as in the main experiments). We then compute the matching distance $\epsilon(\mathbf{x}) = \|\mathbf{x} - \mathbf{r}\|_2$. As a baseline, we also pair each $\mathbf{x}$ with a randomly sampled reference $\tilde{\mathbf{r}}$ from the reference pool and compute $\|\mathbf{x} - \tilde{\mathbf{r}}\|_2$. We report summary statistics over $N = 5{,}000$ query instances.

Table 3 shows that matched pairs have substantially smaller Euclidean distances than random pairs across two large and high-dimension datasets, Adult and Census Income KDD. The differences are statistically significant

Table 3: **Quantitative validation of matching quality.** Comparison of Euclidean distances in feature space between matched pairs (via KNN, $k = 1$) and random pairs across two datasets ($N = 5{,}000$). All differences are statistically significant with $p < 0.001$. The t-statistic and Cohen's d compare matched vs. random distances.

Dataset	Group	Mean Dist.	Std. Dev.	t-statistic	Cohen's d
Adult	Matched	1.56	0.75	101.55	2.03
	Random	3.98	1.51		
KDD	Matched	1.95	1.40	125.69	2.51
	Random	6.32	2.02		

($p < 0.001$) with large effect sizes (Cohen's $d = 2.03$ on Adult and $d = 2.51$ on Census KDD), indicating that our matching strategy consistently retrieves structurally similar references within the local neighborhood of $\mathbf{x}$. Combined with the quadratic scaling $|\mathcal{I}(\mathbf{x})| \leq C\epsilon^2$ in Proposition C.3, these results support the practical validity of the approximation used in Theorem C.2.

Practical Considerations in High Dimensions. While Euclidean distances can become less informative in extremely high-dimensional sparse spaces, the observed large effect sizes indicate that the proposed matching remains discriminative for the datasets considered. For more challenging regimes, we recommend mitigating the curse of dimensionality by: (i) applying dimensionality reduction prior to KNN matching to improve metric reliability, or (ii) substituting standard Optimal Transport with Sliced Optimal Transport, which often provides improved statistical robustness and computational efficiency in high-dimensional feature spaces.

F Computational Complexity of FairSHAP

Setup. The dataset is split by the sensitive attribute into $\mathcal{G} \in \mathbb{R}^{n\times d}$ and $\widetilde{\mathcal{G}} \in \mathbb{R}^{m\times d}$ with d features. Let C_f be the cost of one forward pass of the model f. The permutation Shapley estimator uses M permutations per instance and R reference draws per set-function evaluation. With entropic optimal transport (OT) matching, let I denote the number of Sinkhorn iterations. A *one-side* run means a single call to `FairSHAP` with a fixed target group and the other as reference (e.g., target $= \mathcal{G}$, reference $= \widetilde{\mathcal{G}}$). The *both-sides* setting refers to the full framework that runs `FairSHAP` twice with roles swapped and then concatenates the two modified groups.

Matching. With nearest-neighbor (NN) matching, forming a dense cost matrix over all $n \times m$ pairs in d dimensions costs $O(nmd)$ time; normalizing and taking rowwise argmax adds $O(nm)$. Storing the joint matrix $\mathcal{P}$ uses $O(nm)$ space. With entropic OT, the total time is $O(nmd + Inm)$ and the space remains $O(nm)$.

Shapley estimation (one side). Using a permutation prefix-chain estimator, each of the n target instances performs Md set-function evaluations; each is approximated with R reference draws and one DR evaluation of cost $\mathcal{O}(C_f)$. The time is

$$T_{\text{Shapley}} = \underbrace{\mathcal{O}\big(n\,M\,d\,R\,C_f\big)}_{\text{No more complex than standard SHAP}}, \qquad \text{space } \mathcal{O}(nd)\,.$$

Sampling from $\mathcal{P}(\cdot \mid \mathbf{g}_i)$ can be made $\mathcal{O}(1)$ per draw after $\mathcal{O}(m)$ preprocessing per row, which does not change the asymptotics dominated by model evaluations.

Reference construction and edits. After obtaining the joint matrix $\mathcal{P}$, each target instance i selects a reference index j^* either by (i) deterministic row-wise $\arg\max_j \mathcal{P}_{i,j}$, which costs $\mathcal{O}(nm)$ overall, or (ii) stochastic sampling of R references per row, which costs $\mathcal{O}(nR)$ in total after an $\mathcal{O}(m)$ preprocessing per row. Building the reference table $\mathcal{B}$ by copying the selected reference feature vectors requires $\mathcal{O}(nd)$ time and space. Subsequently, thresholding the Shapley values and replacing the corresponding features in $\mathcal{B}$ adds another $\mathcal{O}(nd)$. Thus the overall cost of this stage in (i) is

$$T_{\text{ref}} = \mathcal{O}(nm + nd)\,, \qquad S_{\text{ref}} = \mathcal{O}(nd)\,,$$

and can be streamed to avoid extra peak memory.

End-to-end cost (one side). With NN matching:

$$T_{\text{one}} = \mathcal{O}\big(nmd + nMdRC_f\big)\,, \qquad S_{\text{one}} = \mathcal{O}\big(nm + nd\big)\,.$$

Overall framework (both sides). The dense pairwise cost/similarity matrix can be computed once and then reused (or transposed) when swapping roles, so the $O(nmd)$ term is incurred once, while Shapley scales with the number of target instances on each side:

$$T_{\text{total}} = \mathcal{O}\big(nmd + (n{+}m)\, MdRC_f\big)\,, \qquad S_{\text{total}} = \mathcal{O}\big(nm + (n{+}m)d\big)\,.$$

Remarks. (i) When model inference is expensive, the Shapley term $\mathcal{O}((n{+}m)MdRC_f)$ dominates; for very large n, m and moderate d, the $\mathcal{O}(nmd)$ matching can dominate.
(ii) Exact Shapley per side requires $\mathcal{O}(n\, 2^d\, R\, C_f)$ time, and the computation grows exponentially with d, quickly becoming infeasible beyond small feature dimensions.
(iii) Compared with existing alternatives, KernelSHAP requires $\mathcal{O}(MC_f + Md^2)$ time per instance, where M is the number of sampled coalitions (Exact Shapley corresponds to the special case $M = 2^d$, which is exponential in d). For tree ensembles, TreeSHAP achieves polynomial-time computation, with complexity $\mathcal{O}(TLD^2)$ in the number of trees T, maximum leaves L, and depth D (often simplified to $\mathcal{O}(d^2)$ when $D \approx d$). Our prefix-chain estimator reduces the per-side cost to $\mathcal{O}(nMdRC_f)$. If more efficient Shapley value approximation methods become available, they can be plugged into the FairSHAP framework to further accelerate attribution without changing the overall structure.

G Threshold Selection

In this section, we justify the choice of the Shapley threshold $T = 0.05$ used throughout our experiments. Recall that T determines which features are considered fairness-critical and thus subject to modification: smaller thresholds yield fewer edits (potentially insufficient for fairness improvement), while larger thresholds increase the number of edits (at the risk of introducing noise and harming data fidelity).

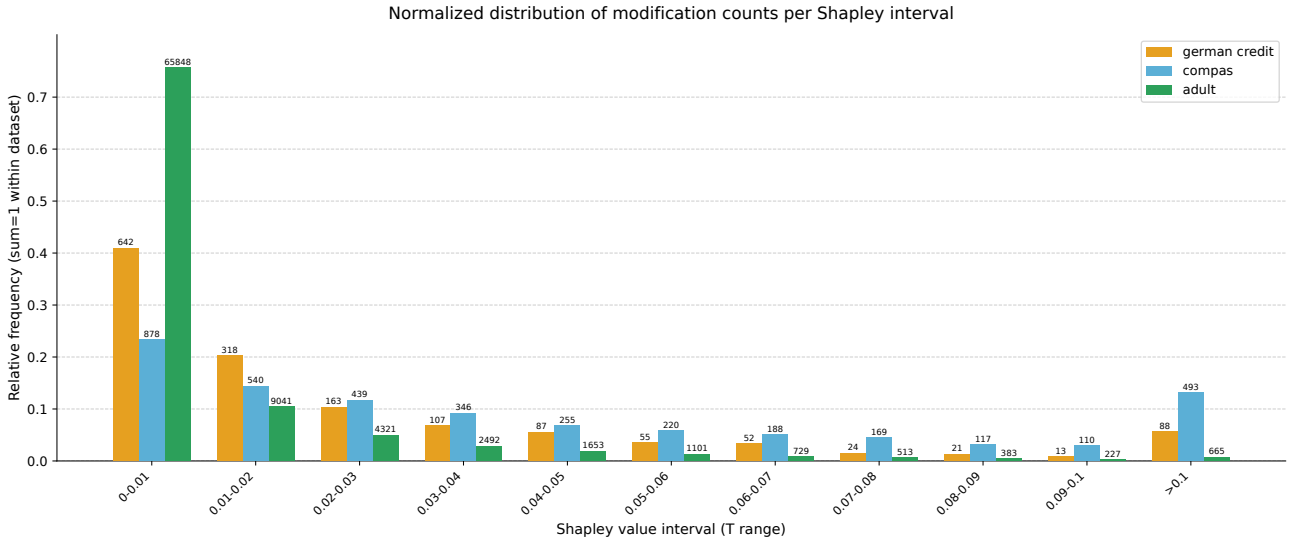

Figure 4: Normalized distribution of modification counts per Shapley interval. Each color corresponds to a dataset (German Credit, COMPAS, Adult). Y-axis shows relative frequency (summing to 1 within each dataset), while numbers above bars indicate the absolute counts of modifications in each interval.

Distribution of modification counts

To better understand the effect of thresholding, we analyze the distribution of feature modifications across different Shapley value intervals. For each dataset, we group the absolute Shapley values into bins of size 0.01 up to 0.10, and a final bin for values larger than 0.10. Figure 4 shows the normalized histograms: the bar heights are relative frequencies (summing to one within each dataset), while the numbers annotated above the bars indicate the raw modification counts.

Observations

From below figure 4, several patterns emerge:

- **Early growth and plateau.** Most edits concentrate within $|\phi| \in [0.01, 0.05]$. Beyond this range, additional intervals contribute relatively few new edits.
- **Stable mid-range.** Across datasets, the distributions flatten around $T = 0.05$, indicating that this value lies in a regime where edits remain meaningful but not excessive.
- **Avoiding extremes.** Very large thresholds (> 0.07) fail to capture sufficient fairness-critical features, while very small thresholds (< 0.03) incorporate many near-zero contributions, which may introduce noise and cost much time.

Choice of $T = 0.05$

Empirical results across datasets confirm that T values in the range $[0.03, 0.07]$ strike a favorable balance between fairness gains and data fidelity. We therefore adopt $T = 0.05$ as a default threshold:

1. it captures the majority of fairness-critical contributions without excessive edits;
2. it avoids the ambiguity of near-zero Shapley values;
3. it yields consistent improvements in both individual and group fairness metrics (DR, DP, EO), with minimal accuracy degradation.

This threshold thus represents a principled and empirically validated trade-off point for FairSHAP across diverse datasets. *We further suggest that practitioners may tune the threshold on a validation set to identify the most suitable value for a given dataset and application.*

H Robustness across different ML models

We acknowledge that Shapley values are inherently dependent on the underlying predictive model. Nevertheless, FairSHAP can naturally adapt to different architectures by identifying and mitigating those feature contributions that most strongly affect fairness for a given predictor. To evaluate the stability of our method across classifiers, we extend our experiments on the German Credit and COMPAS datasets using three representative models: XGBoost [Chen and Guestrin, 2016], SVM [Cortes and Vapnik, 1995], and TabPFN [Hollmann et al., 2025]. The results in Table 4 demonstrate that FairSHAP consistently improves fairness metrics across different model architectures, confirming its robustness to variations in predictive structure.

Table 4: Model performance on German Credit and COMPAS datasets (mean ± std), with sex as the sensitive attribute. SHAP values are computed using KernelSHAP. FairSHAP results that outperform the baseline are highlighted in **bold**.

Model	Method	Accuracy	DR	DP	EO	PQP
German Credit (sex)						
XGBoost	Baseline	0.6650±0.0257	0.0785±0.0211	0.0512±0.0346	0.1287±0.0590	0.1341±0.0486
	FairSHAP	0.6630±0.0275	**0.0243±0.0112**	**0.0301±0.0347**	**0.1126±0.0783**	0.1852±0.1074
SVM	Baseline	0.7170±0.0175	0.0368±0.0139	0.0374±0.0327	0.0929±0.0948	0.3728±0.1957
	FairSHAP	0.7100±0.0240	**0.0306±0.0099**	**0.0335±0.0643**	0.1138±0.1049	**0.3185±0.0885**
TabPFN	Baseline	0.7170±0.0236	0.0354±0.0118	0.0609±0.0292	0.1429±0.1317	0.2863±0.1158
	FairSHAP	0.7130±0.0271	**0.0220±0.0057**	**0.0477±0.0628**	**0.1010±0.1204**	0.3027±0.1938
COMPAS (sex)						
XGBoost	Baseline	0.6698±0.0051	0.0883±0.0064	0.1548±0.0241	0.1243±0.0510	0.0492±0.0084
	FairSHAP	0.6609±0.0106	**0.0629±0.0091**	**0.1326±0.0407**	**0.0985±0.0603**	**0.0452±0.0383**
SVM	Baseline	0.6759±0.0101	0.0284±0.0039	0.1688±0.0096	0.1717±0.0378	0.0658±0.0246
	FairSHAP	0.6738±0.0129	**0.0263±0.0056**	**0.1542±0.076**	**0.1557±0.0359**	**0.0598±0.0326**
TabPFN	Baseline	0.6876±0.0103	0.0548±0.0092	0.1981±0.0289	0.1771±0.0500	0.0234±0.0030
	FairSHAP	0.6842±0.0122	**0.0465±0.0044**	**0.1614±0.0225**	**0.1416±0.0527**	0.0458±0.0162

I Robustness across different SHAP kernels

We further examined robustness under different Shapley-value approximations. Specifically, we compared KernelSHAP [Roshan and Zafar, 2022], SamplingSHAP [Lundberg and Lee, 2017], and PermutationSHAP [Strumbelj and Kononenko, 2010] on German Credit and COMPAS with XGBoost fixed as the base model. As shown in Table 5, FairSHAP yields consistent fairness gains regardless of explainer, with KernelSHAP offering the best trade-off between accuracy, fairness, and runtime. These results indicate that FairSHAP's effectiveness is not tied to a particular estimator and can readily benefit from future advances in scalable Shapley-value computation.

Table 5: Comparison of FairSHAP performance using different Shapley value estimators on the German Credit and COMPAS datasets (mean ± std). All results are computed based on XGBoost, with sex as the sensitive attribute. Shapley values are approximated via KernelSHAP, SamplingSHAP, and PermutationSHAP, respectively.

Method	**Accuracy**	**DR**	**DP**	**EO**	**PQP**	**Time (s)**
			German Credit (sex)			
Baseline	0.665±0.026	0.079±0.021	0.051±0.035	0.129±0.059	0.134±0.049	—
Kernel explainer	0.663±0.028	0.024±0.011	0.030±0.035	0.113±0.078	0.185±0.107	19
Sampling explainer	0.668±0.039	0.027±0.011	0.043±0.036	0.169±0.086	0.178±0.137	224
Permutation explainer	0.660±0.027	0.024±0.014	0.043±0.031	0.118±0.100	0.195±0.103	42
			COMPAS (sex)			
Baseline	0.670±0.005	0.088±0.006	0.155±0.024	0.124±0.051	0.049±0.008	—
Kernel explainer	0.661±0.011	0.063±0.009	0.133±0.041	0.099±0.060	0.045±0.038	40
Sampling explainer	0.654±0.012	0.062±0.006	0.142±0.025	0.114±0.041	0.046±0.039	556
Permutation explainer	0.661±0.014	0.060±0.003	0.141±0.027	0.115±0.044	0.046±0.040	382