Bridging Explainability and Fairness in Machine Learning: From DR[1] to FairSHAP[2,3]

Yijun Bian

Department of Computer Science
University of Copenhagen

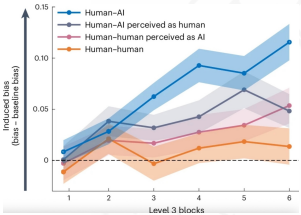
13 May 2026

[1] ***Yijun Bian****, "Improving fairness with ensemble combination: Margin-dependent bounds," in *the 9th annual ACM Conference on Fairness, Accountability, and Transparency (FAccT 2026)*, accepted.

[2] Lin Zhu†, ***Yijun Bian***†, and Lei You*, "FairSHAP: Preprocessing for fairness through attribution-based data augmentation," in *the 29th International Conference on Artificial Intelligence and Statistics (AISTATS 2026)*, accepted.

[3] †Equal contribution, *corresponding author. Joint work (Technical University of Denmark & University of Copenhagen)

AI/ML is everywhere now



Social identity biases exist not only in human psychology and social behaviour, but also are present in artificial intelligence (AI) systems.[4]

When humans and AI interact, even minute perceptual, emotional and social biases—originating either from AI systems or humans—leave human beliefs more biased, potentially forming a feedback loop.[5]

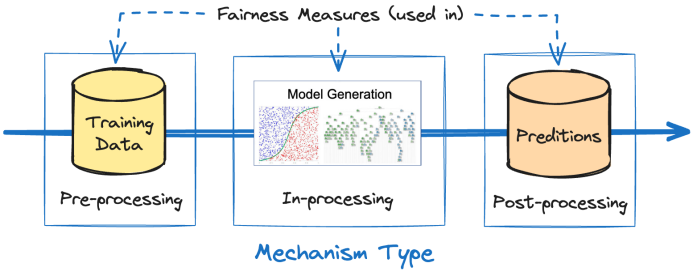
[4] Ziad Obermeyer et al. "Dissecting racial bias in an algorithm used to manage the health of populations". In: *Science* 366.6464 (2019), pp. 447–453. DOI: 10.1126/science.aax2342; Tiancheng Hu et al. "Generative language models exhibit social identity biases". In: *Nat Comput Sci* (2024), pp. 1–11; Richard J Chen et al. "Algorithmic fairness in artificial intelligence for medicine and healthcare". In: *Nat Biomed Eng* 7.6 (2023), pp. 719–742.

[5] Moshe Glickman and Tali Sharot. "How human-AI feedback loops alter human perceptual, emotional and social judgements". In: *Nat Hum Behav* 9 (2025), pp. 345–359. DOI: 10.1038/s41562-024-02077-2; Madalina Vlasceanu and David M Amodio. "Propagation of societal gender inequality by internet search algorithms". In: *Proc Natl Acad Sci U.S.A.* 119.29 (2022), e2204529119.

Existing work about fairness

Many sources of bias [6]

Mechanisms to enhance fairness [7]



Types of fairness measures

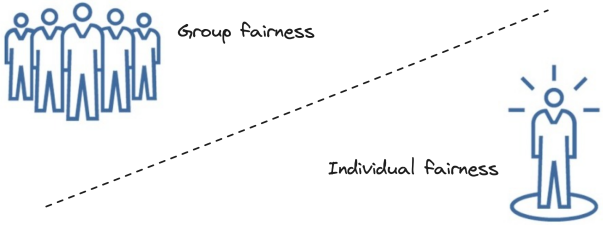
[6] *Unintential*: Limited and coarse features, sample size disparity (less data by definition about minority populations), skewed sample (feedback loops), tainted examples, features that act as proxies; *Intentional*: conscious prejudice.

[7] ***Pre-*** and ***post-processing*** *mechanisms* normally function by manipulating input or output, while ***inprocessing*** *mechanisms* introduce fairness constraints into training procedures or algorithmic objectives.

Existing work about fairness

Many sources of bias
Mechanisms to enhance fairness

Types of fairness measures [6,7]



Challenging:
incompatibility,[8] multi-attribute protection, etc.

[6] ***Distributive*** *fairness*: group fairness, individual fairness, counterfactual fairness, etc.; ***Procedural*** *fairness*

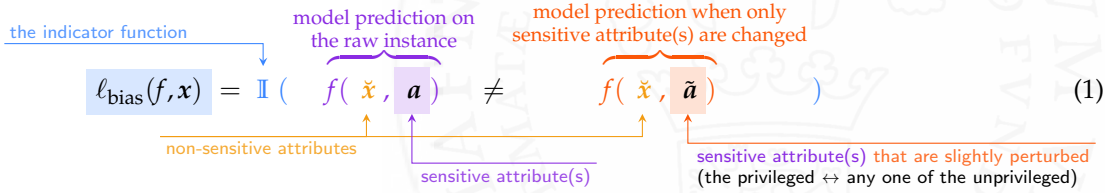
[7] *Group fairness* focuses on statistical/demographic equality among groups defined by sensitive attributes, while *individual fairness* follows a principle that "similar individuals should be evaluated or treated similarly."

[8] Tensions between notions of fairness, between fairness and accuracy, between different methods for achieving fairness

Discriminative risk (DR) [11]

—from an individual-level aspect

Following the principle of individual fairness *(the treatment/evaluation on one instance should not change solely due to minor changes in its sensitive attributes)*, with an instance denoted by $\boldsymbol{x} = (\check{\boldsymbol{x}}, \boldsymbol{a})$, the fairness quality of one hypothesis[9] $f(\cdot)$ could be evaluated by



similarly to the 0/1 loss, where $\tilde{\boldsymbol{a}}$ is a perturbed[10] $\boldsymbol{a} = [a_1, ..., a_{n_a}]^\top$, $a_i \in \mathcal{A}_i$, $n_a \geqslant 1$, and $|\mathcal{A}_i| \geqslant 2$. Note that Eq. (1) is evaluated on only one instance with sensitive attributes $\boldsymbol{x}$.

[9] The hypothesis used in this equation could indicate an individual classifier or an ensemble classifier.

[10] A member belonging to the marginalised group(s) will be perturbed into the privileged group, while a member of the privileged group is perturbed randomly into one of the marginalised groups.

[11] Yijun Bian. "Improving fairness with ensemble combination: Margin-dependent bounds". In: *FAccT*. Accepted. 2026.

Discriminative risk (DR) [11]

—from an individual-level aspect
—from a group-level aspect

To describe this characteristic of the hypothesis on multiple instances, then the empirical discriminative risk on one dataset S is expressed as $\hat{\boldsymbol{L}}_{\text{bias}}(f, S) = \frac{1}{n}\sum_{i=1}^{n} \ell_{\text{bias}}(f, \boldsymbol{x}_i)$, and the true discriminative risk[9] of the hypothesis over a data distribution is $\boldsymbol{L}_{\text{bias}}(f) = \mathbb{E}_{(\boldsymbol{x},y)\sim\mathcal{D}}[\ell_{\text{bias}}(f, \boldsymbol{x})]$, respectively. Note that the empirical DR on S is an unbiassed estimation of the true DR.

$$\ell_{\text{bias}}(f, \boldsymbol{x}) = \mathbb{I}(f(\check{\boldsymbol{x}}, \boldsymbol{a}) \neq f(\check{\boldsymbol{x}}, \tilde{\boldsymbol{a}}))$$

$$\text{DR}(f) = \mathbb{E}_{\mathcal{D}}[\mathbb{I}(f(\check{\boldsymbol{x}}, \boldsymbol{a}) \neq f(\check{\boldsymbol{x}}, \tilde{\boldsymbol{a}}))]$$

- Widely applicable, allowing one or more SAs, and each SA allowing binary or multiple values
- Different from existing (group/individual/counterfactual) fairness measures
- Limitations: small values; computational results may be affected somehow by a randomness factor[10]

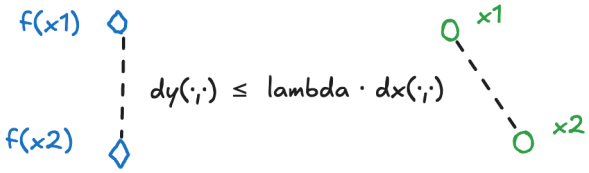
[9] The instances from S are independent identically distributed (i.i.d.) drawn from an input/feature-output/label space $\mathcal{X}\times\mathcal{Y}$ according to an unknown distribution $\mathcal{D}$.

[10] With p probability, the instance will be perturbed into another group.

[11] Bian, see n. 11.

Differences from existing fairness measures

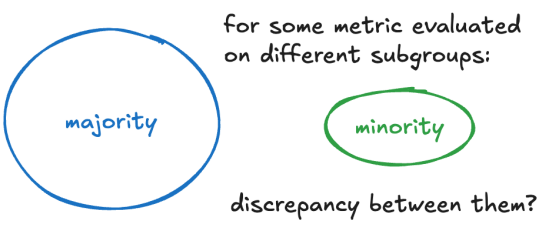
- Two distinctions from *individual fairness* measures
  1. relies on the choice of similarity/distance metric
  2. instance pairs in comparison coming from original data



- Two distinctions from *group fairness* measures
- Four distinctions from *causal fairness*

Differences from existing fairness measures

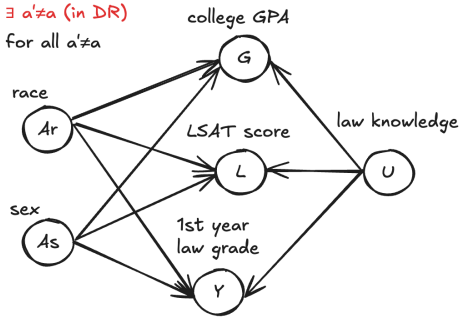
- Two distinctions from *individual fairness* measures
- Two distinctions from *group fairness* measures
  1. works for only one sensitive attribute (usually bi-valued)
  2. computing separately for each subgroup, then difference



- Four distinctions from *causal fairness*

Differences from existing fairness measures

- Two distinctions from *individual fairness* measures
- Two distinctions from *group fairness* measures
- Four distinctions from *causal fairness*
  1. works for only one sensitive attribute (although possibly multi-valued)
  2. based on causal models/graphs, not a quantitative measure
  3. non-sensitive attributes may vary with it in counterfactual fairness
  4. conditions for achieving them are stronger



Differences from existing fairness measures

- Two distinctions from *individual fairness* measures
- Two distinctions from *group fairness* measures
- Four distinctions from *causal fairness*

$$\ell_{\text{bias}}(f, \boldsymbol{x}) = \mathbb{I}\big(f(\check{\boldsymbol{x}}, \boldsymbol{a}) \neq f(\check{\boldsymbol{x}}, \tilde{\boldsymbol{a}})\big)$$

$$\hat{L}_{\text{bias}}(f, S) = \frac{1}{n}\sum_{i=1}^{n} \ell_{\text{bias}}(f, \boldsymbol{x}_i)$$

$$L_{\text{bias}}(f) = \mathbb{E}_{(\boldsymbol{x},y)\sim\mathcal{D}}\big[\, \ell_{\text{bias}}(f, \boldsymbol{x}) \,\big]$$

$$L'_{\text{bias}}(f) = \big|\mathbb{E}_{(\boldsymbol{x},y)\sim\mathcal{D}|\boldsymbol{a}=1}[\ell_{\text{bias}}(f, \boldsymbol{x})] - \mathbb{E}_{(\boldsymbol{x},y)\sim\mathcal{D}|\boldsymbol{a}=0}[\ell_{\text{bias}}(f, \boldsymbol{x})]\big|$$

- Similarities that *DR* shares with the existing fairness measures
  - follows the same principle as *individual fairness* measures
  - is computed over a group of instances (like one dataset or a data distribution)
  - indicates the discrimination level from a statistical/demographic perspective

Validating *DR*, a fairness quality measure

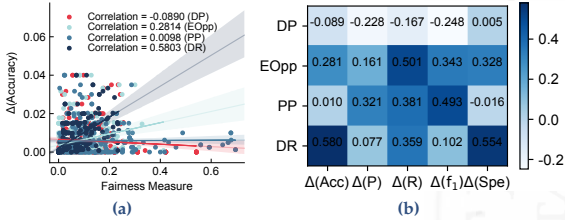


Figure 1. Comparison of the proposed discriminative risk (DR) with three group fairness measures, that is, DP, EOpp, and PP.
(a) Scatter diagrams with the degree of correlation, where the x- and y-axes are different fairness measures and the variation of accuracy between the raw and disturbed data. (b) Correlation among multiple criteria. Note that correlation here is calculated based on the results from all datasets.

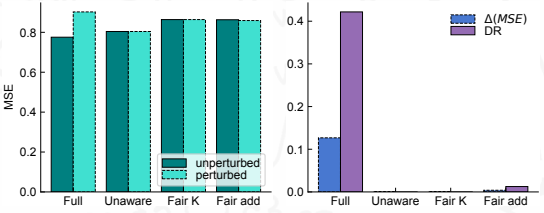


Figure 2. Example: law school success.
(a) Test MSE of different models, where 'unperturbed' and 'perturbed' denote the results obtained from the original and disturbed data respectively. (b) The comparison between the change in MSE and *DR*, which suggests that DR$\approx$0 when the corresponding model satisfies or nearly satisfies counterfactual fairness.

Improving fairness with ensemble combination

Consider *weighted voting* as the ensemble combination.[12] For brevity, we record $\sum_{j=1}^{m} w_j \mathbb{I}(f_j(\boldsymbol{x}) = y)$ as $q_y(\boldsymbol{x})$, and $y^* \triangleq \mathbf{wv}_\rho(\boldsymbol{x})$. Then we define the *top-1 voting margin* as

$$\gamma_\rho(\boldsymbol{x}) \triangleq q_{y^*}(\boldsymbol{x}) - \max_{y \neq y^*} q_y(\boldsymbol{x}) \in [0, 1]\,.$$

If the ensemble exhibits discrimination risk, it means that *the leading gap in the prediction class has been reversed*, in other words, at least half of the votes for the prediction class will be altered. That is to say, the total weight of the inconsistency between the original prediction and the perturbed prediction has to hold

$$\phi_\rho(\boldsymbol{x}) \triangleq \mathbb{E}_\rho[\ell_{\text{bias}}(f, \boldsymbol{x})] = \mathbb{E}_\rho[\mathbb{I}\big(f(\check{\boldsymbol{x}}, \boldsymbol{a}) \neq f(\check{\boldsymbol{x}}, \tilde{\boldsymbol{a}})\big)] \geqslant \tfrac{1}{2}\gamma_\rho(\boldsymbol{x})\,.$$

Therefore, we have the discriminative risk of one ensemble

$$\ell_{\text{bias}}(\mathbf{wv}_\rho, \boldsymbol{x}) \leqslant \mathbb{I}\big(\phi_\rho(\boldsymbol{x}) \geqslant \tfrac{1}{2}\gamma_\rho(\boldsymbol{x})\big) = \mathbb{I}\big(\phi_\rho(\boldsymbol{x})/\gamma_\rho(\boldsymbol{x}) \geqslant \tfrac{1}{2}\big)\,, \tag{2}$$

and can discuss some bounds concerning fairness for the weighted vote, inspired by prior work [13].

[12] where the prediction by an ensemble of m individual classifiers, parameterised by a weight vector $\rho = [w_1, ..., w_m]^\intercal \in [0, 1]^m$, is given by $\mathbf{wv}_\rho(\boldsymbol{x}) = \text{argmax}_{y \in \mathcal{Y}} \sum_{j=1}^{m} w_j \mathbb{I}(f_j(\boldsymbol{x}) = y)$, such that $\sum_{j=1}^{m} w_j = 1$, and w_j is the weight of individual classifier $f_j(\cdot)$.
It could be viewed as $w_j = 1/m$ for all $j \in \{1, 2, ..., m\}$ in plurality voting and majority voting, meaning that all individual classifiers will be taken into account equally. Note that ties will be resolved arbitrarily, and both parametric and non-parametric models can serve as individual classifiers.

[13] By applying Markov's inequality, second-order Markov's inequality, and Chebyshev-Cantelli inequality (Andrés R Masegosa et al. "Second order PAC-Bayesian bounds for the weighted majority vote". In: *NeurIPS*. vol. 33. Curran Associates, Inc., 2020, pp. 5263–5273)

Verification of the proposed *margin-dependent* bounds

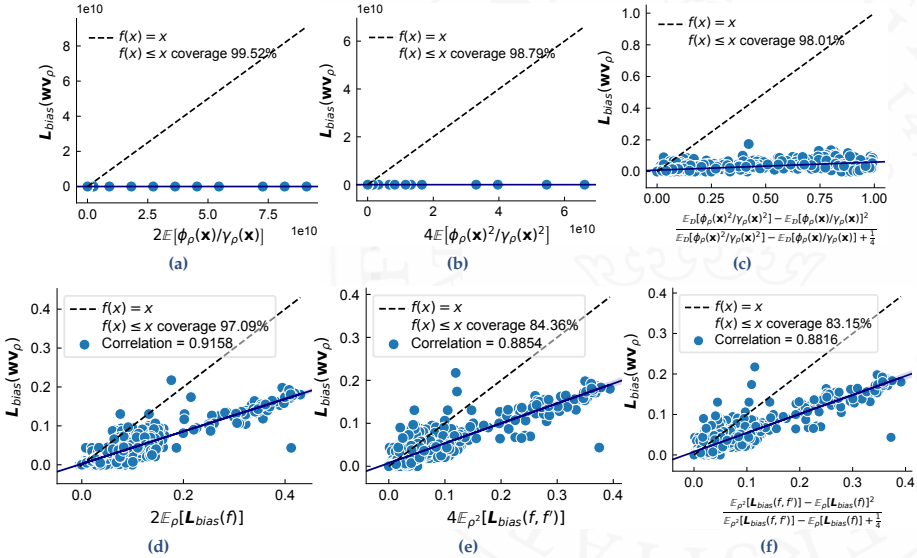


Figure 3. Verification of the proposed margin-dependent bounds, in comparison with prior work

Verification of the proposed *margin-dependent* bounds

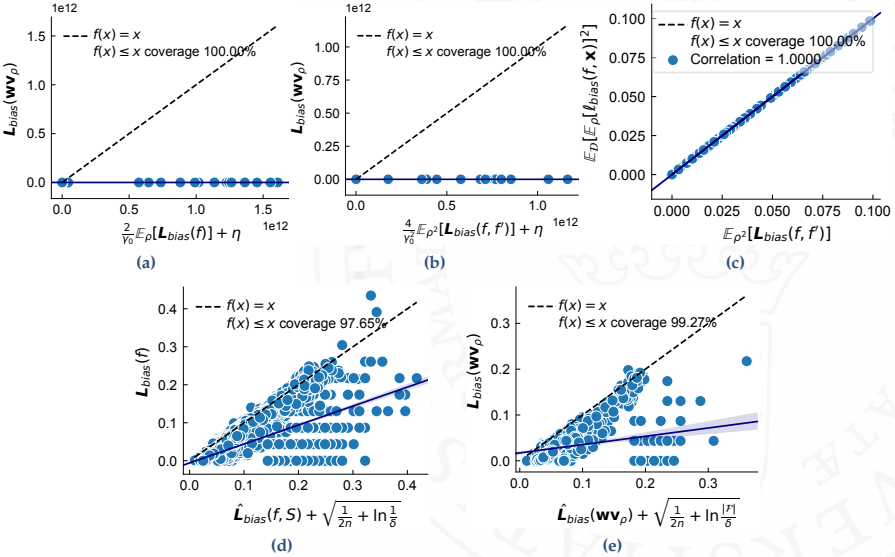
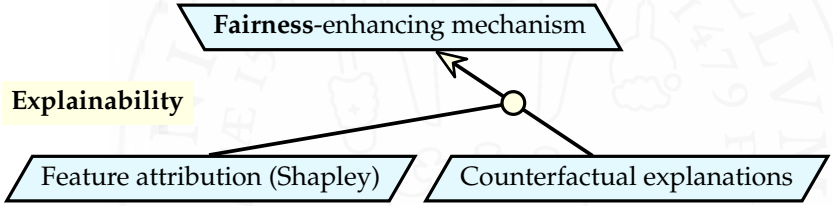


Figure 3. Verification of the proposed margin-dependent bounds's relaxation and PAC bounds

FairSHAP


Fairness-enhancing mechanism
Explainability
Feature attribution (Shapley)
Counterfactual explanations

Preliminaries

Shapley value

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

$$\phi_k(v) = \sum_{S \subseteq \mathcal{F} \setminus \{k\}} \frac{|S|!(n - |S| - 1)!}{n!} \big(v(S \cup \{k\}) - v(S) \big). \tag{3}$$

Discriminative risk

With $\mathbf{x} = (\check{\mathbf{x}}, A)$, 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 = unprivileged, 1 = privileged), discriminative risk (DR) is defined as:

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

FairSHAP

For instance i, define the coalition value on features $S \subseteq \mathcal{F}$ via an individual-fairness loss, where $\mathcal{P}$ is the joint probability obtained by nearest neighbour matching:

$$v^{(i)}(S) = \mathrm{E}_{\tilde{\boldsymbol{g}} \sim \mathcal{P}(\tilde{\boldsymbol{g}} | \boldsymbol{g}_i)} \Big[L_{\mathrm{DR}} \big(\boldsymbol{g}_{i,S}; \tilde{\boldsymbol{g}}_{\mathcal{F} \setminus S} \big) \Big] - \mathrm{E}_{\tilde{\boldsymbol{g}} \sim \mathcal{P}(\tilde{\boldsymbol{g}} | \boldsymbol{g}_i)} \big[L_{\mathrm{DR}} \big(\tilde{\boldsymbol{g}}_{\mathcal{F}} \big) \big], \tag{4a}$$

$$\text{s.t.} \quad \mathcal{P}(\boldsymbol{g}, \tilde{\boldsymbol{g}}) = \mathcal{M}_{\text{method}}(\mathcal{G}, \widetilde{\mathcal{G}}). \tag{4b}$$

Framework

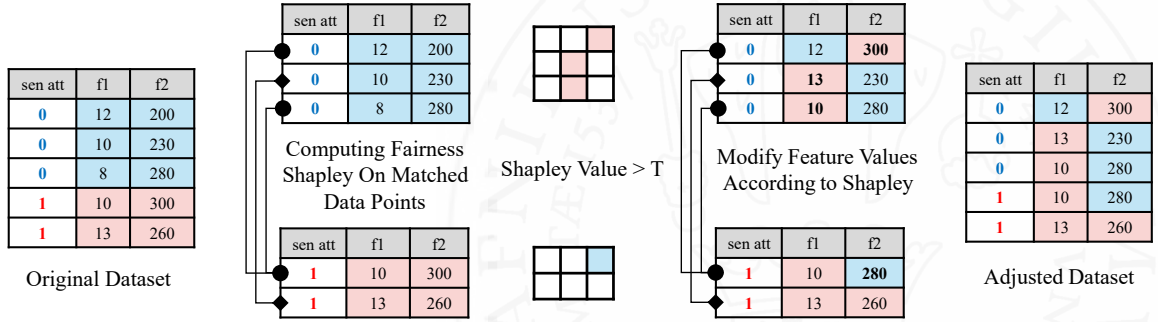
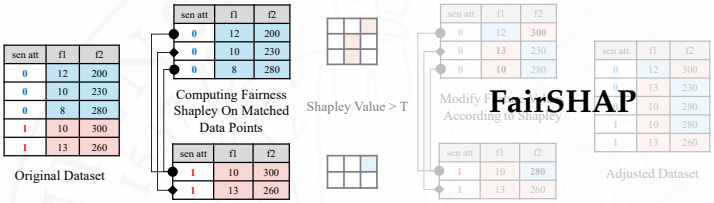


Figure 4. Overall framework of FairSHAP.[14]

• ***Left:*** *Partition the original dataset by the sensitive attribute into privileged andunprivileged groups, then perform nearest-neighbour matching between two groups.* • ***Middle:*** *Computed Shapley value matrix on the matched samples.* • ***Right:*** *Modify feature values according to the Shapley attributions and synthesize the adjusted dataset.*

[14] leverages Shapley value attribution to improve both individual and group fairness; *model-agnostic* and transparent; is broadly applicable to tabular data, supports various models and SHAP algorithms, and can be seamlessly integrated into existing ML pipelines

Algorithm 1



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 $\in \{\text{NearestNeighbour}, \text{OptimalTransport}\}$

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

3: $\widetilde{\mathcal{G}}' \leftarrow \texttt{FairSHAP}(\text{target} = \widetilde{\mathcal{G}},\ \text{non-target} = \mathcal{G},\ \text{model} = f,\ T,\ \mathcal{M}_{\text{method}})$ // 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** $\mathcal{D}_{\text{new}}$

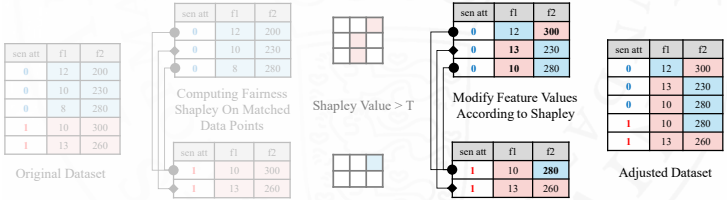
Algorithm 2

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 $\mathcal{M}_{\text{method}}(\mathcal{G}, \widetilde{\mathcal{G}})$ to obtain joint probability $\mathcal{P}(\mathbf{g}, \tilde{\mathbf{g}}) \in \mathbb{R}^{n \times m}$
Use Eqs. (3) and (4) to obtain Shapley value matrix $\boldsymbol{\phi} \in \mathbb{R}^{n \times d}$
Initialize reference data $\mathcal{B} \leftarrow \mathbf{0}_{n \times d}$
for $i = 1$ **to** n **do**
  $j^* \leftarrow \arg\max_{1 \le j \le m} \mathcal{P}_{i,j}$
  $\mathcal{B}_{i,:} \leftarrow \widetilde{\mathcal{G}}_{j^*,:}$
Let $\mathcal{G}' \leftarrow \mathcal{G}(\mathcal{G}' \in \mathbb{R}^{n \times d})$
for $i = 1$ **to** n **do**
  for $k = 1$ **to** d **do**
    if $\phi_{i,k} \geq T$ **then**
      $\mathcal{G}'_{i,k} \leftarrow \mathcal{B}_{i,k}$
return $\mathcal{G}'$



Qualitative results

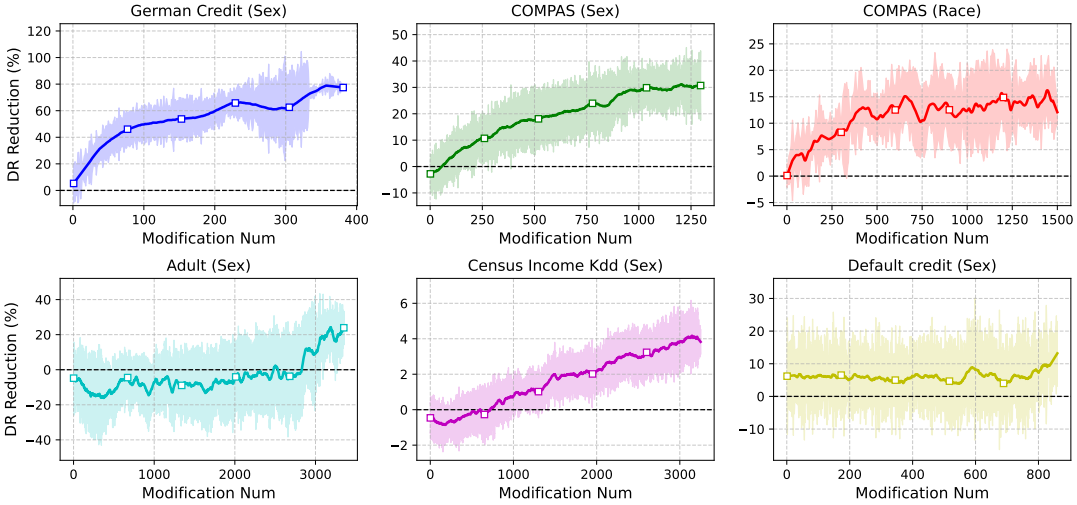


Figure 5. Percentage reduction in the DR across different datasets. The x-axis denotes the number of modifications applied (up to the maximum required under a fairness threshold $T = 0.05$), while the y-axis indicates the relative in DR, expressed as a percentage of the original value.

Qualitative results

Table 1. Compare FairSHAP with other fairness mitigation methods across different datasets.[15] Here, "s.a." denotes the sensitive attribute used in fairness evaluation. 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.

Dataset (s.a.)	Methods	Accuracy	DR	DP	EOpp	PP	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**

[15] "Kdd" is an abbreviation for the Census Income KDD dataset

Qualitative results

Table 1. Compare FairSHAP with other fairness mitigation methods (cont).

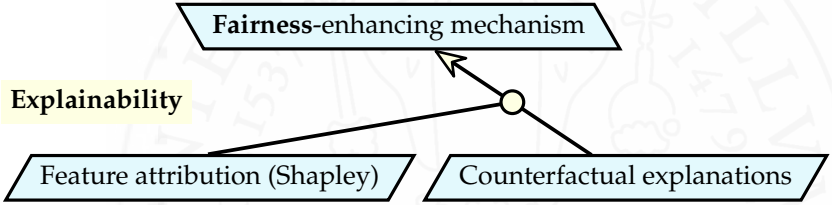
Dataset (s.a.)	Methods	Accuracy	DR	DP	EOpp	PP	Data Fidelity	TrainingAR	TestAN
Adult (sex)	Baseline	**0.8722**±0.0033	0.0315±0.0037	0.1805±0.0066	0.0735±0.0275	0.0275±0.0321	—	—	No
	RSA	0.8721±0.0024	—	0.1770±0.0066	0.0678±0.0277	**0.0272**±0.0252	—	—	Yes
	CR	0.8706±0.0029	**0.0000**±0.0000	0.1824±0.0055	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**

Qualitative results

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, EOpp, PP) are **bold**.

Dataset	Methods	Accuracy	DR	DP	EOpp	PP
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

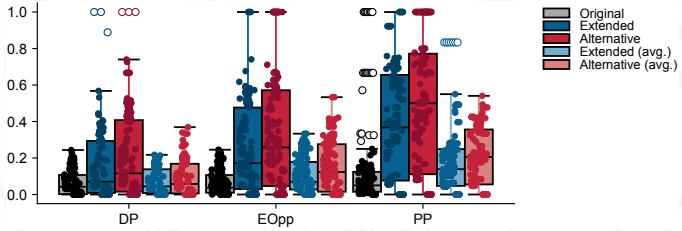
Takeaways



- Prioritising individual fairness can offer stronger and more flexible leverage than focusing only on group fairness
- Connecting explainability with fairness makes mitigation more interpretable, targeted, and practical

Additionally[15]

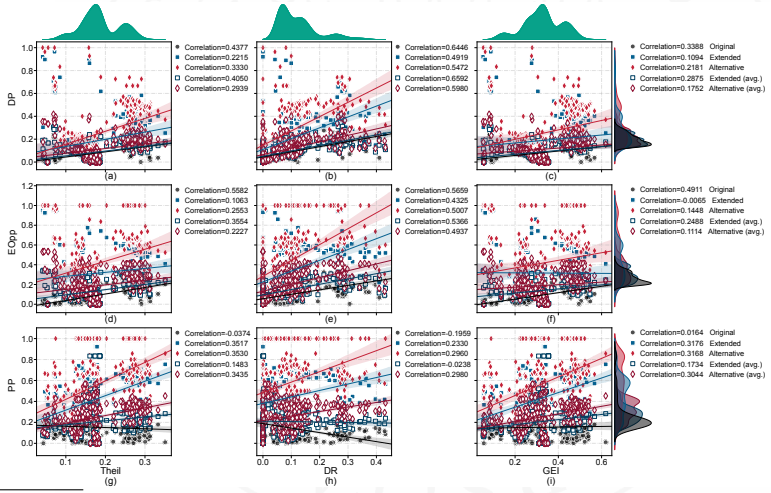
- Binarlisation systematically underestimates discrimination



[15] Yijun Bian et al. "Revisiting some misconceptions and limitations in algorithmic fairness". In: (2025).

Additionally[15]

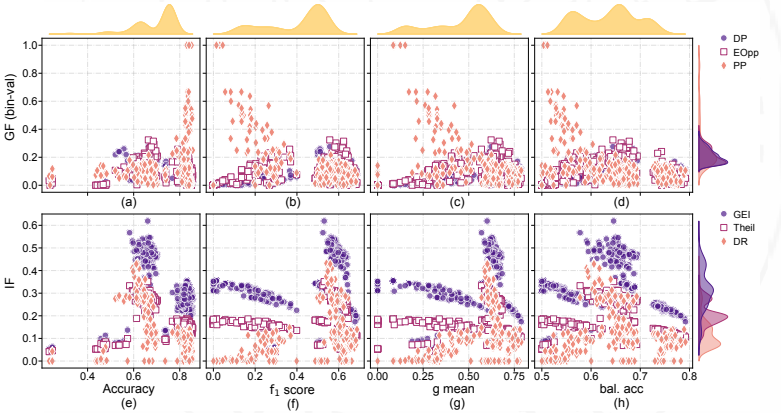
- Binarlisation systematically underestimates discrimination
- The myth of incompatibility between individual and group fairness



[15] Bian et al., see n. 15.

Additionally[15]

- Binarlisation systematically underestimates discrimination
- The myth of incompatibility between individual and group fairness
- The elusive trade-off between utility performance and fairness



[15] Bian et al., see n. 15.