HIFI: Explaining and Mitigating Algorithmic Bias through the Lens of Game-Theoretic Interactions

Lingfeng Zhang
East China Normal University
lanford217@gmail.com

Zhaohui Wang
East China Normal University
sternstund22@gmail.com

Yueling Zhang†
East China Normal University
ylzhang@sei.ecnu.edu.cn

Min Zhang†
East China Normal University
mzhang@sei.ecnu.edu.cn

Jiangtao Wang
East China Normal University
jtwang@sei.ecnu.edu.cn

***Abstract*—Machine Learning (ML) algorithms are increasingly used in decision-making process across various social-critical domains, but they often somewhat inherit and amplify bias from their training data, leading to unfair and unethical outcomes. This issue highlights the urgent need for effective methods to detect, explain, and mitigate bias to ensure the fairness of ML systems. Previous studies are prone to analyze the root causes of algorithmic bias from a statistical perspective. However, to the best of our knowledge, none of them has discussed how sensitive information inducing the final discriminatory decision is encoded by ML models. In this work, we attempt to explain and mitigate algorithmic bias from a game-theoretic view. We mathematically decode an essential and common component of sensitive information implicitly defined by various fairness metrics with Harsanyi interactions, and on this basis, we propose an in-processing method HIFI for bias mitigation. We conduct an extensive evaluation of HIFI with 11 state-of-the-art methods, 5 real-world datasets, 4 fairness criteria, and 5 ML performance metrics, while also considering intersectional fairness for multiple protected attributes. The results show that HIFI surpasses state-of-the-art in-processing methods in terms of fairness improvement and fairness-performance trade-off, and also achieves notable effectiveness in reducing violations of individual fairness simultaneously.**

***Index Terms*—algorithmic bias, fairness, bias mitigation, game-theoretic interaction, explainable artificial intelligence**

I. Introduction

Machine Learning (ML) models, as the enablers of Artificial Intelligence (AI) systems, are applied in social-critical scenarios, such as job recommendations [1], hiring [2, 3], and criminal justice [4]. However, ML models are vulnerable to unwanted algorithmic bias, which can harm human rights [5] and bring software engineers additional legal risk [6, 7].

In the context of decision-making, fairness is the absence of any prejudice or favoritism toward an individual or group based on their inherent or acquired characteristics [5]. A variety of metrics have been designed and justified for fairness evaluation, and we will present some representative definitions in II-A. Most of existing bias mitigation methods focus on group fairness. Pre-processing methods posit that bias originates from discriminatory data [8, 9, 10], in-processing methods mitigate bias during the training procedure [11, 12, 13, 14], and post-processing methods calibrate model predictions to ensure fair outcomes [15, 16, 17]. For individual fairness improvement, the discriminatory samples generated by fairness testing methods are augmented to the training data for retraining or adversarial training [18, 19, 20, 21, 22, 23].

Nevertheless, there are trade-offs between model utilities and fairness [7, 24, 25, 26, 27, 28], between different group fairness metrics [28, 29, 30], and between group fairness and individual fairness [28, 31, 32, 33, 34]. Moreover, it is widely recognized that fairness through blindness (i.e., simply ignoring protected attributes or preventing models from explicitly encoding information about protected attributes) not only fails to effectively reduce discrimination due to redundant encoding, but also significantly harms model performance [31, 35]. Hence, we speculate that the representations encoded by ML models for making decisions are complex and blended, protected attributes may contain task-related information, and non-sensitive attributes might also conceal sensitive information. In this work, we divide the encoded information used for inference into **sensitive information** that causes algorithmic bias, and **task-related information** that is totally and reasonably relevant to the classification task.

The existing debiasing methods have not thoroughly investigated how ML models encode sensitive information and how to separate it from task-related information. To figure out these questions, we attempt to explain ML bias from a game-theoretic perspective. Harsanyi interaction [36] decomposes the output of a model into contributions of AND relationships between input variables, and the utility of an interaction pattern is equally assigned to the involved variables in the computation of Shapley values [36, 37, 38]. Harsanyi interaction has been utilized to define, extract, and quantify symbolic concepts encoded by Deep Neural Networks (DNNs) [36, 39, 40, 41, 42]. We believe that such interactive concepts, as a general explainability framework with a relatively comprehensive theoretical foundation, should also be effective for elucidating the root causes of ML bias. Therefore, in Sec. III, we decompose the most popular group fairness metrics into interactions to analyze which kinds of AND relationships encode most of the sensitive information for inference, and we further suppress the parts that are natural to separate from task-related information to improve model fairness without metric conflicts. We call our in-processing bias mitigation method *Harsanyi Interaction-*

†Corresponding authors.

1558-1225/25/$31.00 ©2025 IEEE
DOI 10.1109/ICSE55347.2025.00221

based Fairness Improvement (HIFI), and HIFI also implies HIgh FIdelity in terms of model utilities. Note that compared to pre-processing methods, previous in-processing methods tend to fall into the poorer performance-fairness trade-off level [27]. However, our method has partially changed this situation.

Overall, the main contributions of our paper are:

- We extract a large portion of sensitive information encoded by ML models, and completely separate it from the task-related information, which boosts the understanding and explanation of algorithmic bias.
- We suppress the sensitive information in an in-processing fashion without substantially damaging the task-related information, and we have implemented our method HIFI with GPU acceleration support[1]. HIFI can be applied to any model that can be easily trained using gradient-based optimization methods, if sensitive attributes are specified.
- We comprehensively compare HIFI with 11 state-of-the-art debiasing methods that support intersectional fairness scenarios on 5 real-world datasets. Our experimental results indicate that HIFI outperforms other in-processing methods in enhancing group fairness, balancing the performance-fairness trade-off, and resolving conflicts between individual and group fairness.
- We also find that HIFI can be combined with other distinguished methods like REW [8] , MAAT [43], and FairMask [44] to further enhance debiasing effectiveness.

II. Preliminaries

A. Algorithmic Bias

Fairness, as a non-functional property, is gaining increasing attention from both the ML and Software Engineering (SE) communities. During the classification, certain personal attributes need to be protected against discrimination. These attributes are referred to as protected attributes, also known as sensitive attributes. Common protected attributes include sex, race, age, religion, disability status, and national origin [27].

Group fairness usually focuses on the statistical parity of different demographic groups. In this paper, we adopt the extended definitions of the most typical metrics, i.e., *Statistical Parity Difference* (SPD) [45], *Average Odds Difference* (AOD) [16], and *Equal Opportunity Difference* (EOD) [16], from the empirical study [27]. The mathematical definitions of them are shown in Table I, where A represents the set of protected attribute(s), S is the set of value combinations of the attribute(s) in A, Y denotes the groud truth label, $\hat{Y}$ denotes the predicted label, 1 is the favorable class, and 0 is the unfavorable class. Such definitions are generalizable to single or multiple sensitive attributes.

Individual fairness ensures that similar individuals are treated similarly [31]. The most widely adopted metric in fairness testing is causal fairness, and it requires a model to produce the same outcomes for every two individuals who differ only in sensitive attributes [46]. Formally, for any given input x and any two valuations of the sensitive attribute(s) $s, s' \in S$: $\hat{Y}(x, s) = \hat{Y}(x, s')$. We further define the ratio of violations of causal fairness on the testing data as *Causal Fairness Violation Ratio* (CFVR). For all the above metrics, the larger the value, the higher the degree of discrimination.

[1]https://github.com/LingfengZhang98/HIFI

Metric	Definition
SPD	$\max_{s\in S} P[\hat{Y}=1 \mid A=s] - \min_{s\in S} P[\hat{Y}=1 \mid A=s]$
AOD	$\frac{1}{2}\Big[\begin{array}{l}\max_{s\in S}(P[\hat{Y}=1 \mid A=s,Y=0]+P[\hat{Y}=1 \mid A=s,Y=1]) \\ -\min_{s\in S}(P[\hat{Y}=1 \mid A=s,Y=0]+P[\hat{Y}=1 \mid A=s,Y=1])\end{array}\Big]$
EOD	$\max_{s\in S} P[\hat{Y}=1 \mid A=s,Y=1] - \min_{s\in S} P[\hat{Y}=1 \mid A=s,Y=1]$

TABLE I: Representative group fairness metrics.

B. Harsanyi Interaction

Given an input sample x and a classification model $v(\cdot)$ to predict the probability that an input is classified as a favorable label, let us first define the interaction concepts that emerge in the model inference. Given the sentence '*sit down and take it easy*', the co-appearance of a set of words $C = \{$*take, it, easy*$\}$ forms a coalition and causes the meaning of '*calm down*', contributing significantly to the output. Such a combination of words is termed an interactive concept. Each interactive concept of C represents an AND relationship between the set of words in C, and only their co-appearance will contribute an additional effect I_C to the model output [36, 38].

Given any input sample, all the interactive concepts can be organized into a three-layer causal graph like Fig. 1. Each source node in the bottom layer represents the binary state of whether the variable is masked or not. Each intermediate node I_C represents the causal effect of the AND relationship between variables in C. If such a causal pattern is activated, i.e., any variable involved in the pattern is not masked, we denote $ACT_C = 1$; otherwise, $ACT_C = 0$. We denote the set of all input variables as N, and $\Omega = 2^N = \{C : C \subseteq N\}$ includes all the $2^{|N|}$ coalitions. Therefore, the output Y_{CG} of the causal graph can be specified by a Structural Causal Model (SCM) [47], which sums up all triggered causal effects, i.e., $Y_{CG}(x) = \sum_{C\in\Omega} I_C \cdot ACT_C$. The previous work [36] proved that if the causal effect I_C is measured by the *Harsanyi dividend* [37], i.e.,

$$I_C \overset{\text{def}}{=} \sum_{C'\subseteq C} (-1)^{|C|-|C'|} \cdot v(x_{C'}), \tag{1}$$

then the causal graph faithfully encodes the inference logic of the model, as $\forall C \subseteq N, Y_{CG}(x_C) = v(x_C)$, where x_C is implemented by masking all variables in $N\backslash C$ with baseline values [39, 48, 49] on the original input x.

In this work, we refer to the Harsanyi dividend used for explainability research as *Harsanyi interaction*, as game-theoretic interaction studies did. Besides faithfulness, the explainability framework based on Harsanyi interaction also satisfies conciseness (i.e., only a small number of causal patterns are salient) and universality (i.e., various well-trained models on different tasks can all be faithfully and concisely explained by a few salient causal patterns) [36, 39, 40, 41, 42].

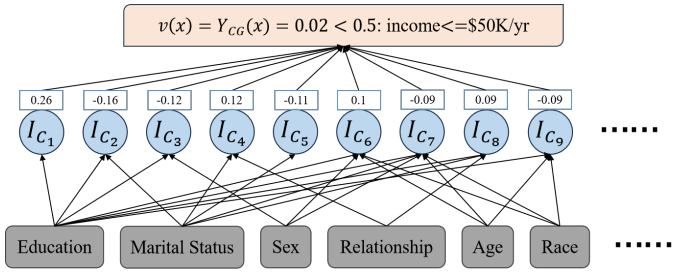


Fig. 1: A real example of causal graph that explains the inference on a sample from Census Income.

C. Related Work

Bias mitigation. We introduce the debiasing methods used in our experiments, and you can refer to the surveys [5, 50] for more on ML fairness. Pre-processing methods include **REW** (Reweighting) [8], **DIR** (Disparate Impact Remover) [9], and **Fair-Smote** [10]. REW adjusts example weights, DIR modifies feature values while preserving rank-ordering, and Fair-Smote generates synthetic samples for balanced distributions. In-processing methods incorporate **META** (Meta Fair Classifier) [13], **ADV** (Adversarial Debiasing) [12], and **PR** (Prejudice Remover) [11]. META designs a meta-algorithm to optimize the preset fairness metric, ADV reduces adversary ability to predict protected attributes, and PR adds a discrimination-aware term. Post-processing methods encompass **EOP** (Equalized Odds Processing) [16], **CEO** (Calibrated Equalized Odds) [17], and **ROC** (Reject Option Classification) [15]. EOP and CEO optimize AOD by modifying output label probabilities, while ROC ensures favorable outcomes for unprivileged groups near decision boundaries. In addition, **MAAT** [43] combines pre-processing and in-processing methods with an ensemble strategy, and **FairMask** [44] applies after training to modify protected attributes in test data..

Explaining root causes of bias. The PR work [11] identifies how sensitive features influence decision-making (prejudice), inadequate training leads to model inaccuracies (underestimation), and biases in training data propagate through models (negative legacy). The Fair-Smote work [10] attributes bias to data imbalance and improper labeling. However, no work has analyzed how sensitive information inducing the final discriminatory decision is encoded by ML models. We will bridge the gap in this work.

Game-theoretic interaction. The system of game-theoretic interactions for Explainable AI (XAI) has been gradually built up. This system focuses on addressing the following problems in XAI: (1) defining, extracting, and quantifying interactive concepts encoded by DNNs [36, 39, 40, 41, 42, 51, 52, 53, 54, 55], (2) using interactive concepts to explain or measure the representation power of DNNs [56, 57, 58, 59, 60, 61], and (3) unifying the common underlying mechanism shared by previous empirical findings [62, 63, 64, 65].

We are the first to explain, quantify, and eliminate discrimination through the lens of game-theoretic interactions.

III. Approach

In this section, we first decompose the group fairness metrics defined in Sec. II-A based on Harsanyi interactions, and empirically study the impacts of different types of interactions on group fairness. Then, we extract the interactions that encode most of the sensitive information and completely separate them from the task-related information. This allows us to directly penalize the strength of these interactions during training to mitigate bias while maintaining the model utilities.

A. Decomposing Group Fairness Criteria

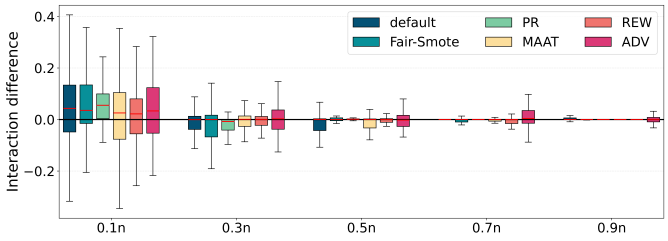
In the real example shown in Fig. 1, the input sample is from a young and well-educated Black woman, and the causal patterns shown in the figure are the most salient ones. As explained by Harsanyi interactions, her good education leads the model to tend towards giving a favorable prediction (I_{C_1}), but when the model considers both her education and gender together, it tends to give a unfavorable treatment (I_{C_3}). Such algorithmic bias is quite common. To trace the root causes of model discrimination, according to the *efficiency* property of Harsanyi interactions [36], i.e., $v(x) = \sum_{C\subseteq N} I_C(x)$, we decompose the group fairness metrics into Harsanyi interactions:

$$\begin{aligned} SPD &= \mathbb{E}_{x\sim P_p} v(x) - \mathbb{E}_{x\sim P_{up}} v(x) \\ &= \mathbb{E}_{x\sim P_p} \sum_{C\subseteq N} I_C(x) - \mathbb{E}_{x\sim P_{up}} \sum_{C\subseteq N} I_C(x) \\ &= \sum_{C\subseteq N} \left(\mathbb{E}_{x\sim P_p} I_C(x) - \mathbb{E}_{x\sim P_{up}} I_C(x)\right), \end{aligned} \tag{2}$$

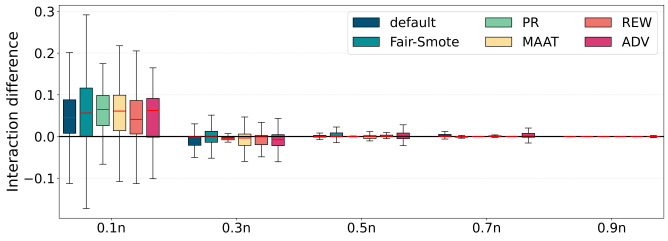
$$\begin{aligned} AOD =& \frac{1}{2}\left(\left(\mathbb{E}_{x\sim P_{p\&f}} v(x) + \mathbb{E}_{x\sim P_{p\&uf}} v(x)\right)\right. \\ & \left. - \left(\mathbb{E}_{x\sim P_{up\&f}} v(x) + \mathbb{E}_{x\sim P_{up\&uf}} v(x)\right)\right) \\ =& \frac{1}{2}\left(\left(\mathbb{E}_{x\sim P_{p\&f}} \sum_{C\subseteq N} I_C(x) + \mathbb{E}_{x\sim P_{p\&uf}} \sum_{C\subseteq N} I_C(x)\right)\right. \\ & \left. - \left(\mathbb{E}_{x\sim P_{up\&f}} \sum_{C\subseteq N} I_C(x) + \mathbb{E}_{x\sim P_{up\&uf}} \sum_{C\subseteq N} I_C(x)\right)\right) \\ =& \frac{1}{2} \sum_{C\subseteq N} \left(\left(\mathbb{E}_{x\sim P_{p\&f}} I_C(x) + \mathbb{E}_{x\sim P_{p\&uf}} I_C(x)\right)\right. \\ & \left. - \left(\mathbb{E}_{x\sim P_{up\&f}} I_C(x) + \mathbb{E}_{x\sim P_{up\&uf}} I_C(x)\right)\right), \end{aligned} \tag{3}$$

$$\begin{aligned} EOD &= \mathbb{E}_{x\sim P_{p\&f}} v(x) - \mathbb{E}_{x\sim P_{up\&f}} v(x) \\ &= \mathbb{E}_{x\sim P_{p\&f}} \sum_{C\subseteq N} I_C(x) - \mathbb{E}_{x\sim P_{up\&f}} \sum_{C\subseteq N} I_C(x) \\ &= \sum_{C\subseteq N} \left(\mathbb{E}_{x\sim P_{p\&f}} I_C(x) - \mathbb{E}_{x\sim P_{up\&f}} I_C(x)\right), \end{aligned} \tag{4}$$

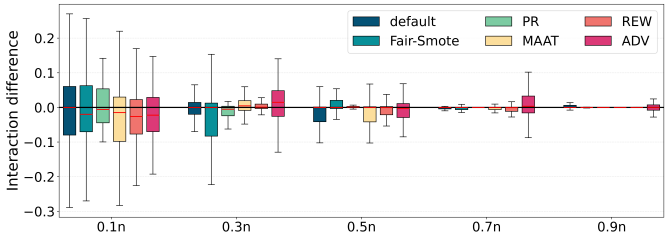
where P_p represents the distribution of the privileged group, i.e., the demographic group getting the maximum favorable class ratio in SPD for example, P_{up} represents the distribution of the unprivileged group, P_f represents the distribution of the group with favorable ground truth label, P_{uf} represents the distribution of the group with unfavorable ground truth



(a) Results for all coalitions $C \subseteq N$.



(b) Results for coalitions in $\{C : C \subseteq N \wedge C \cap A = \emptyset\}$.



(c) Results for coalitions in $\{C : C \subseteq N \wedge C \cap A \neq \emptyset\}$.

Fig. 2: Order-wise distribution of the inter-group interaction differences $\Delta I^{AOD}(C)$. Note that the overall situation is quite similar for $\Delta I^{SPD}(C)$ and $\Delta I^{EOD}(C)$ as well. Here $m = \{0.1n, 0.3n, 0.5n, 0.7n, 0.9n\}$ corresponds to the intervals $\{[0, 0.2n), [0.2n, 0.4n), [0.4n, 0.6n), [0.6n, 0.8n), [0.8n, n]\}$. For each method, each (dataset, classifier, random seed) combination is considered as a scenario. The sum of $\Delta I^{AOD}(C)$ in each order interval in each scenario is used to draw the box plots.

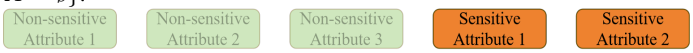
label, and the group-class pairs $P_{p\&f}$, $P_{p\&uf}$, $P_{up\&f}$ and $P_{up\&uf}$ also follow in this manner. Then we define *Inter-Group Interaction Difference* (**IGID**) $\Delta I^{SPD}(C)$, $\Delta I^{AOD}(C)$ and $\Delta I^{EOD}(C)$ as follows:

$$\Delta I^{SPD}(C) \overset{\text{def}}{=} \mathbb{E}_{x \sim P_p} I_C(x) - \mathbb{E}_{x \sim P_{up}} I_C(x), \tag{5}$$

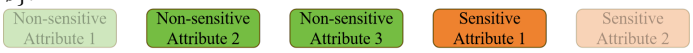
$$\begin{aligned}\Delta I^{AOD}(C) \overset{\text{def}}{=} \frac{1}{2} \big(&\left(\mathbb{E}_{x \sim P_{p\&f}} I_C(x) + \mathbb{E}_{x \sim P_{p\&uf}} I_C(x)\right) \\ &- \left(\mathbb{E}_{x \sim P_{up\&f}} I_C(x) + \mathbb{E}_{x \sim P_{up\&uf}} I_C(x)\right) \big),\end{aligned} \tag{6}$$

$$\Delta I^{EOD}(C) \overset{\text{def}}{=} \mathbb{E}_{x \sim P_{p\&f}} I_C(x) - \mathbb{E}_{x \sim P_{up\&f}} I_C(x). \tag{7}$$

(a) Coalitions NOT involving sensitive attributes: $\{C : C \subseteq N \wedge C \cap A = \emptyset\}$.



(b) Coalitions ONLY involving sensitive attributes: $\{C : C \subseteq A \wedge C \neq \emptyset\}$.



(c) Coalitions containing BOTH non-sensitive attributes and sensitive attributes: $\{C : C \subseteq N \wedge C \cap A \neq \emptyset \wedge C \cap (N - A) \neq \emptyset\}$.

Fig. 3: Examples of co-appearance of various coalitions.

	Spearman			Kendall's Tau		
Part	SPD	AOD	EOD	SPD	AOD	EOD
IGID w/ SA	**0.817***	**0.820***	**0.775***	**0.643***	**0.651***	**0.609***
IGID w/o SA	0.631*	0.534*	0.578*	0.486*	0.398*	0.434*

TABLE II: Correlation analysis between IGIDs with sensitive attributes or IGIDs without sensitive attributes and all IGIDs. * indicates a significant correlation with p-value < 0.05.

B. Extracting Sensitive Information

Furthermore, we divide interactions and differences by order, where the order m of the interaction I_C is defined as the cardinality of the coalition C, and $n = |N|$ is the maximum order. Based on the experimental setup described in Sec. IV-B, we select the default training method and 5 debiasing methods (Fair-Smote [10], PR [11], MAAT [43], REW [8] and ADV [12]) that (1) have reasonable bias mitigation effects, and (2) meet the conditions for the emergence of sparse symbolic concepts [41], and exclude the Default Credit dataset with 23 features due to significant computational costs to compute all 2^{23} interactions. We will comprehensively evaluate these methods in Sec. V. Subsequently, we explore the order-wise distribution of $\Delta I^{AOD}(C)$ in Fig. 2, and the results for $\Delta I^{SPD}(C)$ and $\Delta I^{EOD}(C)$ can be found in our artifacts. All results are approximated on 500 (100 for the Diabetes dataset) inputs randomly sampled from the testing data for each of the 10 random seeds. According to the definitions in Eq. 3 and Eq. 6, the closer the distribution of $\Delta I^{AOD}(C)$ of each order interval is to 0, the lower the absolute value of AOD and the fairer the final result. It is shown that (1) low-order IGIDs contribute the main part of discrimination for well-trained models, and (2) IGIDs involving sensitive attributes (i.e., the part represented by Fig. 3b plus Fig. 3c) have a greater impact on the final fairness compared to IGIDs that do not involve sensitive attributes (i.e., the part represented by Fig. 3a), since the trends in Fig. 2c and Fig. 2a appear more similar compared to Fig. 2b and Fig. 2a, especially when observing the whiskers of the box plots.

To validate the second observation, we calculate the Spearman correlation coefficients [66] and the Kendall's Tau correlation coefficients [67] between the sum of IGIDs involving sensitive attributes and the sum of all IGIDs, as well as between the sum of IGIDs without sensitive attributes and

Metric	Part	default	Fair-Smote	PR	MAAT	REW	ADV
SPD	IOISA	0.097 (0.065)	0.099 (0.064)	0.049 (0.039)	0.094 (0.047)	0.074 (0.044)	0.102 (0.070)
	others	0.051 (0.002)	0.122 (0.002)	0.023 (0.006)	0.042 (0.002)	0.048 (0.001)	0.044 (0.016)
AOD	IOISA	0.097 (0.064)	0.099 (0.064)	0.049 (0.039)	0.093 (0.047)	0.074 (0.044)	0.101 (0.070)
	others	0.053 (0.001)	0.123 (0.002)	0.026 (0.015)	0.043 (0.003)	0.048 (0.001)	0.035 (0.015)
EOD	IOISA	0.096 (0.064)	0.099 (0.063)	0.049 (0.039)	0.093 (0.047)	0.074 (0.044)	0.102 (0.070)
	others	0.059 (0.001)	0.133 (0.002)	0.040 (0.045)	0.044 (0.003)	0.050 (0.000)	0.026 (0.005)

TABLE III: Division of low-order IGIDs involving sensitive attributes into the IGIDs only involving sensitive attributes (denoted as IOISA here) and others. The values in the table are the **average (median)** of the absolute values of the sum of each type of IGID in each scenario.

the sum of all IGIDs, in the low-order interval $[0, 0.2n)$ in all scenarios. In Table II, IGIDs involving sensitive attributes are always more strongly correlated with aggregated IGIDs compared to IGIDs not involving sensitive attributes. Also, the average standard deviation of the former is 1.82 times that of the latter, indicating greater volatility. Therefore, we consider that IGIDs involving sensitive attributes have a greater impact on fairness than IGIDs not involving sensitive attributes. However, IGIDs without sensitive attributes are still positively correlated with all IGIDs, which suggests to some extent that a small portion of sensitive information may be revealed through the interactions of some non-sensitive attributes. In other words, the co-appearance of multiple non-sensitive attributes can implicitly point out the sensitive attributes.

Let's continue to delve into the main components affecting fairness, specifically the low-order IGIDs that involve sensitive attributes, which can be further divided into IGIDs **only** involving sensitive attributes (i.e., the part represented by Fig. 3b) and others (i.e., the part represented by Fig. 3c). IGIDs only involving sensitive attributes are inherently related to sensitive information and completely separate from task-related information, because an interaction on a coalition of sensitive attributes only calculates the additional effect brought by their co-appearance without considering other attributes. In Table III, for all types of IGIDs, the values of IGIDs only involving sensitive attributes are always greater than those of others for the default training method, even though the former has a very small proportion in terms of the number of interactions. Most existing debiasing methods suppress both parts to some extent, but the '**free lunch**' offered by IGIDs only involving sensitive attributes has not been fully utilized. For Fair-Smote and ADV, there are cases where the statistical values of some parts of IGIDs increase compared to the default training method, because these two methods can sometimes make fairness significantly worse as we will thoroughly evaluate in Sec. V, and the extreme values finally distort the statistical results.

At this point, we recognize that setting aside other sensitive information that is difficult to separate from task-related information in terms of model representations, we can still achieve a reasonable debiasing effect without significantly compromising the model utilities by directly suppressing the intrinsically sensitive information represented by the IGIDs only involving sensitive attributes, because this part accounts for a significant proportion of the sensitive information and is inherently separate from the task-related information.

C. Harsanyi Interaction-Based Fairness Improvement

Based on the above understanding, we develop the fairness loss as an improvement of model fairness, and apply this loss to the general training process, as follows.

$$\text{Loss} = \text{Loss}_{\text{classification}} + \eta * \text{Loss}_{\text{fairness}}, \tag{8}$$

where $\eta > 0$ is the weight of the fairness loss. To suppress the IGIDs only involving sensitive attributes, a simple method is to punish the interaction strength of coalitions only containing sensitive attributes across the entire training set, without considering which group the input belongs to. This will also bring an additional benefit of suppressing the model's explicit encoding of the intrinsically sensitive information, leading to better results on various fairness metrics simultaneously, without the need to predefine specific ones. In V, we will demonstrate the excellent performance of our method in improving causal fairness while mitigating group discrimination. Thus, the fairness loss on a training batch B can be formulated as

$$\text{Loss}_{\text{fairness}} = \frac{1}{|B|} \sum_{x \in B} \sum_{C \subseteq A, C \neq \emptyset} |I_C(x)|. \tag{9}$$

We name our in-processing debiasing method ***Harsanyi Interaction-based Fairness Improvement*** **(HIFI)**. HIFI enables us to optimize fairness metrics of different categories while maintaining model utilities. HIFI can be applied to any model that can be trained using gradient-based optimization methods, if sensitive attributes are explicitly defined.

Example. Given a tabular dataset with 5 attributes like Fig. 3, if we want to optimize the intersectional fairness of the subject model with HIFI in terms of Sensitive Attribute 1 (SA1) and Sensitive Attribute 2 (SA2), the coalition C in Eq. 9 can be [SA1], [SA2] and [SA1, SA2].

IV. Evaluation

In this section, we illustrate how our method HIFI is comprehensively evaluated, and partially reference the evaluation framework of the latest and most comprehensive empirical study [27].

A. Research Questions

RQ1: How does our method HIFI perform in improving model fairness? This RQ evaluates the effectiveness of HIFI in improving group fairness when compared to state-of-the-art bias mitigation methods.

RQ2: What fairness-performance trade-off does our method HIFI achieve? This RQ evaluates the effectiveness of HIFI in fairness-performance trade-off when compared to state-of-the-art bias mitigation methods.

Name	#Samples	#Features	Protected attributes	Task	Favorable label (Proportion)	URL
Census Income	30162	12	race, sex	Income prediction	income>50K yr (24.9%)	[68]
UFRGS	43303	11	gender, race	GPA prediction	GPA≥3 (46.8%)	[69]
COMPAS	7215	7	sex, race	Re-offence prediction	no recidivism (54.9%)	[70]
Diabetes	769	8	age	Readmission prediction	no diabetes (65.1%)	[71]
Default Credit	30000	23	sex, age	Credit risk prediction	default (22.1%)	[72]

TABLE IV: Datasets.

η	0.001	0.01	0.1	0.25	0.5	0.75	1	2	5	10	100	1000
Performance	+0.04%	-0.4%	-1.18%	-1.06%	-0.02%	-0.38%	-1.6%	-2.64%	-4.28%	-6.02%	-17.8%	-22.82%
Discrimination	+1.17%	-2.67%	-16.6%	-30.37%	-32.4%	-37.3%	-36.87%	-41.67%	-45.23%	-48.9%	-62.57%	-64.47%

TABLE V: The respective average of the relative change rate of performance and fairness metrics for different η.

RQ3: How well does our method HIFI apply to different decision tasks and ML models? This RQ provides a detailed supplement to RQ1 and RQ2, and explores whether HIFI is widely applicable.

RQ4: How does individual fairness change when our method HIFI improves group fairness? This RQ examines the ability of our method to simultaneously improve different types of fairness and explores the impact of the strength of interactions only involving sensitive attributes on the model's individual fairness.

B. Experimental Setup

Datasets and models. We consider 5 real-world tabular datasets [2] that are widely used in the literature of algorithmic fairness [50, 73]. These datasets are described in Table IV. For each dataset, we train 4 ML models, including Logistic Regression (**LR**), Support Vector Machine (**SVM**), Neural Network (**NN**), and Random Forest (**RF**). LR, SVM, and RF use default configurations, while the NN employs a fully connected architecture with 5 hidden layers, containing 64, 32, 16, 8, and 4 units, respectively, as the study [27] did [3].

Baseline debiasing methods. We employ 11 state-of-the-art bias mitigation methods summarized in [27] for comparisons, of which 8 are from the ML community and have been integrated into the IBM AIF360 toolkit [74], and 3 are from the SE community. We have briefly introduced them in Sec. II-C. We keep the configurations used in [27] for these methods. Note that the three in-processing methods implemented by AIF360 [74], including META, ADV and PR, are only applicable to the LR model, but our method HIFI can be used for LR, SVM and NN.

Evaluation metrics. We use 4 fairness metrics introduced in Sec. II-A, including **SPD**, **AOD**, and **EOD** for group fairness, and **CFVR** for individual fairness. We also use 5 common ML performance metrics for evaluation: **accuracy**, **precision**, **recall**, **F1-score**, and **MCC** (Matthews Correlation Coefficient) [75], as previous work [26, 27, 43] did. For all the 5 metrics, larger values indicate better ML performance. To assess the fairness-performance trade-off, we employ the benchmark set up by **Fairea** [7], which establishes the trade-off baseline by connecting fairness-performance points of the original model and a set of models mutated to varying degrees. Fairea classifies the trade-off effectiveness into 5 levels by comparing the trade-off achieved by the subject methods with the baseline: the **win-win** trade-off level increases both fairness and performance; the **good** trade-off level reasonably trades performance for fairness; the **inverted** trade-off level trades fairness for performance; the **poor** trade-off level improves fairness but sacrifices performance too much; and the **lose-lose** trade-off decreases both fairness and performance. For each combination of (dataset, ML model, coalition of sensitive attribute(s), group fairness metric, performance metric), we establish a trade-off baseline.

Statistical analysis. We adopt 4 non-parametric statistical analysis methods: **Mann Whitney U-test** [76], **Cliff's δ** [77], **Spearman correlation coefficient** [66], and **Kendall's Tau correlation coefficient** [67]. In RQ1, RQ2 and RQ3, we use the Mann Whitney U-test and Cliff's δ to assess whether fairness improvement methods significantly impact fairness or performance. To confirm statistical significance, we follow the empirical study [27] to consider a p-value of Mann Whitney U-test lower than 0.05 and an absolute value of δ greater than or equal to 0.428. In RQ4, we use the Spearman correlation coefficient and the Kendall's Tau correlation coefficient to explore the relationship between the strength of interactions only involving sensitive attributes and CFVR. These two coefficients ranges from -1 to 1, with 1 representing a perfect positive correlation. A correlation is deemed statistically significant only when the coefficient yield a p-value lower than 0.05.

Configurations. The fairness loss term enables us to explicitly control the penalty of intrinsically sensitive information by adjusting the weight η in Eq. 8. The weight η needs to be carefully set, since a small value of η does not make much difference, and a large value of η may lead to under-fitting. We calculate the average performance and fairness for different values of η in Table V. It is shown that as η increases, the debiasing effect of HIFI enhances, and the model performance deteriorates. However, at $\eta = 0.5$ or $\eta = 0.75$, a good fairness-performance trade-off is achieved. We will set $\eta = 0.75$ for HIFI in the experiments. Additionally, the baseline value of

[2]We have also considered other datasets like Bank Marketing and Heart Health. However, the former shows fair performance on original models (average fairness metric 0.011), while the latter is too small, making some methods like Fair-Smote with K-Nearest Neighbor inapplicable.

[3]In this paper, we refer to models trained with default methods as 'default' in the charts, and this should not be confused with the Default Credit dataset.

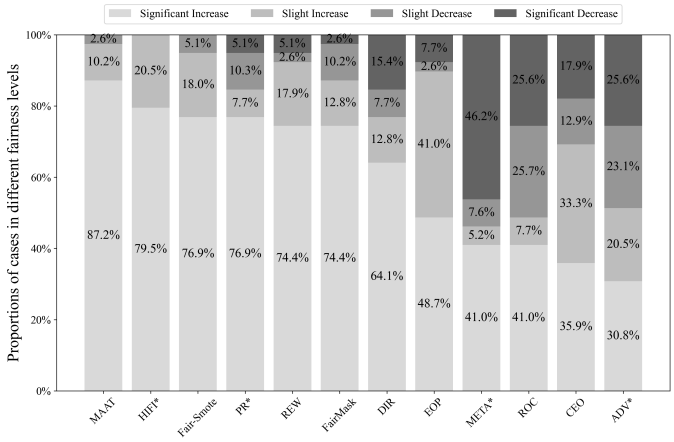


Fig. 4: (RQ1) Fairness level distributions for the LR model. * indicates in-processing methods.

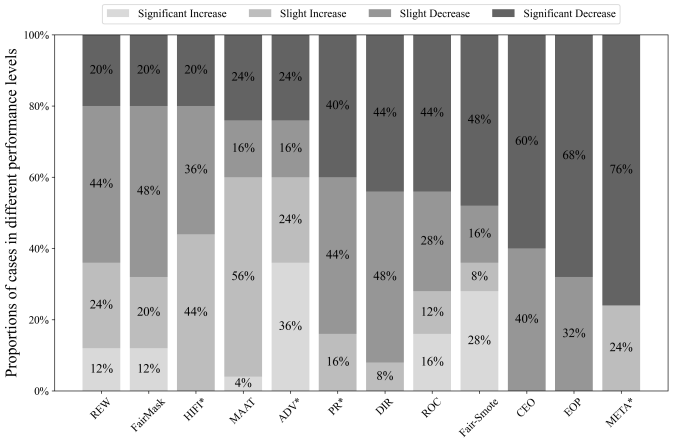


Fig. 5: (RQ2) Performance level distributions for the LR model. * indicates in-processing methods.

each input variable is set to the mean value of this variable over all samples [48] when computing Eq. 1.

All expeirments were run on a laptop with 64 GB RAM, AMD Ryzen 9 7945HX CPU and NVIDIA RTX 4060 GPU for 10 rounds with different random seeds [4].

V. Results

This section answers RQs based on the experimental results. Note that if the subject datasets have multiple sensitive attributes, we optimize fairness on them simultaneously to achieve intersectional fairness [27], and the results of group fairness evaluation are averaged or calculated on all non-empty subsets of the set of considered sensitive attributes. We answer RQ1 and RQ2 with experiments on the LR model, and extend the evaluation to other models when answering RQ3. RQ4 considers all the subject models.

A. RQ1: Fairness Improvement

This RQ investigates whether HIFI effectively suppresses the sensitive information encoded by Harsanyi interactions only involving sensitive attributes, so that the model fairness is significantly improved.

We consider each combination of (dataset, classifier, coalition of sensitive attribute(s), group fairness metric, random seed) as a scenario, and explore the proportions of scenarios in which the subject methods significantly increase fairness, slightly increase fairness, slightly decrease fairness, and significantly decrease fairness, respectively. Fig. 4 shows the fairness level distributions for the LR model, where the methods are sorted by descending order in the proportion of cases significantly increasing fairness. For the LR model, HIFI improves fairness in all scenarios, with 79.5% of these improvements being significant, outperforming the best existing in-processing method PR. The other two in-processing methods, META and ADV even significantly decrease fairness in 46.2% and 25.6% of scenarios, respectively.

[4]To show the competency of the original classifiers, we report the mean metric values of them: accuracy (0.748), recall (0.680), precision (0.724), f1-score (0.687), MCC (0.397), false alarm rate (0.214), SPD (0.220), AOD (0.180), and EOD (0.190).

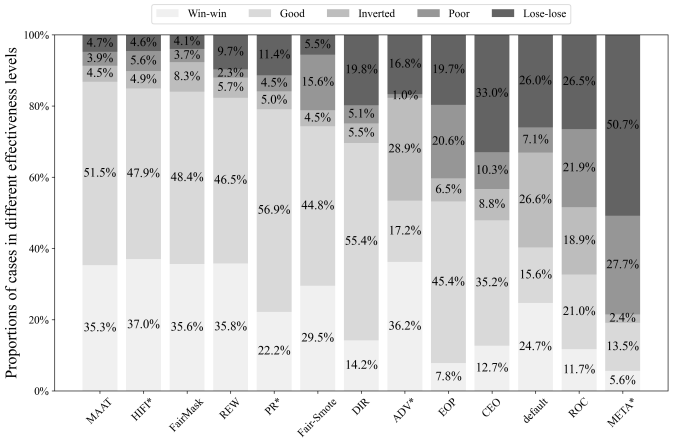


Fig. 6: (RQ2) Fairea effectiveness level distributions for the LR model. * indicates in-processing methods.

To gain an intuitive understanding of the extent of changes in fairness after applying debiasing methods, we calculate the average absolute values and relative changes of group fairness metrics for the LR model in Table VI. HIFI achieves top-3 bias mitigation effectiveness across all metrics. For the LR model, HIFI reduces discrimination by 40.3%, 52.8%, and 55% in the SPD, AOD, and EOD metrics, respectively, significantly better than the existing in-processing methods.

Ans. to RQ1: HIFI achieves the best effectiveness in fairness improvement for the LR model among in-processing methods. The results validate the reliability of the sensitive information we extracted.

Method	Accuracy ↑	Recall ↑	Precision ↑	F1 Score ↑	MCC ↑	SPD ↓	AOD ↓	EOD ↓
default	0.749	0.673	0.731	0.683	0.394	0.232	0.200	0.207
REW	0.749 (-0.1%)	0.669 (-0.6%)	0.733 (+0.3%)	0.677 (-0.8%)	0.391 (-0.8%)	0.175 (-24.8%)	0.138 (-31.1%)	0.140 (-32.3%)
DIR	0.739 (-1.4%)	0.656 (-2.6%)	0.720 (-1.4%)	0.657 (-3.8%)	0.357 (-9.4%)	0.117 (-49.6%)	0.094 (-52.8%)	0.124 (-40.3%)
Fair-Smote	0.720 (-4.0%)	0.696 (+3.4%)	0.685 (-6.3%)	0.688 (+0.7%)	0.380 (-3.6%)	0.114 (-51.0%)	0.084 (-58.1%)	0.095 (-54.1%)
META*	0.701 (-6.5%)	0.611 (-9.2%)	0.707 (-3.2%)	0.592 (-13.3%)	0.288 (-27.0%)	0.218 (-6.2%)	0.229 (+14.7%)	0.318 (+53.7%)
ADV*	0.705 (-5.9%)	0.646 (-4.1%)	0.689 (-5.7%)	0.649 (-5.0%)	0.330 (-16.3%)	0.218 (-6.3%)	0.196 (-1.8%)	0.210 (+1.4%)
PR*	0.729 (-2.7%)	0.641 (-4.8%)	0.724 (-0.9%)	0.642 (-6.0%)	0.346 (-12.3%)	0.144 (-38.2%)	0.123 (-38.1%)	0.125 (-39.5%)
EOP	0.718 (-4.2%)	0.643 (-4.5%)	0.690 (-5.6%)	0.650 (-4.9%)	0.324 (-17.8%)	0.180 (-22.3%)	0.153 (-23.5%)	0.160 (-23.0%)
CEO	0.724 (-3.4%)	0.633 (-6.0%)	0.712 (-2.5%)	0.638 (-6.6%)	0.331 (-16.2%)	0.242 (+4.0%)	0.219 (+9.9%)	0.181 (-12.4%)
ROC	0.710 (-5.3%)	0.689 (+2.3%)	0.694 (-5.0%)	0.666 (-2.5%)	0.380 (-3.6%)	0.190 (-18.0%)	0.158 (-21.0%)	0.140 (-32.3%)
MAAT	0.749 (-0.0%)	0.665 (-1.2%)	0.737 (+0.9%)	0.676 (-1.0%)	0.390 (-1.1%)	0.186 (-20.0%)	0.148 (-25.7%)	0.150 (-27.7%)
FairMask	0.744 (-0.8%)	0.671 (-0.4%)	0.722 (-1.1%)	0.680 (-0.4%)	0.385 (-2.4%)	0.154 (-33.6%)	0.110 (-44.7%)	0.108 (-47.7%)
HIFI*	0.742 (-0.9%)	0.666 (-1.1%)	0.722 (-1.1%)	0.669 (-2.1%)	0.380 (-3.8%)	0.139 (-40.3%)	0.094 (-52.8%)	0.093 (-55.0%)

TABLE VI: (RQ1 & RQ2) The absolute values and relative changes of performance and group fairness metrics averaged on all combinations of (dataset, random seed) for the LR model. The top three values in each column are highlighted. * indicates in-processing methods.

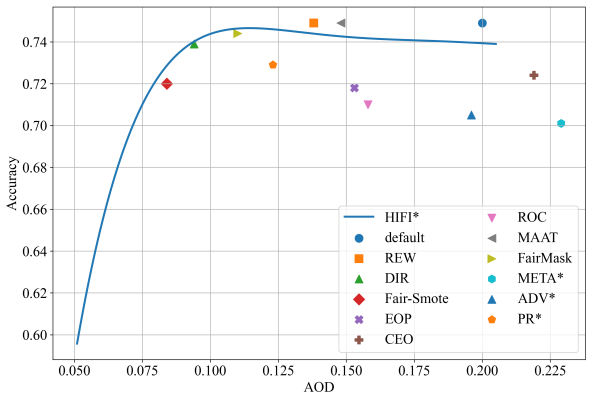


Fig. 7: (RQ2) AOD-accuracy trade-off on the LR model. For HIFI, we vary the weight η (as shown in Table V) to record the performance, and the solid curve is the fitted polynomial with order 5. The closer a dot or a curve to the upper-left corner, the better the method is. * indicates in-processing methods.

B. RQ2: Fairness-perfromance Trade-off

This RQ examines whether the sensitive information encoded by Harsanyi interactions only involving sensitive attributes is completely separate from the task-related information, so that the penalty of the fairness loss term in Eq. 9 will not substantially damage the model utilities.

Fig. 5 shows the performance level distribution for the LR model, where each combination of (dataset, classifier, performance metric, random seed) is considered as a scenario, and the methods are sorted by ascending order in the proportion of cases significantly degrading performance. HIFI only causes a significant decline in model performance in 20% of cases, which is identical to the best methods in this regard, REW and FairMask.

We also measure the average absolute values and relative changes of performance metrics for the LR model in Table VI. HIFI markedly outperforms the existing in-processing methods in terms of maintaining model performance.

To comprehensively evaluate the efficacy of our method in terms of the fairness-performance trade-off, we further assess our method with the Fairea benchmark. Fig. 6 shows the results, and the methods are sorted by descending order in the proportions of 'win-win' and 'good' regions. HIFI exhibits a substantial competence in fairness-performance trade-off, not only significantly outperforming existing in-processing methods but also approaching the effectiveness of the state-of-the-art method, MAAT.

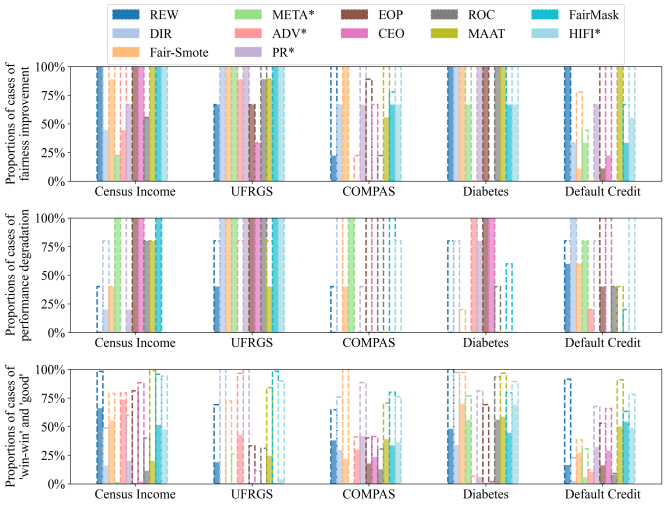
Moreover, we plot the trade-off between AOD and accuracy in Fig. 7 for intuitive understanding. The fitted curve of HIFI on varying weight η is positioned in close proximity to the upper-left corner. HIFI not only achieves a better balance between accuracy and AOD compared to other in-processing methods, but it also allows for flexible customization of different preferences by simply adjusting the weight of the fairness loss term.

> **Ans. to RQ2:** HIFI achieves commendable effectiveness in fairness-performance trade-off, surpassing existing in-processing methods. The results verify that the separation of sensitive information we extracted from task-related information is sufficiently high.

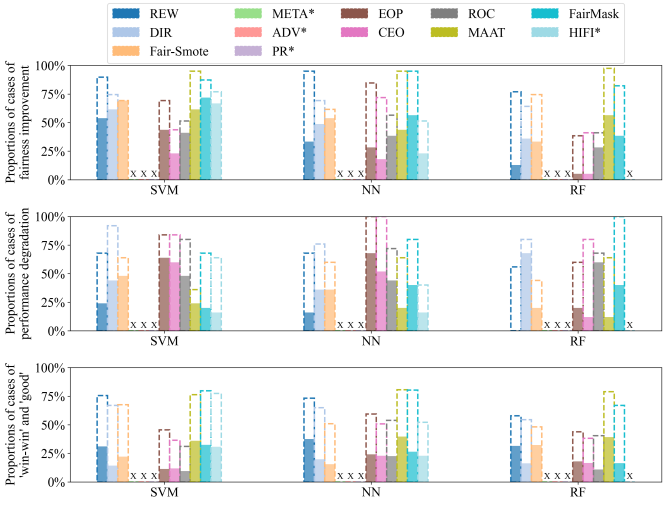
C. RQ3: Applicability

This RQ aims to evaluate the applicability of HIFI, and we compare HIFI with the existing methods on different decision tasks and more ML models.

Fig. 8a presents the dataset-wise proportions of scenarios of fairness improvement, performance degradation, and 'win-win' and 'good' regions in the Fairea benchmark for each method. There is no method that consistently outperforms others across all datasets, so we calculate the ranks (from 1 to 12) of each method across different datasets, and take the mean rank to assess its overall effectiveness and stability in each subplot. For fairness improvement, PR, HIFI and MAAT achieve the top-3 best mean ranks (3.0, 3.2, and 3.2) in terms of significant fairness improvement, and HIFI is the only method that improves fairness in 100% of scenarios across every dataset. For preventing significant model performance degradation, HIFI, MAAT and REW attain the top-3 best mean



(a) Dataset-wise effectiveness for the LR model.



(b) Model-wise effectiveness.

Fig. 8: (RQ3) Proportions of scenarios of fairness improvement, performance degradation, and 'win-win' or 'good' in Fairea of each method across various datasets and ML models. For the figures of fairness improvement and performance degradation, the dashed sections indicate the proportion of changes, while the color-filled sections indicate significantly changed parts. For the figure of Fairea trade-off, the color-filled sections indicate 'win-win' regions, and the remaining dashed sections indicate 'good' regions. 'X' indicates that the method is not applicable to the current model. * indicates in-processing methods.

ranks (1.6, 2.4, and 2.8). On the Fairea benchmark, REW, HIFI and MAAT secure the top-3 best mean ranks (4.4, 4.6, and 4.6) for the combined "win-win" and "good" categories.

We advocate for the use of multiple datasets of different sizes and tasks when evaluating methods. Overall, the UFRGS dataset poses the greatest challenge for balancing

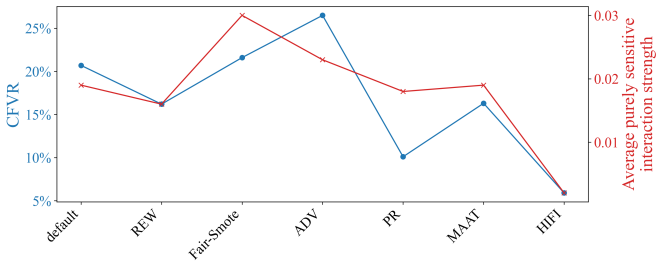


Fig. 9: (RQ4) The CFVR and average absolute value of interactions that only involve sensitive attributes for each method. The Spearman correlation and Kendall's tau correlation between the two are as high as 0.919* and 0.787*, respectively, where * indicates a significant correlation with p-value < 0.05.

model performance with existing methods. We recommend including this dataset when assessing bias mitigation methods. Additionally, many methods exhibit a significant decline in bias mitigation effectiveness on relatively fair datasets like Default Credit, whose mean of group fairness metrics is 0.048 for the default training method.

Similarly, Fig. 8b demonstrates the model-wise proportions for other models besides LR, and we also calculate the mean ranks on the 3 additional ML models for assessment. Note that 'X' indicates that the method is not applicable to the model, so we assign it the worst rank (12). Considering the 3 models besides LR, FairMask achieves the best mean rank (1.3) for significant fairness improvement, REW attains the best mean rank (1.7) for preventing significant performance degradation, and MAAT and FairMask secure the best mean ranks (1.7 for both) on the Fairea benchmark.

Although HIFI cannot be applied to the RF model either, it can be easily adapted to SVM and NN compared to other in-processing methods. HIFI consistently achieves excellent results in maintaining the original model performance across the three supported models. HIFI performs well on the two simple models, LR and SVM, but like most methods, its performance declines on the NN model. We will further discuss the causes of the decline in Sec. VI-C, and explore the potential of HIFI to combine with other approaches for improved debiasing effectiveness in Sec. VI-B.

> **Ans. to RQ3:** On the LR model, HIFI demonstrates outstanding and stable performance in bias mitigation and fairness-performance trade-off across various datasets. For relatively more complex models like NN, using HIFI alone does not ensure satisfactory debiasing results.

D. RQ4: Fairness Compatibility

This RQ probes whether the sensitive information encoded by Harsanyi interactions that only involve sensitive attributes essentially represents an important and common component

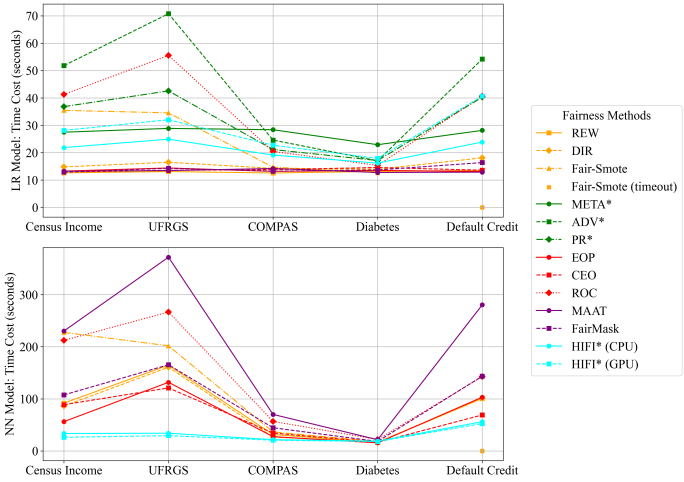


Fig. 10: Time cost of each fairness method on the LR model and the NN model across various datasets. * indicates in-processing methods.

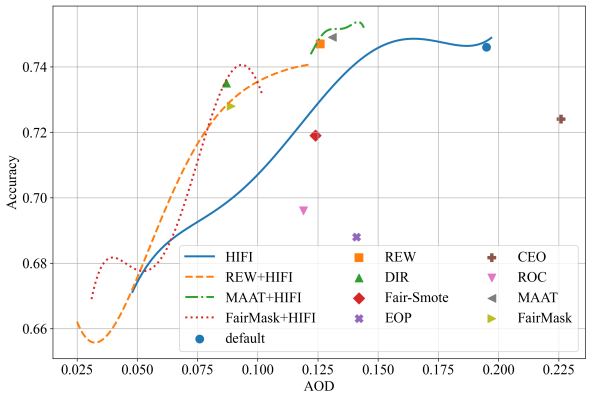


Fig. 11: AOD-accuracy trade-off on the NN model. For HIFI and extended integration, we vary the weight η (as shown in Table V) to record the performance. The curves correspond to fitted polynomials of degree 5. The closer a dot or a curve to the upper-left corner, the better the method is.

of the sensitive information implicitly defined by various fairness metrics. We have verified the effectiveness in bias mitigation by suppressing such interactions in Sec. V-A, and we continue to empirically prove the effectiveness of HIFI in bias mitigation under another kind of fairness definition, namely individual fairness.

Fig. 9 shows the CFVR and average strength of the interactions only involving sensitive attributes for the default training method, the methods considered in Sec. III-B and our method. HIFI reduces such strength from 0.019 to 0.002. Correspondingly, CFVR decreases from 20.7% to 5.9%. We also find a strongly positive correlation between the CFVR and the average strength, which validates the above hypothesis.

Ans. to RQ4: HIFI not only effectively mitigates group discrimination, but also performs well in improving individual fairness. The results corroborate that the sensitive information encoded by Harsanyi interactions that only involve sensitive attributes essentially represents an important and common component of the sensitive information implicitly defined by various fairness metrics.

VI. Discussion

Building upon the findings from our experiments, this section delves deeper into the practical considerations of our method, HIFI. Specifically, we explore aspects such as computational efficiency (e.g., time cost and GPU acceleration), the potential synergy between HIFI and other methods, and the inherent limitations and validity of our approach, providing a comprehensive understanding of HIFI's contributions.

A. Time Cost

In Fig. 10, we record the runtime for each method across different datasets, running 10 rounds with different random seeds simultaneously in a multi-processing setup. On the LR model, although HIFI is over 10 seconds slower than the advanced methods such as REW, MAAT, and FairMask when running on the CPU, it consistently outperforms existing in-processing methods in terms of speed. On the NN model, the runtime for many methods increases several times compared to the LR model, whereas HIFI's runtime only rises from 21.4 seconds to 32.8 seconds on average, making it significantly faster than other methods. In addition, the time cost of each method is influenced not only by the complexity of the model but also by the size of the dataset and the number of features. However, HIFI demonstrates relatively stable runtime performance.

In our implementation, we allow users to easily choose between running on the CPU or enabling GPU acceleration. We find that with GPU acceleration, HIFI actually runs slower on simple models like LR, while on more complex models like NN, it can leverage the GPU's advantages in parallel computation, higher bandwidth, and specialized optimizations over the CPU. With GPU acceleration, HIFI is 10.58% faster compared to running solely on the CPU for the subject NN model. We believe that if HIFI is applied to larger datasets and models, GPU acceleration will become even more important.

In Sec. V, MAAT demonstrates exceptional overall effectiveness. However, due to its use of the ensemble strategy, which requires training multiple models, the training time increases significantly as the model complexity grows. Additionally, when the ensemble model trained by MAAT is used for inference, its computational resource consumption can be several times higher than that of HIFI.

B. Method Synergy

HIFI can be integrated with some other techniques to enhance effectiveness. DIR is an exception because it re-

moves sensitive attributes in the final stage of preprocessing. To explore how HIFI performs alongside other established approaches, we select the most distinguished bias mitigation methods (REW, MAAT and FairMask) for synergy, and propose REW+HIFI, MAAT+HIFI, and FairMask+HIFI. REW+HIFI simply applies HIFI after REW preprocesses the training data. MAAT+HIFI includes HIFI as an additional fair model for ensemble. FairMask+HIFI involves using the model trained with HIFI to make predictions on the test data processed by FairMask.

In Fig. 11, we present the trade-off between AOD and accuracy of the combined methods on the NN model. These combined methods not only enhance HIFI's debiasing effectiveness on the NN model, but also allow for explicit control over debiasing strength while maintaining the trade-off level of the subject methods, which further extends the capabilities of these approaches. Therefore, for complex models, we recommend combining HIFI with other advanced debiasing methods to achieve better effectiveness and flexibility for bias mitigation.

C. Limitations

We assume sensitive attributes are explicitly provided, limiting the direct applicability of our approach to unstructured datasets like images or texts. As an in-processing method, it requires access to training data and the ability to adjust the training objective function, making it unsuitable for proprietary model mechanisms. While compatible with models trained via gradient-based techniques, it does not apply to models like Decision Tree or Random Forest, which do not support gradient-based training. We observe reduced effectiveness on NN models compared to LR models, and its efficacy on larger models requires further evaluation. We hypothesize that the reduced effectiveness of HIFI on NN models may stem from unfaithful baseline values or the complex encoding of sensitive information, with HIFI addressing only part of this complexity.

D. Threats to Validity

Baseline values. To punish the fairness loss term in Eq. 9, we have to compute model outputs on masked inputs according to Eq. 1. However, it is still an open problem to estimate the baseline values to faithfully represent the absence state of an input variable. Here are the available methods: **mean** baseline values [48], **zero** baseline values [49, 78], **blurring** [79, 80], **marginal distribution** [81], **conditional distribution** [82, 83] and **causality**-based baseline values [39]. The mean baseline is considered relatively reasonable among methods with low computational costs. In our cases, the average interaction strength on $\emptyset$, i.e., the input with all variables masked as the baseline values, is 0.381, which does not deviate much from the midpoint 0.5 of the range [0,1] for the predicted probability of the positive label in our settings. If more faithful lightweight methods emerge in future, HIFI will also be further improved.

Datasets. We use widely studied datasets of different tasks and sizes, but potential limitations exist [84]. Our method has not been applied to unstructured datasets, and we focus on sex, race, and age as protected attributes. One could replicate this study with more datasets and protected attributes based on our public artifacts.

ML models. For evaluation, we select four representative ML models commonly used in group fairness literature [26, 27, 43]. HIFI is compatible with gradient-based optimization methods and implemented with GPU acceleration for application to more complex models.

Bias mitigation methods. Incorporating all bias mitigation methods is challenging. We choose to utilize the 11 state-of-the-art methods summarized in the empirical study [27], and they are adapted to simultaneously optimize multiple sensitive attributes for intersectional fairness, if any.

Evaluation criteria. We follow previous studies [27, 43] to use three group fairness metrics and five ML performance criteria, including AOD, which is regarded as the most promising and representative metric for evaluating model fairness [28, 35]. We also incorporate the most commonly used definition of individual fairness, i.e, causal fairness [46, 50, 85]. Moreover, we take both fairness on a single sensitive attribute and intersectional fairness into consideration. We have considered a relatively extensive range in related work.

Access to protected attributes. Our method requires access to the protected attributes of interest, and we have not yet addressed the issue of fairness without demographics [86].

VII. Conclusion

We extract and separate the sensitive information encoded by Harsanyi interactions that only involves protected attributes from task-related information, and propose an in-processing bias mitigation method HIFI by suppressing the strength of such interactions. We have empirically proved that this intrinsically sensitive information essentially represents an important and common component of the sensitive information implicitly defined by various fairness metrics. HIFI surpasses all the existing in-processing methods in terms of fairness improvement and fairness-performance trade-off, and also performs well in reducing violations of individual fairness simultaneously. Moreover, HIFI could be naturally combined with other distinguished methods like REW, MAAT and FairMask to strengthen the effectiveness of debiasing.

Acknowledgment

This work was supported by the Shanghai Trusted Industry Internet Software Collaborative Innovation Center and the "Digital Silk Road" Shanghai International Joint Lab of Trustworthy Intelligent Software (Grant No. 22510750100), and by the National Natural Science Foundation of China (NSFC) for Young Scientists (Grant No. 62202166).

References

[1] J. Zhao, Y. Zhou, Z. Li, W. Wang, and K.-W. Chang, "Learning gender-neutral word embeddings," *arXiv preprint arXiv:1809.01496*, 2018.
[2] J. Chan and J. Wang, "Hiring preferences in online labor markets: Evidence of a female hiring bias," *Management Science*, vol. 64, no. 7, pp. 2973–2994, 2018.
[3] M. Bogen and A. Rieke, "Help wanted: An examination of hiring algorithms, equity, and bias," *Upturn, December*, vol. 7, 2018.

[4] R. Berk, H. Heidari, S. Jabbari, M. Kearns, and A. Roth, "Fairness in criminal justice risk assessments: The state of the art," *Sociological Methods & Research*, vol. 50, no. 1, pp. 3–44, 2021.

[5] N. Mehrabi, F. Morstatter, N. Saxena, K. Lerman, and A. Galstyan, "A survey on bias and fairness in machine learning," *ACM computing surveys (CSUR)*, vol. 54, no. 6, pp. 1–35, 2021.

[6] J. Chen, N. Kallus, X. Mao, G. Svacha, and M. Udell, "Fairness under unawareness: Assessing disparity when protected class is unobserved," in *Proceedings of the conference on fairness, accountability, and transparency*, 2019, pp. 339–348.

[7] M. Hort, J. M. Zhang, F. Sarro, and M. Harman, "Fairea: A model behaviour mutation approach to benchmarking bias mitigation methods," in *Proceedings of the 29th ACM joint meeting on European software engineering conference and symposium on the foundations of software engineering*, 2021, pp. 994–1006.

[8] F. Kamiran and T. Calders, "Data preprocessing techniques for classification without discrimination," *Knowledge and information systems*, vol. 33, no. 1, pp. 1–33, 2012.

[9] M. Feldman, S. A. Friedler, J. Moeller, C. Scheidegger, and S. Venkatasubramanian, "Certifying and removing disparate impact," in *proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining*, 2015, pp. 259–268.

[10] J. Chakraborty, S. Majumder, and T. Menzies, "Bias in machine learning software: Why? how? what to do?" in *Proceedings of the 29th ACM joint meeting on european software engineering conference and symposium on the foundations of software engineering*, 2021, pp. 429–440.

[11] T. Kamishima, S. Akaho, H. Asoh, and J. 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*. Springer, 2012, pp. 35–50.

[12] B. H. Zhang, B. Lemoine, and M. Mitchell, "Mitigating unwanted biases with adversarial learning," in *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, 2018, pp. 335–340.

[13] L. E. Celis, L. Huang, V. Keswani, and N. K. Vishnoi, "Classification with fairness constraints: A meta-algorithm with provable guarantees," in *Proceedings of the conference on fairness, accountability, and transparency*, 2019, pp. 319–328.

[14] J. Yang, L. Zhang, and M. Zhang, "Making fair classification via correlation alignment," in *ECAI 2024*. IOS Press, 2024, pp. 842–849.

[15] F. Kamiran, A. Karim, and X. Zhang, "Decision theory for discrimination-aware classification," in *2012 IEEE 12th international conference on data mining*. IEEE, 2012, pp. 924–929.

[16] M. Hardt, E. Price, and N. Srebro, "Equality of opportunity in supervised learning," *Advances in neural information processing systems*, vol. 29, 2016.

[17] G. Pleiss, M. Raghavan, F. Wu, J. Kleinberg, and K. Q. Weinberger, "On fairness and calibration," *Advances in neural information processing systems*, vol. 30, 2017.

[18] P. Zhang, J. Wang, J. Sun, G. Dong, X. Wang, X. Wang, J. S. Dong, and T. Dai, "White-box fairness testing through adversarial sampling," in *Proceedings of the ACM/IEEE 42nd international conference on software engineering*, 2020, pp. 949–960.

[19] L. Zhang, Y. Zhang, and M. Zhang, "Efficient white-box fairness testing through gradient search," in *Proceedings of the 30th ACM SIGSOFT International Symposium on Software Testing and Analysis*, 2021, pp. 103–114.

[20] H. Zheng, Z. Chen, T. Du, X. Zhang, Y. Cheng, S. Ji, J. Wang, Y. Yu, and J. Chen, "Neuronfair: Interpretable white-box fairness testing through biased neuron identification," in *Proceedings of the 44th International Conference on Software Engineering*, 2022, pp. 1519–1531.

[21] Z. Wang, M. Zhang, J. Yang, B. Shao, and M. Zhang, "Maft: Efficient model-agnostic fairness testing for deep neural networks via zero-order gradient search," in *Proceedings of the IEEE/ACM 46th International Conference on Software Engineering*, 2024, pp. 1–12.

[22] G. Tao, W. Sun, T. Han, C. Fang, and X. Zhang, "Ruler: discriminative and iterative adversarial training for deep neural network fairness," in *Proceedings of the 30th acm joint european software engineering conference and symposium on the foundations of software engineering*, 2022, pp. 1173–1184.

[23] J. Chen, Y. Zhang, L. Zhang, M. Zhang, C. Wan, T. Su, and G. Pu, "Fipser: Improving fairness testing of dnn by seed prioritization," in *Proceedings of the 39th IEEE/ACM International Conference on Automated Software Engineering*, ser. ASE '24. New York, NY, USA: Association for Computing Machinery, 2024, p. 1069–1081. [Online]. Available: https://doi.org/10.1145/3691620.3695486

[24] S. Corbett-Davies, E. Pierson, A. Feller, S. Goel, and A. Huq, "Algorithmic decision making and the cost of fairness," in *Proceedings of the 23rd acm sigkdd international conference on knowledge discovery and data mining*, 2017, pp. 797–806.

[25] M. Wick, J.-B. Tristan *et al.*, "Unlocking fairness: a trade-off revisited," *Advances in neural information processing systems*, vol. 32, 2019.

[26] Z. Chen, J. M. Zhang, F. Sarro, and M. Harman, "A comprehensive empirical study of bias mitigation methods for machine learning classifiers," *ACM Transactions on Software Engineering and Methodology*, vol. 32, no. 4, pp. 1–30, 2023.

[27] ——, "Fairness improvement with multiple protected attributes: How far are we?" in *Proceedings of the IEEE/ACM 46th International Conference on Software Engineering*, 2024, pp. 1–13.

[28] Z. Ji, P. Ma, S. Wang, and Y. Li, "Causality-aided trade-off analysis for machine learning fairness," in *2023 38th IEEE/ACM International Conference on Automated Software Engineering (ASE)*. IEEE, 2023, pp. 371–383.

[29] J. Kleinberg, S. Mullainathan, and M. Raghavan, "Inherent trade-offs in the fair determination of risk scores," *arXiv preprint arXiv:1609.05807*, 2016.

[30] A. D. Selbst, D. Boyd, S. A. Friedler, S. Venkatasubramanian, and J. Vertesi, "Fairness and abstraction in sociotechnical systems," in *Proceedings of the conference on fairness, accountability, and transparency*, 2019, pp. 59–68.

[31] C. Dwork, M. Hardt, T. Pitassi, O. Reingold, and R. Zemel, "Fairness through awareness," in *Proceedings of the 3rd innovations in theoretical computer science conference*, 2012, pp. 214–226.

[32] R. Binns, "On the apparent conflict between individual and group fairness," in *Proceedings of the 2020 conference on fairness, accountability, and transparency*, 2020, pp. 514–524.

[33] S. A. Friedler, C. Scheidegger, and S. Venkatasubramanian, "The (im) possibility of fairness: Different value systems require different mechanisms for fair decision making," *Communications of the ACM*, vol. 64, no. 4, pp. 136–143, 2021.

[34] S. Ruggieri, J. M. Alvarez, A. Pugnana, F. Turini *et al.*, "Can we trust fair-ai?" in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 37, no. 13, 2023, pp. 15 421–15 430.

[35] J. Yang, J. Jiang, Z. Sun, and J. Chen, "A large-scale empirical study on improving the fairness of deep learning models," *arXiv preprint arXiv:2401.03695*, 2024.

[36] J. Ren, M. Li, Q. Chen, H. Deng, and Q. Zhang, "Defining and quantifying the emergence of sparse concepts in dnns," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2023, pp. 20 280–20 289.

[37] J. C. Harsanyi and J. C. Harsanyi, "A simplified bargaining model for the n-person cooperative game," *Papers in game theory*, pp. 44–70, 1982.

[38] M. Li and Q. Zhang, "Defining and quantifying and-or interactions for faithful and concise explanation of dnns," *arXiv preprint arXiv:2304.13312*, 2023.

[39] J. Ren, Z. Zhou, Q. Chen, and Q. Zhang, "Can we faithfully represent absence states to compute shapley values on a DNN?" in *The Eleventh International Conference on Learning Representations*, 2023. [Online]. Available: https://openreview.net/forum?id=YV8tP7bW6Kt

[40] M. Li and Q. Zhang, "Does a neural network really encode symbolic concepts?" in *International conference on machine learning*. PMLR, 2023, pp. 20 452–20 469.

[41] Q. Ren, J. Gao, W. Shen, and Q. Zhang, "Where we have arrived in proving the emergence of sparse symbolic concepts in ai models," *arXiv preprint arXiv:2305.01939*, 2023.

[42] D. Liu, H. Deng, X. Cheng, Q. Ren, K. Wang, and Q. Zhang, "Towards the difficulty for a deep neural network to learn concepts of different complexities," *Advances in Neural Information Processing Systems*, vol. 36, 2024.

[43] Z. Chen, J. M. Zhang, F. Sarro, and M. Harman, "Maat: a novel ensemble approach to addressing fairness and performance bugs for machine learning software," in *Proceedings of the 30th ACM joint european software engineering conference and symposium on the foundations of software engineering*, 2022, pp. 1122–1134.

[44] K. Peng, J. Chakraborty, and T. Menzies, "Fairmask: Better fairness via model-based rebalancing of protected attributes," *IEEE Transactions on Software Engineering*, vol. 49, no. 4, pp. 2426–2439, 2022.

[45] T. Calders, F. Kamiran, and M. Pechenizkiy, "Building classifiers with independency constraints," in *2009 IEEE international conference on data mining workshops*. IEEE, 2009, pp. 13–18.

[46] S. Galhotra, Y. Brun, and A. Meliou, "Fairness testing: testing software for discrimination," in *Proceedings of the 2017 11th Joint meeting on foundations of software engineering*, 2017, pp. 498–510.

[47] J. Pearl, *Causality*. Cambridge university press, 2009.

[48] P. Dabkowski and Y. Gal, "Real time image saliency for black box classifiers," *Advances in neural information processing systems*, vol. 30, 2017.

[49] M. Ancona, C. Oztireli, and M. Gross, "Explaining deep neural networks with a polynomial time algorithm for shapley value approximation," in *International Conference on Machine Learning*. PMLR, 2019, pp. 272–281.

[50] Z. Chen, J. M. Zhang, M. Hort, M. Harman, and F. Sarro, "Fairness testing: A comprehensive survey and analysis of trends," *ACM Transactions on Software Engineering and Methodology*, vol. 33, no. 5, pp. 1–59, 2024.

[51] H. Zhang, Y. Xie, L. Zheng, D. Zhang, and Q. Zhang, "Interpreting multivariate shapley interactions in dnns," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, no. 12, pp. 10 877–10 886, May 2021. [Online]. Available: https://ojs.aaai.org/index.php/AAAI/article/view/17299

[52] D. Zhang, H. Zhang, H. Zhou, X. Bao, D. Huo, R. Chen, X. Cheng, M. Wu, and Q. Zhang, "Building interpretable interaction trees for deep nlp models," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, no. 16, pp. 14 328–14 337, May 2021. [Online]. Available: https://ojs.aaai.org/index.php/AAAI/article/view/17685

[53] H. Zhang, S. Li, Y. Ma, M. Li, Y. Xie, and Q. Zhang, "Interpreting and boosting dropout from a game-theoretic view," in *International Conference on Learning Representations*, 2021. [Online]. Available: https://openreview.net/forum?id=Jacdvfjicf7

[54] L. Chen, S. Lou, K. Zhang, J. Huang, and Q. Zhang, "Harsanyinet: Computing accurate shapley values in a single forward propagation," 2023. [Online]. Available: https://arxiv.org/abs/2304.01811

[55] L. Chen, S. Lou, B. Huang, and Q. Zhang, "Defining and extracting generalizable interaction primitives from DNNs," in *The Twelfth International Conference on Learning Representations*, 2024. [Online]. Available: https://openreview.net/forum?id=OCqyFVFNeF

[56] X. Wang, S. Lin, H. Zhang, Y. Zhu, and Q. Zhang, "Interpreting attributions and interactions of adversarial attacks," in *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021, pp. 1075–1084.

[57] X. Cheng, C. Chu, Y. Zheng, J. Ren, and Q. Zhang, "A game-theoretic taxonomy of visual concepts in dnns," *ArXiv*, vol. abs/2106.10938, 2021. [Online]. Available:

https://api.semanticscholar.org/CorpusID:235489713
[58] H. Deng, Q. Ren, H. Zhang, and Q. Zhang, "DISCOVERING AND EXPLAINING THE REPRESENTATION BOTTLENECK OF DNNS," in *International Conference on Learning Representations*, 2022. [Online]. Available: https://openreview.net/forum?id=iRCUlgmdfHJ
[59] H. Zhou, H. Zhang, H. Deng, D. Liu, W. Shen, S.-H. Chan, and Q. Zhang, "Explaining generalization power of a dnn using interactive concepts," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, no. 15, pp. 17 105–17 113, Mar. 2024. [Online]. Available: https://ojs.aaai.org/index.php/AAAI/article/view/29655
[60] Q. Ren, H. Deng, Y. Chen, S. Lou, and Q. Zhang, "Bayesian neural networks avoid encoding complex and perturbation-sensitive concepts," 2023. [Online]. Available: https://arxiv.org/abs/2302.13095
[61] X. Cheng, X. Wang, H. Xue, Z. Liang, and Q. Zhang, "A hypothesis for the aesthetic appreciation in neural networks," 2021. [Online]. Available: https://arxiv.org/abs/2108.02646
[62] J. Ren, D. Zhang, Y. Wang, L. Chen, Z. Zhou, Y. Chen, X. Cheng, X. Wang, M. Zhou, J. Shi, and Q. Zhang, "A unified game-theoretic interpretation of adversarial robustness," in *Proceedings of the 35th International Conference on Neural Information Processing Systems*, ser. NIPS '21. Red Hook, NY, USA: Curran Associates Inc., 2024.
[63] X. Wang, J. Ren, S. Lin, X. Zhu, Y. Wang, and Q. Zhang, "A unified approach to interpreting and boosting adversarial transferability," *CoRR*, vol. abs/2010.04055, 2020. [Online]. Available: https://arxiv.org/abs/2010.04055
[64] Q. Zhang, X. Wang, J. Ren, X. Cheng, S. Lin, Y. Wang, and X. Zhu, "Proving common mechanisms shared by twelve methods of boosting adversarial transferability," 2022. [Online]. Available: https://arxiv.org/abs/2207.11694
[65] H. Deng, N. Zou, M. Du, W. Chen, G. Feng, Z. Yang, Z. Li, and Q. Zhang, "Understanding and unifying fourteen attribution methods with taylor interactions," 2023. [Online]. Available: https://arxiv.org/abs/2303.01506
[66] J. L. Myers, A. D. Well, and R. F. Lorch Jr, *Research design and statistical analysis*. Routledge, 2013.
[67] M. G. Kendall, "A new measure of rank correlation," *Biometrika*, vol. 30, no. 1-2, pp. 81–93, 1938.
[68] B. Becker and R. Kohavi, "Adult," UCI Machine Learning Repository, 1996, DOI: https://doi.org/10.24432/C5XW20.
[69] B. C. da Silva, "UFRGS Entrance Exam and GPA Data," 2019. [Online]. Available: https://doi.org/10.7910/DVN/O35FW8
[70] ProPublica, "Compas analysis," 2016, accessed: 2024-07-05. [Online]. Available: https://github.com/propublica/compas-analysis
[71] M. Kahn, "Diabetes," UCI Machine Learning Repository, DOI: https://doi.org/10.24432/C5T59G.
[72] I.-C. Yeh, "Default of Credit Card Clients," UCI Machine Learning Repository, 2016, DOI: https://doi.org/10.24432/C55S3H.
[73] V. Monjezi, A. Trivedi, G. Tan, and S. Tizpaz-Niari, "Information-theoretic testing and debugging of fairness defects in deep neural networks," in *2023 IEEE/ACM 45th International Conference on Software Engineering (ICSE)*. IEEE, 2023, pp. 1571–1582.
[74] R. K. E. Bellamy, K. Dey, M. Hind, S. C. Hoffman, S. Houde, K. Kannan, P. Lohia, J. Martino, S. Mehta, A. Mojsilović, S. Nagar, K. N. Ramamurthy, J. Richards, D. Saha, P. Sattigeri, M. Singh, K. R. Varshney, and Y. Zhang, "Ai fairness 360: An extensible toolkit for detecting and mitigating algorithmic bias," *IBM Journal of Research and Development*, vol. 63, no. 4/5, pp. 4:1–4:15, 2019.
[75] B. W. Matthews, "Comparison of the predicted and observed secondary structure of t4 phage lysozyme," *Biochimica et Biophysica Acta (BBA)-Protein Structure*, vol. 405, no. 2, pp. 442–451, 1975.
[76] H. B. Mann and D. R. Whitney, "On a test of whether one of two random variables is stochastically larger than the other," *The annals of mathematical statistics*, pp. 50–60, 1947.
[77] B. Kitchenham, L. Madeyski, D. Budgen, J. Keung, P. Brereton, S. Charters, S. Gibbs, and A. Pohthong, "Robust statistical methods for empirical software engineering," *Empirical Software Engineering*, vol. 22, pp. 579–630, 2017.
[78] M. Sundararajan, A. Taly, and Q. Yan, "Axiomatic attribution for deep networks," in *International conference on machine learning*. PMLR, 2017, pp. 3319–3328.
[79] R. C. Fong and A. Vedaldi, "Interpretable explanations of black boxes by meaningful perturbation," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 3429–3437.
[80] R. Fong, M. Patrick, and A. Vedaldi, "Understanding deep networks via extremal perturbations and smooth masks," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 2950–2958.
[81] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," *Advances in neural information processing systems*, vol. 30, 2017.
[82] C. Frye, D. de Mijolla, T. Begley, L. Cowton, M. Stanley, and I. Feige, "Shapley explainability on the data manifold," *arXiv preprint arXiv:2006.01272*, 2020.
[83] I. Covert, S. M. Lundberg, and S.-I. Lee, "Understanding global feature contributions with additive importance measures," *Advances in Neural Information Processing Systems*, vol. 33, pp. 17 212–17 223, 2020.
[84] F. Ding, M. Hardt, J. Miller, and L. Schmidt, "Retiring adult: New datasets for fair machine learning," *Advances in neural information processing systems*, vol. 34, pp. 6478–6490, 2021.
[85] M. Zhang and J. Sun, "Adaptive fairness improvement based on causality analysis," in *Proceedings of the 30th ACM Joint European Software Engineering Conference and Symposium on the Foundations of Software Engineering*, 2022, pp. 6–17.
[86] P. Lahoti, A. Beutel, J. Chen, K. Lee, F. Prost, N. Thain, X. Wang, and E. Chi, "Fairness without demographics through adversarially reweighted learning," *Advances in neural information processing systems*, vol. 33, pp. 728–740, 2020.