AIM: Attributing, Interpreting, Mitigating Data Unfairness

Zhining Liu
liu326@illinois.edu
University of Illinois
Urbana-Champaign
Urbana, IL, USA

Ruizhong Qiu
rq5@illinois.edu
University of Illinois
Urbana-Champaign
Urbana, IL, USA

Zhichen Zeng
zhichenz@illinois.edu
University of Illinois
Urbana-Champaign
Urbana, IL, USA

Yada Zhu
yzhu@us.ibm.com
IBM Research
Yorktown Heights, NY, USA

Hendrik Hamann
hendrikh@us.ibm.com
IBM Research
Yorktown Heights, NY, USA

Hanghang Tong
htong@illinois.edu
University of Illinois
Urbana-Champaign
Urbana, IL, USA

Abstract

Data collected in the real world often encapsulates historical discrimination against disadvantaged groups and individuals. Existing fair machine learning (FairML) research has predominantly focused on mitigating discriminative bias in the model prediction, with far less effort dedicated towards exploring how to trace biases present in the data, despite its importance for the transparency and interpretability of FairML. To fill this gap, we investigate a novel research problem: discovering samples that reflect biases/prejudices from the training data.. Grounding on the existing fairness notions, we lay out a sample bias criterion and propose practical algorithms for measuring and countering sample bias. The derived bias score provides intuitive **sample-level attribution** and **explanation** of historical bias in data. On this basis, we further design two FairML strategies via sample-bias-informed minimal data editing. They can **mitigate both group and individual unfairness** at the cost of **minimal or zero predictive utility loss**. Extensive experiments and analyses on multiple real-world datasets demonstrate the effectiveness of our methods in explaining and mitigating unfairness. Code is available at https://github.com/ZhiningLiu1998/AIM.

CCS Concepts

• **Computing methodologies** → **Machine learning**; **Artificial intelligence**; • **Social and professional topics**;

Keywords

FairML, Group Fairness, Individual Fairness, Bias Attribution

ACM Reference Format:
Zhining Liu, Ruizhong Qiu, Zhichen Zeng, Yada Zhu, Hendrik Hamann, and Hanghang Tong. 2024. AIM: Attributing, Interpreting, Mitigating Data Unfairness. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD '24), August 25–29, 2024, Barcelona, Spain.* ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/3637528.3671797

KDD '24, August 25–29, 2024, Barcelona, Spain

ACM ISBN 979-8-4007-0490-1/24/08
https://doi.org/10.1145/3637528.3671797

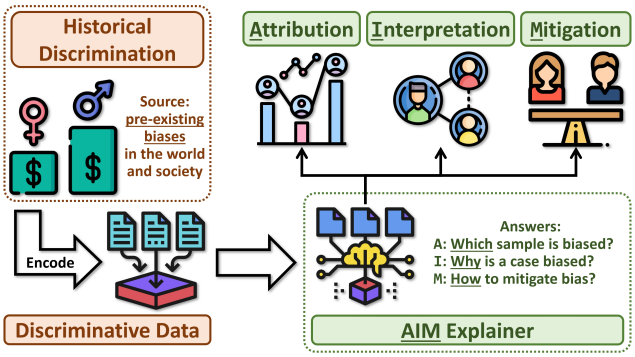

Figure 1: Concept and applications of the proposed AIM (Bias Attribution, Interpretation, Mitigation) framework.

1 Introduction

Machine learning techniques are increasingly used in high-stake scenarios such as loans, recruitment, and policing strategies. Despite the benefits of automated decision-making, data-driven models are susceptible to biases that render their decisions potentially unfair, reflecting racism, ageism, and sexism [10, 41]. With the increasing demand for equitable and responsible use of artificial intelligence, Fair Machine Learning (FairML) has gained significant attention in both research and practice [6, 31].

Existing FairML research mainly focuses on how to devise learning algorithms to guarantee that the model has no prejudice or favoritism toward an individual or group based on their inherent or acquired characteristics [41]. In essence, the bias in machine learning models is inherited from their training data: even with perfect sampling and feature selection, data collected from the real world inevitably contains historical bias [41, 54], which is a result of the pre-existing biases and socio-technical issues in the world [19, 32]. For example, discrimination against female/minority by loan providers recorded in databases can be inherited by the learning model for loan screening, and eventually leads to outcomes that are unfavorable for female/minority applicants [22]. Similar situations arise in various domains of daily living, such as employment, housing, insurance, credit scoring, and many more [19, 22]. Identifying and understanding such biases encoded in the data is therefore crucial for achieving more transparent, interpretable, and equitable machine learning systems.

To this end, we delve into an under-explored aspect of FairML: discovering samples that reflect biases/prejudices from the training data. Practically, this can assist human experts in understanding the bias within the data, locating and scrutinizing discriminatory samples, and developing more trustworthy FairML techniques based on these insights. With its root in social, ethical and legal literature on fairness [8, 17, 20, 54], we posit that a good sample bias criterion should be able to *robustly capture various prejudices encoded in the data, whether targeted towards specific individuals or demographic groups.* Grounding on this, we consider a sample exhibits bias if its comparable samples from other groups receive *different* and *credible* treatments. As a practical example, we consider a female applicant being rejected for a loan as being discriminated if (1) a male applicant with similar conditions gets the loan (different), and (2) the approval is not by chance (credible). Intuitively, it captures both individual-level and group-level fairness and prevents incidental events or noise in the real world from disturbing bias estimates.

Our criterion is grounded in the overarching principles behind popular algorithmic fairness notions such as group [11, 17], individual [17, 27], and counterfactual fairness [35], but without depending on intricate causal modeling or requiring additional expert knowledge. Existing fairness research also largely overlooks the imperfections in the data: due to various unexpected subjective and objective factors in data generation and processing, the data collected from the real world often contains unavoidable noise and errors [21, 25], which can significantly disturb the FairML process [57, 58]. To address this, we introduce sample credibility in FairML to obtain robust bias estimates, and practical algorithms are further proposed for estimating sample bias and credibility. Our bias criterion is also self-explanatory: the bias of a sample can be naturally explained by corresponding other-group samples that receive different and credible treatments, which provides further insights for human experts to inspect the bias in the data.

Our bias and confidence estimation require a sample similarity metirc. In principle, any reasonable similarity measure on the feature space can seamlessly integrate with our framework. However, due to the complexity of real-world data, practical applications typically require human experts to manually design/annotate similarities for each task, resulting in significant costs [18, 20, 43]. To this end, we propose a practical and intuitive similarity measurement that requires the minimal user input based on two key concepts: (i) creating a comparability graph to capture local similarities between input samples; and (ii) applying graph proximity measures on the comparability graph to capture global similarities reflecting the manifold structure of the input data. This approach yields *localized*, *intuitive*, and *interpretable* sample similarities to better support practical bias attribution and interpretation.

Finally, we explore the potential of mitigating unfairness based on bias attribution. We propose two strategies to mitigate unfairness with informed minimal data editing, namely unfairness removal and fairness augmentation. By discarding a small fraction of samples with high bias or augmenting samples with low bias, the proposed methods can *mitigate group and individual unfairness* at the cost of *minimal or zero predictive utility loss.* Extensive experiments and analyses on multiple real-world FairML tasks demonstrate the effectiveness of the proposed unfairness mitigation strategies.

Figure 1 shows the concept and applications of AIM. To sum up, our AIM (Bias Attribution, Interpretation, Mitigation) framework can assist practitioners in addressing the following problems:

- **Attribution**: **Which** samples exhibit bias?
- **Interpretation**: **Why** is a particular sample biased?
- **Mitigation**: **How** to counter unfairness with auditable data editing and minimal utility loss?

Our contributions are 3-fold:

(1) **Problem Formulation.** We formulate a novel problem of identifying samples that encode discrimination in the data, which is crucial for achieving more transparent FairML.
(2) **Algorithm Design.** We propose a novel framework AIM. Armed with credibility-aware sample bias criterion and similarity based on user-defined comparable constraints, AIM offers robust, practical, and self-explanatory sample-level bias attribution. The results further support efficient unfairness mitigation with minimal data editing and utility loss.
(3) **Experimental Evaluation.** We provide comprehensive experiments and analyses on real-world datasets to validate the effectiveness of AIM in explaining and mitigating unfairness.

2 Preliminaries

Notations. Real-world data usually contains both numerical (e.g., age, income) and categorical features (e.g., job type, residence). In this paper, we consider the attribute vector $\mathbf{x}$ contains numerical features with real values $\mathbf{r} = [\mathbf{r}^{(1)}, \mathbf{r}^{(2)}, \cdots, \mathbf{r}^{(n_r)}]$, and categorical features with discrete values $\mathbf{d} = [\mathbf{d}^{(1)}, \mathbf{d}^{(2)}, \cdots, \mathbf{d}^{(n_d)}]$, where n_r/n_d denote the number of numerical/categorical features. We use $\mathbf{r_x}/\mathbf{d_x}$ to denote the numerical/categorical part of feature vector $\mathbf{x}$. For simplicity, we assume numerical features $\mathbf{r}$ contain values in $[0, 1]$ after some scaling/normalization operation. We consider binary labels and demographic groups following the common setting in algorithm fairness [6, 10, 41]. Without loss of generality, we consider label space $\mathcal{Y} := \{0, 1\}$ and sensitive group membership space $\mathcal{S} := \{0, 1\}$, with positive value representing the favorable outcome/treatment (e.g., loan approval) and the advantaged group (e.g., gender/race with favoritism). A dataset with n samples is denoted as $\mathcal{D} : \{(\mathbf{x}_i, y_i, s_i)|i = 0, 1, \cdots, n\}$ with the i-th data instance in $\mathcal{D}$ as $(\mathbf{x}_i, y_i, s_i)$. We provide the problem definition as follows:

Problem 2.1 (*Unfairness Attribution, Interpretation, Mitigation*)**.** Given a tabular dataset $\mathcal{D} : \{(\mathbf{x}_i, y_i, s_i)|i = 0, 1, \cdots, n\}$ containing historical discrimination, we aim to solve the following problems. *Unfairness Attribution*: quantifying the historical bias carried by each instance $(\mathbf{x}, y, s)$. *Unfairness Interpretation*: providing samples as justifications to explain why a specific instance $(\mathbf{x}, y, s)$ is biased/unbiased. *Unfairness Mitigation*: debiasing the dataset such that the predictive model trained on the debiased $\mathcal{D}$ inherits as little discrimination as possible while retaining the predictive utility.

3 Methodology

In this section, we present our AIM (Bias Attribution, Interpretation, Mitigation) framework. We first introduce our sample bias criterion for unfairness *attribution* and *interpretation*, then discuss its rationales and connections to existing fairness notions. In short,

our criterion captures both individual-level and group-level unfairness and prevents incidental events or noise in the real world from disturbing bias estimates. Then, we propose a novel similarity measure based on user-defined comparable constraints to support reasonable attribution and interpretation of sample bias. It allows practical, configurable, and interpretable similarity computing in complex heterogeneous feature spaces without relying on human prior moral judgment. Finally, we propose two practical unfairness *mitigation* strategies based on the bias attribution results, namely unfairness removal (AIM_{REM}) and fairness augmentation (AIM_{AUG}). By removing/augmenting a small fraction of unfair/fair samples, our mitigation algorithms can alleviate both group and individual unfairness with minimal utility loss.

3.1 Sample Bias Criterion

We first introduce the definition of sample bias and then discuss the rationale for our design. Aligned with the philosophy that fairness is the *absence of any prejudice or favoritism towards an individual or group based on their inherent or acquired characteristics* [41], our goal is to establish a sample bias definition that can effectively characterize the prejudices encoded in data, be it directed towards specific individuals or demographic groups. Specifically, assuming given (i) an appropriate similarity function $\sigma_{\mathcal{X}}(\cdot,\cdot) : \mathcal{X}\times\mathcal{X} \mapsto [0,1]$ defined on the input feature space, and (ii) the credibility $c_i \in [0,1]$ of each data instance $(\mathbf{x}_i, y_i, s_i)$, we define the criterion of sample bias as follows.

Definition 3.1 (*Sample Bias*)**.** A data sample $(\mathbf{x}_i, y_i, s_i)$ is biased if its similar samples specified by $\sigma_{\mathcal{X}}(\cdot,\cdot)$ from the other sensitive group $\mathcal{D}_{s_j \neq s_i} := \{(\mathbf{x}_j, y_j, s_j) | s_j \neq s_i\}$ receive different (i.e., $y_j \neq y_i$) and credible (i.e., with high credibility c_i) treatments.

In an ideal world, the credibility of samples could be assessed by domain experts reviewing each data instance. However, this would incur substantial costs, making it generally impractical in real-world applications. When human-evaluated credibility is not available, we propose to use the following more practical definition of credibility that can be straightforwardly computed:

Definition 3.2 (*Sample Credibility*)**.** A sample $(\mathbf{x}_i, y_i, s_i)$ with label y_i is credible if its similar samples specified by $\sigma_{\mathcal{X}}(\cdot,\cdot)$ from the same sensitive group $\mathcal{D}_{s_j = s_i} := \{(\mathbf{x}_j, y_j, s_j) | s_j = s_i\}$ received same treatments (i.e., $y_j = y_i$).

Remark 3.3 (*Intuition and Example*). The intuition behind Definitions 3.1 and 3.2 is that if an individual a from group A received treatment y with high credibility (i.e., other same-group individuals similar to a also receive the same treatment y.), then an individual b from group B with similar attributes to a should also receive the same treatment y. For a practical example, if a male applicant a is approved for a loan, and this decision is credible (i.e., not caused by random rare events or data errors), then a female applicant b with similar conditions (income, education, etc.) to a should also have her loan approved. If this is not the case, then we consider b was being discriminated and thus the data sample documenting her application case exhibits historical bias.

Connection to Existing Fairness Notions. Our sample bias criterion is grounded in the overarching principles behind popular algorithmic fairness notions, including group [17], individual [17, 27], and counterfactual fairness [35]. Specifically, as a widely used fairness notion, group fairness (GF) promotes equitable outcomes for different groups in terms of statistics such as positive rates. However, GF has been criticized for lacking guarantees on the treatment of individual cases [17, 27] since it is defined on the group average. Alternatively, individual fairness (IF) is based on the consensus that "similar individuals should be treated similarly", but the absence of group constraints makes it challenging for IF to characterize systematic discrimination and related implicit bias in data [8, 20]. Our bias is defined on individuals but goes beyond just considering the consistency of similar samples. It simultaneously takes into account demographic membership to ensure fairness across different groups. Further, counterfactual fairness (CF) is based on the intuition that "an individual should receive same decision in both the actual world and a counterfactual world where the individual belonged to a different demographic group". Nevertheless, CF generally relies on causal models that require substantial domain knowledge and cost to obtain unobserved variables and construct the associated causal graph. Even with such efforts, causal models can only be built under strong assumptions [35]. Interestingly, recent research suggests that CF is largely equivalent to demographic parity [47], a basic group fairness constraint.

Our bias definition is partially inspired by CF but has been reasonably simplified, and moreover, taking into account the credibility of the samples, a point that is largely overlooked by the previous works but is crucial for robust sample bias attribution. This makes our definition not reliant on expert knowledge and strong assumptions for constructing causal models and (ways of estimating) latent variables [35, 47], while also preventing random events/noise in the real world from disrupting bias estimation. To sum up, compared with existing fairness notions, our sample bias definition can characterize individual-level and group-level fairness *in a practical, intuitive, and robust manner*, while also allowing self-explanatory bias attribution.

3.2 Unfairness Attribution and Interpretation

Computing Self-Explanatory AIM Bias Score. We now formally describe our AIM bias attribution algorithm. To further illustrate the practical implications of our sample bias (Definition 3.1) and credibility (Definition 3.2) criterion, we present probabilistic definitions for bias and credibility based on the underlying data distribution, and show that our criterion can be seen as utilizing a weighted local regression [13–15] to estimate sample bias/credibility on the observed data. To start with, we give the probabilistic definition of sample bias b_i for a data instance $(\mathbf{x}_i, y_i, s_i)$:

$$b_i := \Pr[Y = \neg y_i \mid S = \neg s_i, X = \mathbf{x}_i], \tag{1}$$

i.e., the true probability that a sample with identical attributes $\mathbf{x}_i$ but from another sensitive group $\neg s_i$ has the opposite label $\neg y_i$. A larger b_i signifies that the sample is more likely to have received treatment inconsistent with similar samples from different groups, thus carrying more historical bias.

However, it is evident that b_i cannot be directly calculated as the underlying distribution $\Pr[Y|S,X]$ is usually unknown. Alternatively, we can utilize weighted local regression (WLR) [13] as a non-parametric statistical method to estimate b_i based on the observed data. The general idea of WLR is to do regression on samples that are in the neighborhood of the point being estimated, thereby providing a more accurate and efficient estimation of the target based on local data patterns. Here we employ a soft variant, where the similarity $\sigma_{\mathcal{X}}(\cdot,\cdot)$ specifies the more local samples by assigning higher weights instead of doing hard nearest-neighbor selection [13] as in the original WLR. For robust bias estimation, we incorporate sample credibility $\hat{c}_j$ as the weighting function, which will be defined later in this section. Intuitively, if a sample j has low credibility, say $\hat{c}_j = 0$, it will have no effect in estimating the bias. Specifically, recall that for any event A, we have $\Pr[A \mid S, X] = \mathbb{E}[\mathbb{1}[A] \mid S, X]$. We can estimate b_i by doing regression w.r.t. indicators:

$$\hat{b}_i := \arg\min_{b\in\mathbb{R}} \sum_{j\in\mathcal{D}} \mathbb{1}[s_j = \neg s_i]\sigma_{\mathcal{X}}(\mathbf{x}_j,\mathbf{x}_i)\hat{c}_j(b - \mathbb{1}[y_j = \neg y_i])^2, \quad (2)$$

where $\mathbb{1}[s_j = \neg s_i]$ implies that we use different-group samples for regression, $\sigma_{\mathcal{X}}(\mathbf{x}_j,\mathbf{x}_i)$ specifies the local samples, and $\hat{c}_j$ serves as the weighting function which gives high weights to more credible data points. Eq. (2) has closed form solution

$$\hat{b}_i = \frac{\sum_{j\in\mathcal{D}} \mathbb{1}[s_j = \neg s_i]\mathbb{1}[y_j = \neg y_i]\hat{c}_j\sigma_{\mathcal{X}}(\mathbf{x}_j,\mathbf{x}_i)}{\sum_{j\in\mathcal{D}} \mathbb{1}[s_j = \neg s_i]\hat{c}_j\sigma_{\mathcal{X}}(\mathbf{x}_j,\mathbf{x}_i)} \in [0,1], \quad (3)$$

which can be seen as an realization of our Definition 3.1 for sample bias: we consider a sample is biased if its similar (with high $\sigma_{\mathcal{X}}(\mathbf{x}_j,\mathbf{x}_i)$) samples from the other sensitive group ($\mathbb{1}[s_j = \neg s_i]$) receive different ($\mathbb{1}[y_j = \neg y_i]$) and credible (with high $\hat{c}_j$) treatments. Similarly, the sample credibility can be defined as the true probability that sample $(\mathbf{x}_i, s_i)$ should having label y_i:

$$c_i := \Pr[Y = y_i \mid S = s_i, X = x_i]. \quad (4)$$

A larger c_i indicates that the label of this sample is more consistent with the underlying data distribution, and therefore has higher credibility. We can estimate c_i in a similar way by solving

$$\hat{c}_i := \arg\min_{c\in\mathbb{R}} \sum_{j\in\mathcal{D}} \mathbb{1}[s_j = s_i]\sigma_{\mathcal{X}}(\mathbf{x}_j,\mathbf{x}_i)(c - \mathbb{1}[y_j = y_i])^2. \quad (5)$$

The solution of Eq. (5) gives our credibility estimation:

$$\hat{c}_i = \frac{\sum_{j\in\mathcal{D}} \mathbb{1}[s_j = s_i]\mathbb{1}[y_j = y_i]\sigma_{\mathcal{X}}(\mathbf{x}_j,\mathbf{x}_i)}{\sum_{j\in\mathcal{D}} \mathbb{1}[s_j = s_i]\sigma_{\mathcal{X}}(\mathbf{x}_j,\mathbf{x}_i)} \in [0,1], \quad (6)$$

which is also well-aligned with our sample credibility criterion given in Definition 3.2: a sample is credible if its similar (with high $\sigma_{\mathcal{X}}(\mathbf{x}_j,\mathbf{x}_i)$) samples from the same group ($\mathbb{1}[s_j = s_i]$) received same ($\mathbb{1}[y_j = y_i]$) treatment.

Interpreting the AIM Bias Score. As mentioned earlier, and as readily observed from the bias criterion (Definition 3.1) and estimation (Eq. (3)), our derived AIM sample bias score is self-explanatory: the bias score $\hat{b}$ of each sample can be naturally explained by the corresponding samples that have contributed to $\hat{b}$. Specifically, as implicated by Eq. (3), given the bias score $\hat{b}_i$ of a sample $(\mathbf{x}_i, y_i, s_i)$, the bias contribution of sample $(\mathbf{x}_j, y_j, s_j)$ to $\hat{b}_i$ is

$$\hat{b}^{\texttt{contr}}_{ij} = \frac{\mathbb{1}[s_j = \neg s_i]\mathbb{1}[y_j = \neg y_i]\hat{c}_j\sigma_{\mathcal{X}}(\mathbf{x}_j,\mathbf{x}_i)}{\sum_{j\in\mathcal{D}} \mathbb{1}[s_j = \neg s_i]\hat{c}_j\sigma_{\mathcal{X}}(\mathbf{x}_j,\mathbf{x}_i)} \in [0,1]. \quad (7)$$

Intuitively, for sample i from group A, a sample j from group B has large bias contribution $b^{\texttt{contr}}_{ij}$ if it is *highly similar* to i (with high $\sigma_{\mathcal{X}}(\mathbf{x}_j,\mathbf{x}_i)$) and *highly credible* (with large $\hat{c}_j$), yet received *different treatment* ($\mathbb{1}[y_j = \neg y_i]$). Practitioners can discover and audit discrimination present in the data by examining the bias score and interpretation of each sample. We conduct experiments and case studies on real-world data in Section 4.2 to validate the quality and soundness of AIM bias attribution and interpretation results.

Practical Similarity Computation. We now discuss how to determine the similarity metric $\sigma_{\mathcal{X}}(\cdot,\cdot) : \mathcal{X}\times\mathcal{X} \mapsto [0,1]$ in practice. In principle, *any similarity measure that satisfies the above definition can be seamlessly integrated with our framework*. However, finding an appropriate similarity metric is not always easy, as real-world data can exhibit complex structure in heterogeneous feature space that contains both numerical (e.g., age, income) and categorical (e.g., residence, occupation) values. It often requires human experts to design task-specific similarity functions based on domain knowledge, or to directly judge the similarity of sample pairs in the data, both incurring significant costs in practice [20, 36]. To address this, we present an *intuitive* and *practical* similarity measure that requires *minimum user input* based on two key ideas: (i) creating a comparability graph to capture the local similarity between input samples; and (ii) applying a graph proximity measure on the comparability graph to capture the global similarity that reflects the manifold structure of the input data.

To start with, we first define the comparability between samples by limiting the maximum allowed disparity in numerical/categorical features. Let $\mathbf{r}_\mathbf{x}/\mathbf{d}_\mathbf{x}$ represent the numerical/categorical part of feature vector $\mathbf{x}$, and given user-defined numerical/categorical disparity thresholds $t_r, t_d > 0$, we define sample comparability and the comparability constraint $\Psi_{t_r,t_d} : \mathcal{X}\times\mathcal{X} \mapsto \{0,1\}$ as follows:

Definition 3.4 (*Sample Comparability*)**.** Two samples $\mathbf{x}_1$ and $\mathbf{x}_2$ have comparability under thresholds t_r and t_d if both holds: (i) the disparity in any numerical feature is smaller than or equal to t_r, and (ii) at most t_d categorical features are different. Formally, the above conditions can be write as a comparability constraint function:

$$\Psi_{t_r,t_d}(\mathbf{x}_1,\mathbf{x}_2) = \begin{cases} 1 & \text{if } \begin{array}{l} \Pi_{i=1}^{n_r} \mathbb{1}[|\mathbf{r}^{(i)}_{\mathbf{x}_1} - \mathbf{r}^{(i)}_{\mathbf{x}_2}| \le t_r] = 1 \\ \text{and } \Sigma_{i=1}^{n_d} \mathbb{1}[\mathbf{d}^{(i)}_{\mathbf{x}_1} \neq \mathbf{d}^{(i)}_{\mathbf{x}_2}] \le t_d; \end{array} \\ 0 & \text{otherwise.} \end{cases} \quad (8)$$

Practitioners can set Ψ_{t_r,t_d} based on the application scenario and feature importance to ensure semantic similarity among comparable samples. This realization prevents costly human evaluation for a large number of sample pairs, and avoids the complexity of finding appropriate distance functions in heterogeneous feature spaces. The sample comparability defines a comparability graph $\mathbf{A}[i,j] = \Psi_{t_r,t_d}(\mathbf{x}_i,\mathbf{x}_j), \forall 1 \le i,j \le n$ over the input data based on *local similarity*. To capture the *global similarity* that reflects the manifold structure of the input data, we further utilize a graph proximity measure. In this study, we employ random walk with restart (RWR) [44, 55] due to (i) its effectiveness in capturing the global graph structure and (ii) its flexibility in adjusting the locality of the similarity. We first remove the self-loops in $\mathbf{A}$, then with symmetric normalization $\tilde{\mathbf{W}} \leftarrow \mathbf{D}^{-\frac{1}{2}}\mathbf{A}\mathbf{D}^{-\frac{1}{2}}$ and damping factor $p \in [0,1]$, the RWR similarity matrix can be derived by $\mathbf{Q} = (1-p)(\mathbf{I} - p\tilde{\mathbf{W}})^{-1}$ [44],

Algorithm 1 AIM: `Unfairness Attribution`

Input: Dataset $\mathcal{D} : \{(\mathbf{x}_i, y_i, s_i)|i = 0, 1, \cdots, n\}$, Comparability Constraint $\Psi : \mathcal{X} \times \mathcal{X} \mapsto \{0, 1\}$, Damping Factor $p \in [0, 1]$;
$\mathbf{A} \leftarrow [\Psi(\mathbf{x}_i, \mathbf{x}_j)]_{1 \le i,j \le n}$ (construct comparable graph)
$\tilde{\mathbf{W}} \leftarrow \mathbf{D}^{-\frac{1}{2}}\mathbf{A}\mathbf{D}^{-\frac{1}{2}}$ (symmetric normalization)
$\mathbf{Q} \leftarrow (1-p)(\mathbf{I} - p\tilde{\mathbf{W}})^{-1}$ (obtain similarity by RWR [55])
for $i = 1$ to n **do**
    $\hat{c}_i \leftarrow \frac{\sum_{j\in\mathcal{D}} \mathbb{1}[s_j=s_i]\mathbb{1}[y_j=y_i]\mathbf{Q}[i,j]}{\sum_{j\in\mathcal{D}} \mathbb{1}[s_j=s_i]\mathbf{Q}[i,j]}$ (estimate credibility);
end for
for $i = 1$ to n **do**
    $\hat{b}_i \leftarrow \frac{\sum_{j\in\mathcal{D}} \mathbb{1}[s_j=\neg s_i]\mathbb{1}[y_j=\neg y_i]\hat{c}_j\mathbf{Q}[i,j]}{\sum_{j\in\mathcal{D}} \mathbb{1}[s_j=\neg s_i]\hat{c}_j\mathbf{Q}[i,j]}$ (estimate bias);
end for
Return: The sample bias vector $\hat{\mathbf{b}} : [\hat{b}_1, \hat{b}_2, \cdots \hat{b}_n]$;

a smaller p means higher restart probability and thus more locality. We refer the readers to [44, 55] and references therein for more details on the properties of RWR. With this practical similarity measure, we can use $\mathbf{Q}[i, j]$ as $\sigma_{\mathcal{X}}(\mathbf{x}_j, \mathbf{x}_i)$. We now summarize the process of AIM unfairness attribution in Algorithm 1.

3.3 AIM for Unfairness Mitigation

Our bias attribution framework can also facilitate unfairness mitigation. We introduce two strategies for mitigating unfairness through informed minimal data editing: unfairness removal (AIM_{REM}) and fairness augmentation (AIM_{AUG}). By removing a small fraction of samples exhibiting high bias or augmenting samples with low bias, these methods can effectively *mitigate both group and individual unfairness* while incurring *minimal to zero loss in predictive utility*.

AIM_{REM}: Unfairness Removal. The first intuitive approach to mitigating data unfairness is simply to delete samples from the dataset that exhibit high bias (i.e., carry historical discriminatory information). This can be achieved by simply sorting the training samples $\hat{b}_i$ and removing the top-K samples with the highest bias scores from the training set. Additionally, considering the inevitable information loss from discarding samples, to achieve fairness with minimal sample removal while also alleviating the impact of class imbalance, we adaptively select a subgroup for removal based on the class distribution. Specifically, we remove majority class samples to alleviate class imbalance. If the positive class (i.e., favorable treatment) is the majority, we select samples for removal from the privileged group (e.g., gender/race with favoritism), and vice versa. Users can control the number of removed samples through a sample removal budget k. This approach is straightforward to implement and requires little additional computational cost.

AIM_{AUG}: Fairness Augmentation. Despite the simplicity and effectiveness of AIM_{REM}, it may still lead to some potential information loss. Thus we further propose an augmentation-based approach to promote fairness. Specifically, instead of discarding unfair samples, we suggest synthesizing more fair data instances through neighborhood mixup. This approach can augment the pattern of fair samples, compelling the model to focus more on learning the fair patterns, and thus mitigating unfairness without deleting information from the original data. Similarly to the above, we augment the minority class in order to alleviate class imbalance. If the positive class (i.e., favorable treatment) is the minority, we select fair samples from the protected group (i.e., gender/race being discriminated) for augmentation, and vice versa.

Existing research indicates that simple perturbation (such as adding Gaussian noise or arbitrarily changing categorical features) may generate unrealistic samples that escape the data manifold [36], e.g., here we quote a good example from [36]: "sample with age 5 or 10 but holding a doctoral degree or getting $80K annual income". To ensure the semantic coherence of synthetic samples, we propose neighborhood-based mixup for sample synthesis. Specifically, we first use $1 - \hat{b}_i$ as a weight (where low-bias samples have high weights) to randomly select a fair seed sample $(\mathbf{x}_s, y_s, s_s)$ for augmentation. Then, based on the similarity $\mathbf{Q}$, we choose the most similar n same-group samples and randomly select one, say $(\mathbf{x}_t, y_t, s_t)$, as the mixup target. Subsequently, we sample the mixup weight $\lambda \sim \text{Uniform}(0, 1)$. Denoting the synthetic sample as $(\mathbf{x}^*, y^*, s^*)$, it has the same group membership and label as the seed, i.e., $y^* = y_s, s^* = s_s$. For numerical features, we simply perform linear mixup $\mathbf{r}_{x^*} = \lambda\mathbf{r}_{x_s} + (1-\lambda)\mathbf{r}_{x_t}$. While for each categorical feature $\mathbf{d}^{(i)}$, we sample value from seed/target w.r.t. a Bernoulli distribution, i.e., $\mathbf{d}^{(i)}_{x^*} = \mathbf{d}^{(i)}_{x_s}$ with probability λ, and $\mathbf{d}^{(i)}_{x^*} = \mathbf{d}^{(i)}_{x_t}$ with probability $1 - \lambda$. It is worth noting that seed and target are similar to each other, meaning that the disparity between each numerical attributes are small and only a few categorical features are different. Such neighborhood-based mixup ensures the semantic coherence of the synthetic fair samples. We validate on real-world data in Section 4.1 that both AIM_{REM} and AIM_{AUG} can mitigate group and individual unfairness with *minimal or zero predictive utility loss*.

4 Experiments and Analysis

In this section, we conduct experiments on real-world datasets to answer the following research questions.

- **RQ1 (mitigation)**: To what extent can AIM alleviate various forms of discrimination against groups/individuals?
- **RQ2 (attribution)**: Can AIM capture sample biases encoding unfair/discriminatory aspects in the data?
- **RQ3 (interpretation)**: How can AIM provide intuitive and reasonable explanations for attributed sample biases?

We first introduce the datasets, experiment protocol, and baselines, and then present the empirical results and corresponding analysis.

Datasets. We conduct experiments on census dataset *Adult* [34], criminological dataset *Compas* [4], educational dataset *LSA* [50] (Law School Admission), and medical dataset *MEPS* [16] (Medical Expenditure Panel Survey) to validate the effectiveness of the proposed AIM framework in various application domains. For each dataset, we choose one or two attributes related to ethics as sensitive attributes that exhibit significant group and individual unfairness in standard training. More details can be found in Appendix A.1.

Experiment Protocol. To obtain reliable results, a 5-fold cross-validation is employed and we report the average test score to eliminate randomness. In each run, 3/1/1 folds of the data are used as the training/validation/test set (i.e., 60%/20%/20% split). We transform the categorical features into one-hot encoded features, and standardize the numerical features into the range of $[0, 1]$. We compare

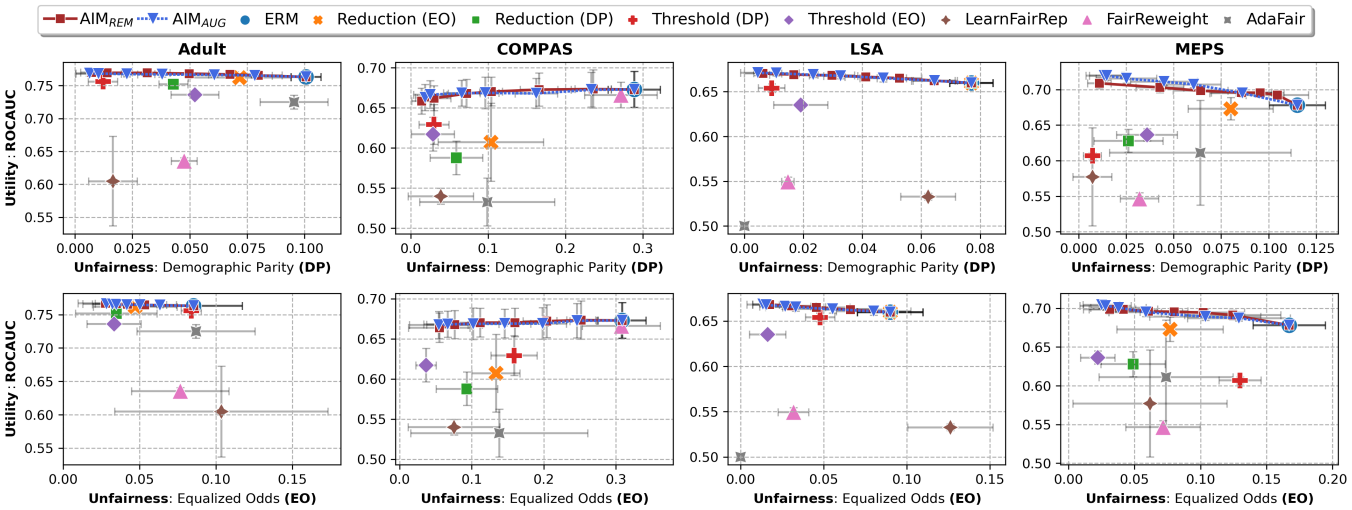

Figure 2: Compare AIM$_{\text{REM}}$ and AIM$_{\text{AUG}}$ with group fairness baselines. We show the trade-off between utility (x-axis) and unfairness metrics (y-axis) on 4 real-world FairML tasks. Results close to the *upper-left corner have better trade-offs, i.e., with low unfairness (x-axis) and high utility (y-axis).* Each column corresponds to a FairML task, and each row corresponds to a utility-unfairness metric pair. As AIM's utility-unfairness trade-off can be controlled by the sample removal/augmentation budget, we show its performance with line plots. We show error bars for both utility and unfairness metrics.

AIM with various FairML methods proposed for group/individual fairness, considering their ability to mitigate unfairness while maintaining predictive utility. For utility, we consider the area under the Receiver Operating Characteristic Curve (**ROC**) and Average Precision (**AP**) for unbiased utility evaluation due to class imbalance in occupation proportions in the data. Four popular measures of group and/or individual (un)fairness are used. For group fairness, we adopt the widely used Demographic Parity (**DP**) [17] and Equalized Odds (**EO**) [26]. For individual fairness, we use Prediction Consistency (**PC**) following Yurochkin et al. [60], Yurochkin and Sun [61]. It measures the sensitivity of model to individual's group membership by testing whether $\Pr[\hat{Y}|X=x, S=s] = \Pr[\hat{Y}|X=x, S=\neg s]$ for each test instance. This is also known as test fairness [41] or predictive parity [12]. We also adopt Generalized Entropy (**GE**) [53], a comprehensive metric that measures group and individual unfairness simultaneously with inequality indices.

Baselines. We have the following 10 FairML baselines: (i) **Reduction** [2] reduces fair classification to a sequence of cost-sensitive classification problems, returning the classifier with the lowest empirical error subject to fair constraints. (ii) **Threshold** [26] applies group-specific thresholds that optimize predictive performance while subjecting to the group fairness constraints. (iii) **FairReweight** [30] weights the examples in each (group, label) combination differently to ensure fairness before classification. (iv) **AdaFair** [29] is an ensemble learning algorithm based on AdaBoost, it takes the fairness into account in each boosting round. (v) **LearnFairRep** [62] finds a latent representation which encodes the data well but obfuscates information about protected attributes. (vi) Sensitive Subspace Robustness (**SenSR**) [60] is an individual fairness algorithm based on Distributionally Robust Optimization (DRO). It finds a sensitive subspace which encodes the sensitive information most, and generates perturbations on this sensitive subspace during optimization. (vii) Sensitive Set Invariance (**SenSeI**) [61] is also based on DRO. It involves distances penalties on both input and model predictions to construct perturbations for training individually fair model. (viii) **FairMixup** [42] transforms fairness objectives into differentiable terms and optimizes them using gradient descent. (ix) Adversarial Debiasing (**AdvFair**) [1, 64] learns a classifier maximizing prediction ability while simultaneously minimizing an adversary's ability to predict sensitive attributes from predictions. (x) Finally, **HSIC** [5] minimizes the Hilbert-Schmidt Independence Criterion between prediction accuracy and sensitive attributes. We consider logistic regression and neural network as base models in our experiments. We use scikit-learn [46] to implement logistic regression. DRO/gradient-based FairML methods (e.g., [1, 5, 42, 60, 61]) that do not compatible with this pipeline will be validated with neural networks implemented with PyTorch [45]. More implementation details can be found in Appendix A.2.

4.1 AIM for Unfairness Mitigation

AIM for group fairness (RQ1). We first compare AIM with five FairML baselines that mitigate group fairness: Reduction, Threshold, FairReweight, AdaFair, and Learn Fair Representation. For a comprehensive evaluation, we show the utility-fairness trade-off between ROC and group unfairness (DP, EO) on 4 real-world FairML tasks from different domains with race being the sensitive attribute. Note that Reduction and Threshold require specifying the group unfairness constraint for optimization, thus we report their performance with DP and EO as target, respectively. We use line plots to present the performance of AIM$_{\text{REM}}$/AIM$_{\text{AUG}}$ with different sample removal/augmentation budget which controls the trade-off of AIM between utility and fairness. The results are detailed in Figure 2. More empirical results with additional utility metrics on more FairML tasks can be found in Appendix B.

Table 1: Compare AIM_{REM} and AIM_{AUG} with individual fairness baselines. We include 3 utility and 3 (un)fairness metrics, with ↑/↓ denoting higher/lower is better. For clarity, we use double/single-underline/bold to highlight the 1st/2nd/3rd best results.

Task	Method	Utility Metrics						Unfairness Metrics					
								Unified		Group		Individual	
		Acc ↑	Δ	ROC ↑	Δ	AP ↑	Δ	GE ↓	Δ	EO ↓	Δ	PC ↑	Δ
gender	Base	84.42±0.35	-	75.68±1.51	-	53.14±1.23	-	8.49±0.37	-	11.56±5.60	-	93.90±0.81	-
	LFR	80.83±3.18	-4.3%	66.89±9.49	-11.6%	42.65±10.07	-19.7%	11.34±2.89	+33.5%	13.52±15.32	+16.9%	97.73±1.59	+4.1%
	SenSR	82.65±0.55	-2.1%	72.76±1.74	-3.9%	48.64±1.71	-8.5%	9.59±0.49	+12.9%	16.33±3.12	+41.3%	**99.92**±0.05	+6.4%
	SenSEI	83.07±0.32	-1.6%	72.46±0.87	-4.3%	49.23±0.83	-7.4%	**9.52**±0.23	+12.0%	15.48±6.42	+33.9%	97.61±0.83	+4.0%
	AdvFair	**84.05**±0.32	-0.4%	**75.23**±1.03	-0.6%	**52.26**±0.78	-1.7%	13.91±1.35	+63.7%	11.73±7.06	+1.5%	92.33±1.06	-1.7%
	FairMixup	82.66±0.39	-2.1%	72.26±1.54	-4.5%	48.49±0.57	-8.7%	14.37±0.34	+69.2%	**10.09**±4.44	-12.7%	97.34±0.50	+3.7%
	HSIC	83.45±0.28	-1.1%	74.70±0.97	-1.3%	50.94±0.93	-4.1%	14.09±0.07	+65.8%	10.83±3.02	-6.3%	97.55±0.58	+3.9%
	AIM_{REM} (Ours)	**83.71**±0.43	-0.8%	**78.25**±1.36	+3.4%	**53.28**±1.21	+0.3%	**8.12**±0.38	-4.4%	**7.30**±2.38	-36.8%	**98.24**±0.19	+4.6%
	AIM_{AUG} (Ours)	**84.37**±0.45	-0.1%	**77.48**±0.97	+2.4%	**53.89**±0.99	+1.4%	**8.15**±0.26	-4.0%	**7.64**±1.80	-33.9%	**98.69**±0.29	+5.1%
race	Base	84.38±0.57	-	75.70±2.32	-	53.08±2.00	-	8.50±0.58	-	9.59±4.13	-	98.16±0.65	-
	LFR	80.29±2.88	-4.9%	66.56±9.39	-12.1%	41.64±9.54	-21.6%	11.56±2.78	+36.0%	6.81±4.32	-29.0%	93.20±4.09	-5.1%
	SenSR	82.69±0.31	-2.0%	71.56±0.57	-5.5%	48.13±0.66	-9.3%	9.81±0.16	+15.4%	**6.02**±3.92	-37.2%	**99.91**±0.10	+1.8%
	SenSEI	83.07±0.31	-1.6%	72.67±1.11	-4.0%	49.33±0.85	-7.1%	**9.47**±0.25	+11.5%	10.87±2.82	+13.4%	98.36±0.63	+0.2%
	AdvFair	**84.61**±0.49	+0.3%	**77.05**±1.30	+1.8%	**54.11**±1.25	+1.9%	14.58±0.44	+71.6%	7.77±5.30	-19.0%	92.99±0.85	-5.3%
	FairMixup	82.73±0.47	-2.0%	70.91±2.08	-6.3%	47.94±1.69	-9.7%	14.44±0.46	+69.9%	7.40±3.22	-22.8%	97.12±1.39	-1.1%
	HSIC	83.42±0.43	-1.1%	75.88±0.74	+0.2%	51.52±0.96	-2.9%	13.94±0.08	+64.0%	7.12±2.49	-25.7%	99.10±0.40	+1.0%
	AIM_{REM} (Ours)	**84.45**±0.56	+0.1%	**77.06**±1.57	+1.8%	**53.81**±1.62	+1.4%	**8.22**±0.44	-3.2%	**6.53**±2.96	-31.9%	**99.21**±0.25	+1.1%
	AIM_{AUG} (Ours)	**84.38**±0.47	-0.0%	**77.97**±0.56	+3.0%	**54.17**±0.86	+2.0%	**8.05**±0.17	-5.3%	**6.37**±2.47	-33.6%	**99.27**±0.18	+1.1%

We summarize the key observations as follows: **(i)** *AIM can mitigate unfairness with minimal/zero utility cost.* Across all settings, AIM achieves the optimal trade-off compared to other group fairness baselines: it either outperforms or matches the best baseline in terms of utility-fairness trade-off (close to the upper-left corner in Figure 2). **(ii)** Since AIM_{REM} and AIM_{AUG} are designed to promote fairness while balancing class distribution, on datasets with significant class imbalance (e.g., *Adult*, *LSA*, *MEPS* with 24.8/27.9/17.2% positive samples), they can *mitigate unfairness without sacrificing predictive performance, and in some cases, may even enhance it* (e.g., ROC on the *MEPS* dataset). **(iii)** At the same fairness level, AIM_{AUG} generally exhibits better classification performance compared to AIM_{REM} as it retains the original training set and promote fairness by adding augmented data. We have similar observations in the following experiments on individual fairness. However, we note that AIM_{AUG} requires a higher sample manipulation budget and computational cost than AIM_{REM}. **(iv)** We notice that some baselines may worsen specific fairness metrics. For example, LearnFairRep can help reduce DP but increase EO in the *Adult* and *LSA* tasks. This is due to the potential incompatibility between fairness notions, and we refer readers to [33, 51] for more related discussions.

AIM for individual fairness (RQ1). We further verify the effectiveness of AIM in promoting individual fairness (IF). 6 FairML baselines that mitigate individual unfairness are included: LearnFairRep (LFR), SenSR, SenSEI, AdvFair, FairMixup, HSIC. Since these methods generally rely on gradient-based optimization, we use neural networks as the base ML model in this section. Following the existing literature [60, 61], We test them on the widely used *Adult* dataset, with gender and race being the sensitive attribute, respectively. To ensure a comprehensive evaluation, we adopt three utility metrics (ACC, ROC, AP) and three metrics for individual and/or group (un)fairness: PC for individual, EO for group, and GE for both. For AIM_{REM} and AIM_{AUG}, we select the sample removal/augmentation budget that maximizes PC (individual fairness) on the validation set. The results are detailed in Table 1.

We summarize our findings as follows: **(i)** AIM_{REM} *and* AIM_{AUG} *simultaneously promote both individual and group fairness.* Compared to the IF-targeted FairML algorithms, AIM demonstrates competitive performance in mitigating individual unfairness, while concurrently achieving better group fairness. **(ii)** AIM *exhibits significantly less (if any) utility losses.* Baseline methods often lead to significant performance declines, whereas AIM achieves minimal utility loss and, in some cases, even enhances certain performance metrics due to its ability to alleviate class imbalance. Taking *Adult*-gender as an example, compared to baseline methods with a relative ACC loss of 1.6-4.3%, AIM incurs only 0.1-0.8% loss. This advantage becomes even more pronounced in ROC/AP, e.g., FairML baselines result in an AP loss of 7.4-19.7%, while AIM brings about a gain of 0.3-1.4%. This align with our earlier experiments on group fairness. **(iii)** *FairML methods devised for individual fairness can potentially lead to group unfairness.* For instance, on *Adult*-gender, while LFR/SenSR/SenSEI can promote individual PC, they also degrade group EO with an increase of 16.9/41.3/33.9%, respectively. GE that reflects "how unequally the outcomes of an algorithm benefit different individuals or groups in a population" [53] better captures this aspect: baseline methods that improve PC actually degrade GE . We refer the readers to [8, 20] for more discussions on the trade-offs and conflicts between individual and group fairness.

Finally, we notice that the majority of existing FairML methods are designed for a specific learning task and/or fairness metric(s), thus optimizing one metric might be at the (unintentional) expense of another metric. In contrast, as demonstrated by the experiments above, AIM avoids such intrinsic tension by focusing on the root of unfairness (i.e., historically biased samples) and thus can lead to a near-universal improvement across different metrics.

4.2 AIM for Bias Attribution and Interpretation

In this section, we conduct experiments and case studies to show the soundness of AIM in unfairness attribution and interpretation.

AIM identifies discriminatory samples (RQ2). We now validate the soundness of AIM unfair attribution by verifying whether the high-bias samples identified by AIM encode discriminatory information from the data, and contribute to the unfairness in model predictions. Specifically, we remove varying quantities of samples with high to low bias scores by AIM_{REM} from the training dataset of *Compas* and observe the (un)fairness metrics of the model trained on the modified data. We compare these results with a naïve method that randomly remove samples from training data. Results are shown in Figure 3. To ensure a comprehensive evaluation, we include DP and EO for group unfairness, Predictive Inconsistency (PI, i.e., 1 - PC) for individual unfairness, and GE for both. It can be observed that randomly removing samples does not alleviate the model's unfairness. At the same time, removing an equal number of high-bias samples identified by AIM significantly reduces the encoded discriminatory information in the data, and effectively promotes both group and individual fairness in model predictions. This verifies the rationality of AIM unfair attribution.

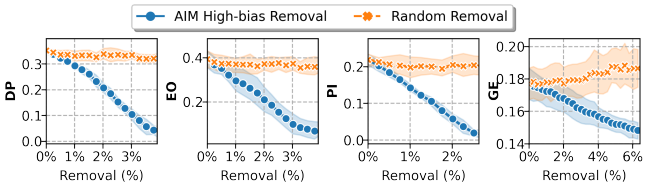

Figure 3: Evaluation of the AIM bias attribution quality. Removing high-bias samples identified by AIM from the data greatly reduces the discrimination in the model prediction.

Visual evaluation of AIM bias attribution (RQ2). To intuitively demonstrate the bias attribution ability of AIM, we further design a series of synthetic datasets with different types of bias for visual evaluation of AIM's ability in capturing group- and individual-level unfairness. In each dataset, we sample two groups from the same distribution, with group #1 as the reference group, and introduce group/individual-level discrimination to group #2. For group unfairness, we altered the class boundary for the target group to simulate group-based discrimination (different groups having different "thresholds" for positive outcomes). For individual unfairness, we randomly selected 10% of samples from the target group and flipped their labels to simulate discrimination against specific individuals (similar individuals not being treated similarly). This approach provides ground truth labels for each sample's bias status, enabling us to visualize the quality of AIM bias attribution.

Figure 4 presents our experimental results. Each row in the figure represents a synthetic dataset, and we label the bias type of the data at the beginning of each row. For each dataset, the 1st and 2nd columns display the data and class distribution of groups #1 and #2, respectively. The 3rd column shows the ground truth bias distribution on the target group (#2), where blue indicates unbiased and red indicates biased. In a similar manner, the 4th column displays the sample bias distribution detected by AIM, with red indicating biased samples (with AIM bias score > 0.5). We also annotate the accuracy of AIM-detected biased samples w.r.t. the ground truth in the 4th-column subplots. It can be observed that AIM accurately detects group/individual-level bias in the data, with very high bias detection accuracy ranging from 97.0% to 99.8%.

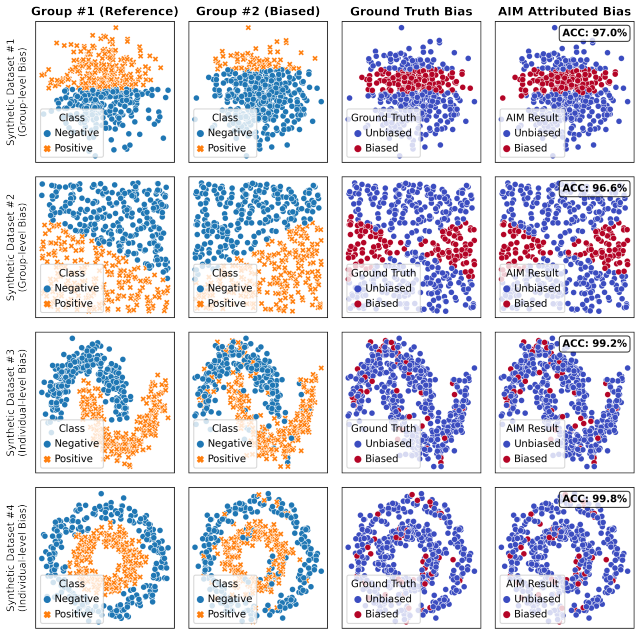

Figure 4: Synthetic bias detection results. AIM (4th column) can accurately detect ground-truth biased samples (3rd columns) under both group- and individual-level unfairness.

AIM reflects the level of discrimination of a dataset (RQ2). From a macro perspective, we further show that the outcomes of bias attribution can also offer insights into how "discriminatory" a dataset is. We inherit the four metrics used in Figure 3, but examine the correlation between the average AIM sample bias scores on the training set and the predictive unfairness of the model on the test set, as shown in Figure 5. It can be observed that the average AIM sample bias score is a good indicator of the level of dataset discrimination, especially in terms of the comprehensive metric GE that captures both group and individual unfairness. This further validates the soundness of AIM bias attribution.

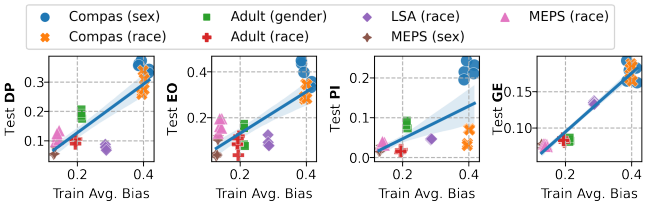

Figure 5: AIM bias score reflect dataset unfairness level. Each dot denotes a combination of datasets, sensitive attributes, and train/test split. We report average training sample bias (x-axis) and test unfairness of the model predictions (y-axis).

Table 2: Case study of AIM bias attribution and interpretation on the *Adult*-gender task.

Sample Type	Individual Feature Attributes									Sensitive	Label	Bias/Contrib	Explanation	Similarity
	Age	Hours	CGain	Edu	Education	MaritalStatus	Occupation	WorkClass	Country	Sex	Income	$\hat{b}$ / $\hat{b}^{\texttt{contr}}$	Credibility $\hat{c}$	to Query
Query	32	45	5013	13	Bachelors	Married-civ-spouse	Prof-specialty	Local-gov	United-States	Female	0	0.9959	-	-
Explanation	29	45	5178	13	Bachelors	Married-civ-spouse	Prof-specialty	Private	United-States	Male	1	0.3089	0.9027	0.3104
	35	45	7298	13	Bachelors	Married-civ-spouse	Prof-specialty	Self-emp-inc	United-States	Male	1	0.2995	0.9916	0.2740
	35	45	7298	13	Bachelors	Married-civ-spouse	Prof-specialty	Private	United-States	Male	1	0.2066	0.9822	0.1908
	33	47	3103	13	Bachelors	Married-civ-spouse	Prof-specialty	Private	United-States	Male	1	0.0455	0.7136	0.0579
	33	43	3103	13	Bachelors	Married-civ-spouse	Prof-specialty	Private	United-States	Male	1	0.0317	0.6546	0.0439
Query	34	40	5178	10	Some-college	Married-civ-spouse	Prof-specialty	Private	United-States	Male	1	0.7091	-	-
Explanation	35	40	7443	10	Some-college	Divorced	Prof-specialty	Private	United-States	Female	0	0.4644	0.9875	0.3114
	31	40	3908	10	Some-college	Married-civ-spouse	Prof-specialty	Private	United-States	Female	0	0.2130	0.5522	0.2555
	38	44	5721	10	Some-college	Divorced	Adm-clerical	Private	United-States	Female	0	0.0126	0.8419	0.0099
	32	40	2597	10	Some-college	Married-civ-spouse	Exec-managerial	Private	Japan	Female	0	0.0018	0.7789	0.0015
	38	40	7443	10	Some-college	Divorced	Adm-clerical	Private	United-States	Female	0	0.0016	0.5340	0.0019

* Hours: working hours per week; CGain: capital gain; Edu: education-num. We show key features with differences here due to space limitation, please refer to [34] for detailed semantics of each feature.

AIM provides reasonable sample bias explanation (RQ3). Finally, we provide case studies on real-world data to intuitively demonstrate the validity of AIM's bias attribution and explanation. In Table 2, we present two high-bias samples detected by AIM from different genders in the *Adult-Gender* task, along with their corresponding top-5 explanations (i.e., the top 5 samples contributing most to their bias). It can be observed that for high-bias samples, AIM can retrieve samples from another group that are similar but have different and credible labels, and quantify their contributions to the bias ($\hat{b}^{\texttt{contr}}$) as explanations.

For example, the bias score of the first query (a female with a negative label) is largely contributed by the first male sample in line 2, who has highly similar attributes but also a different and highly credible label (with $\hat{c} = 0.9027$). Similarly, for the second query sample (a male with a positive label), it is considered biased because similar females have negative labels. This also underscores another practical implication of the AIM bias score: for a sample receiving positive/negative treatment, a high bias score suggests that it may have received an *unfair advantage/disadvantage due to its group membership compared to individuals with similar attributes*. Such sample-level explanation can assist human experts inspect and understand discrimination present in the data, and provide insights into how to design fairer decision criteria in the future.

5 Related Works

Fair Machine Learning. FairML advocates for ethical regulations to rectify algorithms, ensuring non-discrimination against any group or individual [6, 10, 41]. The concept of group fairness (GF) seeks equalized outcomes across sensitive groups concerning statistics like positive rate [26]. Although intuitive, GF falls short in ensuring fairness on an individual level [17, 27]. Individual fairness (IF) is thus proposed, with its main idea being that similar samples should receive similar treatment [17]. However, due to the absence of group constraints, IF cannot capture systematic bias against groups [20]. Counterfactual fairness (CF) examines the consistency of algorithms on a single instance and its counterfactuals when sensitive attributes are altered [35]. However, this notion and its evaluations heavily depend on the causal structure rooted in the data generation process, thus explicit modeling is usually impractical [36]. AIM is grounded on the existing fairness notions and can practically capture various prejudices encoded in the data.

Discrimination Discovery. There are a few works on discrimination discovery, but their scope significantly differs from ours. The discrimination discovery in [48] is aimed at systems based on classification (association) rules. They propose an extended lift measure to assess whether a classification rule might lead to unlawful discrimination against groups protected by law, and further propose a system for discrimination discovery in database [49]. Apparently, this type of discrimination discovery cannot be generalized to non-rule-based ML models. Another line of work [65, 66] models the direct/indirect discrimination as the path-specific effects on the causal network. Like counterfactual fairness [35], this also requires explicit modeling of the causal structure and data generation process. Additionally, their discrimination discovery is achieved by computing a score for the entire dataset, where a high score indicates that the dataset as a whole exhibits discrimination. This is fundamentally different from our sample-level bias attribution.

6 Conclusion

In this work, we investigate the problem of identifying samples carrying historical biases in training data. Building on existing fairness notions, we establish a criterion and propose practical algorithms for measuring and countering sample bias. We propose a practical framework AIM, which supports (i) sample-level bias attribution, (ii) intuitive explanation of the bias of each instance, and (iii) effective unfairness mitigation with minimal or zero predictive utility loss. Extensive experiments and analysis on various real-world datasets demonstrate the efficacy of our approach in attributing, interpreting, and mitigating unfairness.

Acknowledgments

This work is supported by NSF (1939725), AFOSR (FA9550-24-1-0002), the C3.ai Digital Transformation Institute, MIT-IBM Watson AI Lab, and IBM-Illinois Discovery Accelerator Institute. The content of the information in this document does not necessarily reflect the position or the policy of the Government, and no official endorsement should be inferred. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation here on.

References

[1] Tameem Adel, Isabel Valera, Zoubin Ghahramani, and Adrian Weller. 2019. One-network adversarial fairness. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 33. 2412–2420.

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

[3] Mayank Agarwal. 2023. Individual Fairness and inFairness. https://github.com/IBM/inFairness.

[4] Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. 2016. Machine bias. In *Ethics of Data and Analytics*.

[5] Sina Baharlouei, Maher Nouiehed, Ahmad Beirami, and Meisam Razaviyayn. 2019. R\'enyi Fair Inference. *arXiv preprint arXiv:1906.12005* (2019).

[6] Solon Barocas, Moritz Hardt, and Arvind Narayanan. 2023. *Fairness and machine learning: Limitations and opportunities*. MIT Press.

[7] Rachel KE Bellamy, Kuntal Dey, Michael Hind, Samuel C Hoffman, Stephanie Houde, Kalapriya Kannan, Pranay Lohia, Jacquelyn Martino, Sameep Mehta, Aleksandra Mojsilović, et al. 2019. AI Fairness 360: An extensible toolkit for detecting and mitigating algorithmic bias. *IBM Journal of Research and Development* 63, 4/5 (2019), 4–1.

[8] Reuben Binns. 2020. On the apparent conflict between individual and group fairness. In *Proceedings of the 2020 conference on fairness, accountability, and transparency*. 514–524.

[9] Sarah Bird, Miro Dudík, Richard Edgar, Brandon Horn, Roman Lutz, Vanessa Milan, Mehrnoosh Sameki, Hanna Wallach, and Kathleen Walker. 2020. Fairlearn: A toolkit for assessing and improving fairness in AI. *Microsoft, Tech. Rep. MSR-TR-2020-32* (2020).

[10] Simon Caton and Christian Haas. 2020. Fairness in machine learning: A survey. *Comput. Surveys* (2020).

[11] Eunice Chan, Zhining Liu, Ruizhong Qiu, Yuheng Zhang, Ross Maciejewski, and Hanghang Tong. 2024. Group Fairness via Group Consensus. In *The 2024 ACM Conference on Fairness, Accountability, and Transparency*. 1788–1808.

[12] Alexandra Chouldechova. 2017. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big data* 5, 2 (2017), 153–163.

[13] William S Cleveland. 1979. Robust locally weighted regression and smoothing scatterplots. *Journal of the American statistical association* 74, 368 (1979), 829–836.

[14] William S Cleveland and Susan J Devlin. 1988. Locally weighted regression: an approach to regression analysis by local fitting. *Journal of the American statistical association* 83, 403 (1988), 596–610.

[15] William S Cleveland, Eric Grosse, and William M Shyu. 2017. Local regression models. In *Statistical models in S*. Routledge, 309–376.

[16] Joel W Cohen, Steven B Cohen, and Jessica S Banthin. 2009. The medical expenditure panel survey: a national information resource to support healthcare cost research and inform policy and practice. *Medical care* (2009), S44–S50.

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

[18] Cynthia Dwork, Christina Ilvento, Guy N Rothblum, and Pragya Sur. 2020. Abstracting fairness: Oracles, metrics, and interpretability. *arXiv preprint arXiv:2004.01840* (2020).

[19] Maddalena Favaretto, Eva De Clercq, and Bernice Simone Elger. 2019. Big Data and discrimination: perils, promises and solutions. A systematic review. *Journal of Big Data* 6, 1 (2019), 1–27.

[20] Will Fleisher. 2021. What's fair about individual fairness?. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*. 480–490.

[21] Benoît Frénay and Michel Verleysen. 2013. Classification in the presence of label noise: a survey. *IEEE transactions on neural networks and learning systems* 25, 5 (2013), 845–869.

[22] Runshan Fu, Yan Huang, and Param Vir Singh. 2021. Crowds, lending, machine, and bias. *Information Systems Research* 32, 1 (2021), 72–92.

[23] João Gama, Indrė Žliobaitė, Albert Bifet, Mykola Pechenizkiy, and Abdelhamid Bouchachia. 2014. A survey on concept drift adaptation. *ACM computing surveys (CSUR)* 46, 4 (2014), 1–37.

[24] Heitor Murilo Gomes, Jesse Read, Albert Bifet, Jean Paul Barddal, and João Gama. 2019. Machine learning for streaming data: state of the art, challenges, and opportunities. *ACM SIGKDD Explorations Newsletter* 21, 2 (2019), 6–22.

[25] Bo Han, Quanming Yao, Tongliang Liu, Gang Niu, Ivor W Tsang, James T Kwok, and Masashi Sugiyama. 2020. A survey of label-noise representation learning: Past, present and future. *arXiv preprint arXiv:2011.04406* (2020).

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

[27] Tatsunori Hashimoto, Megha Srivastava, Hongseok Namkoong, and Percy Liang. 2018. Fairness without demographics in repeated loss minimization. In *International Conference on Machine Learning*. PMLR, 1929–1938.

[28] Steven CH Hoi, Doyen Sahoo, Jing Lu, and Peilin Zhao. 2021. Online learning: A comprehensive survey. *Neurocomputing* 459 (2021), 249–289.

[29] Vasileios Iosifidis and Eirini Ntoutsi. 2019. Adafair: Cumulative fairness adaptive boosting. In *Proceedings of the 28th ACM international conference on information and knowledge management*. 781–790.

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

[31] Jian Kang and Hanghang Tong. 2021. Fair graph mining. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*. 4849–4852.

[32] Pauline T Kim. 2016. Data-driven discrimination at work. *Wm. & Mary L. Rev.* 58 (2016), 857.

[33] Jon Kleinberg, Sendhil Mullainathan, and Manish Raghavan. 2016. Inherent trade-offs in the fair determination of risk scores. *arXiv preprint arXiv:1609.05807* (2016).

[34] Ronny Kohavi and Barry Becker. 1996. Adult Data Set.

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

[36] Peizhao Li, Ethan Xia, and Hongfu Liu. 2023. Learning antidote data to individual unfairness. In *International Conference on Machine Learning*. PMLR, 20168–20181.

[37] Zhining Liu, Wei Cao, Zhifeng Gao, Jiang Bian, Hechang Chen, Yi Chang, and Tie-Yan Liu. 2020. Self-paced ensemble for highly imbalanced massive data classification. In *2020 IEEE 36th international conference on data engineering (ICDE)*. IEEE, 841–852.

[38] Zhining Liu, Jian Kang, Hanghang Tong, and Yi Chang. 2021. IMBENS: Ensemble class-imbalanced learning in Python. *arXiv preprint arXiv:2111.12776* (2021).

[39] Zhining Liu, Pengfei Wei, Jing Jiang, Wei Cao, Jiang Bian, and Yi Chang. 2020. MESA: boost ensemble imbalanced learning with meta-sampler. *Advances in neural information processing systems* 33 (2020), 14463–14474.

[40] Zhining Liu, Zhichen Zeng, Ruizhong Qiu, Hyunsik Yoo, David Zhou, Zhe Xu, Yada Zhu, Kommy Weldemariam, Jingrui He, and Hanghang Tong. 2023. Topological Augmentation for Class-Imbalanced Node Classification. *arXiv preprint arXiv:2308.14181* (2023).

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

[42] Youssef Mroueh et al. 2021. Fair mixup: Fairness via interpolation. In *International Conference on Learning Representations*.

[43] Debarghya Mukherjee, Mikhail Yurochkin, Moulinath Banerjee, and Yuekai Sun. 2020. Two simple ways to learn individual fairness metrics from data. In *International Conference on Machine Learning*. PMLR, 7097–7107.

[44] Jia-Yu Pan, Hyung-Jeong Yang, Christos Faloutsos, and Pinar Duygulu. 2004. Automatic multimedia cross-modal correlation discovery. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*. 653–658.

[45] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. 2019. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems* 32 (2019).

[46] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. 2011. Scikit-learn: Machine learning in Python. *the Journal of machine Learning research* 12 (2011), 2825–2830.

[47] Lucas Rosenblatt and R Teal Witter. 2023. Counterfactual fairness is basically demographic parity. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 37. 14461–14469.

[48] Salvatore Ruggieri, Dino Pedreschi, and Franco Turini. 2010. Data mining for discrimination discovery. *ACM Transactions on Knowledge Discovery from Data (TKDD)* 4, 2 (2010), 1–40.

[49] Salvatore Ruggieri, Dino Pedreschi, and Franco Turini. 2010. DCUBE: Discrimination discovery in databases. In *Proceedings of the 2010 ACM SIGMOD International Conference on Management of data*. 1127–1130.

[50] Richard Sander. 2009. Law School Admissions Dataset. Project SEAPHE. http://www.seaphe.org/databases.php.

[51] Andrew D Selbst, Danah Boyd, Sorelle A Friedler, Suresh Venkatasubramanian, and Janet Vertesi. 2019. Fairness and abstraction in sociotechnical systems. In *Proceedings of the conference on fairness, accountability, and transparency*. 59–68.

[52] Jack Sherman. 1949. Adjustment of an inverse matrix corresponding to changes in the elements of a given column or a given row of the original matrix. *Annals of mathematical statistics* 20, 4 (1949), 621.

[53] Till Speicher, Hoda Heidari, Nina Grgic-Hlaca, Krishna P Gummadi, Adish Singla, Adrian Weller, and Muhammad Bilal Zafar. 2018. A unified approach to quantifying algorithmic unfairness: Measuring individual &group unfairness via inequality indices. In *Proceedings of the 24th ACM SIGKDD international conference on knowledge discovery & data mining*. 2239–2248.

[54] Harini Suresh and John V Guttag. 2019. A framework for understanding unintended consequences of machine learning. *arXiv preprint arXiv:1901.10002* 2, 8 (2019).

[55] Hanghang Tong, Christos Faloutsos, and Jia-Yu Pan. 2006. Fast random walk with restart and its applications. In *Sixth international conference on data mining (ICDM'06)*. IEEE, 613–622.

[56] Hanghang Tong, Spiros Papadimitriou, Philip S Yu, and Christos Faloutsos. 2008. Proximity tracking on time-evolving bipartite graphs. In *Proceedings of the 2008 SIAM International Conference on Data Mining*. SIAM, 704–715.

[57] Serena Wang, Wenshuo Guo, Harikrishna Narasimhan, Andrew Cotter, Maya Gupta, and Michael Jordan. 2020. Robust optimization for fairness with noisy protected groups. *Advances in neural information processing systems* 33 (2020), 5190–5203.
[58] Han Xu, Xiaorui Liu, Yaxin Li, Anil Jain, and Jiliang Tang. 2021. To be robust or to be fair: Towards fairness in adversarial training. In *International conference on machine learning*. PMLR, 11492–11501.
[59] Yuchen Yan, Lihui Liu, Yikun Ban, Baoyu Jing, and Hanghang Tong. 2021. Dynamic knowledge graph alignment. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 35. 4564–4572.
[60] Mikhail Yurochkin, Amanda Bower, and Yuekai Sun. 2020. Training individually fair ML models with sensitive subspace robustness. In *International Conference on Learning Representations*.
[61] Mikhail Yurochkin and Yuekai Sun. 2021. SenSeI: Sensitive Set Invariance for Enforcing Individual Fairness. In *International Conference on Learning Representations*.
[62] Rich Zemel, Yu Wu, Kevin Swersky, Toni Pitassi, and Cynthia Dwork. 2013. Learning fair representations. In *International conference on machine learning*. PMLR, 325–333.
[63] Zhichen Zeng, Boxin Du, Si Zhang, Yinglong Xia, Zhining Liu, and Hanghang Tong. 2024. Hierarchical multi-marginal optimal transport for network alignment. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 38. 16660–16668.
[64] Brian Hu Zhang, Blake Lemoine, and Margaret Mitchell. 2018. Mitigating unwanted biases with adversarial learning. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*. 335–340.
[65] Lu Zhang and Xintao Wu. 2017. Anti-discrimination learning: a causal modeling-based framework. *International Journal of Data Science and Analytics* 4 (2017), 1–16.
[66] Lu Zhang, Yongkai Wu, and Xintao Wu. 2017. A Causal Framework for Discovering and Removing Direct and Indirect Discrimination. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence*.

A Reproducibility

A.1 Dataset Statistics and Details

***Adult* dataset** The *Adult* dataset [34] contains census personal records with attributes like age, education, race, etc. The task is to determine whether a person makes over $50K a year. ***Compas* dataset** The *Compas* dataset [4] is a criminological dataset recording prisoners' information like criminal history, jail and prison time, demographic, sex, etc. The task is to predict a recidivism risk score for defendants. ***LSA* dataset** *LSA* (Law School Admission) dataset [50] contains admissions data from 25 law schools, features include applicant attributes like LSAT score, undergraduate GPA, residency, race, etc. The target label is the admission decision of each applicant. ***MEPS* dataset** The *MEPS* (Medical Expenditure Panel Survey) dataset [16] comprises demographic features, health status, income, and other attributes of surveyed individuals. The goal is to predict whether the individual has high medical service utilization. We use the AIF360 [7] toolbox[1] to retrieve and process all used datasets. Detailed data statistics are listed in Table 3.

A.2 Implementation Details

Baselines We implement the 10 FairML baselines using standard Python toolkits or official code base provided by the paper authors. Specifically, we use the AIF360 [7] package[1] to implement **Reduction**, **FairReweight**, **LearnFairRep**, **AdvFair**; the fairlearn [9] package[2] for implementing **Threshold**; the inFairness [3] package[3] for implementing **SenSR** and **SenSeI**; and the FBB benchmark[4] for implementing **FairMixup**, **AdvFair**, and **HSIC**. For **AdaFair**, we use the official code base[5] for implementation. Their hyperparameters are fine-tuned to get the best fairness-utility trade-off.

Models We consider logistic regression and neural network as base models in our experiments. We use scikit-learn [46] to implement logistic regression. DRO-based FairML methods (SenSR, SenSEI) that do not compatible with this pipeline are validated with neural networks implemented with PyTorch [45]. For logistic regression, we use the default parameters specified by sklearn. For neural network, we use the implementation provided in the official example in inFairness [3] for SenSR/SenSEI to guarantee fair comparison. The neural network is a three-layer MLP with 100 hidden units and ReLU activation function. We use Adam optimizer with learning rate 1e-3 and cross-entropy loss, and train the MLP for 100 epochs with a mini-batch size 32 until it is converged.

Hyperparameters AIM bias attribution criterion itself does not have any hyperparameters. Similarity computing involves three main parameters: numerical and categorical disparity thresholds t_r and t_d for determining sample comparability, and the damping factor p of RWR used for computing similarity. We choose $t_r = 0.1$ and $t_d = 2$ to offer sufficient comparable samples while maintaining sample semantic comparability. The damping factor p is set to 0.1 (i.e., restart probability for RWR = 0.9) to guarantee locality while capturing global similarity. In general, these three parameters can all be regarded as parameters that directly or indirectly determine the locality of similarity, the smaller they are, the more local the similarity is. Specifically, smaller t_r/t_d results in more strict comparability constraints (also fewer comparable sample pairs) and thus more locality, and smaller p means higher restart probability in RWR, thus also more locality in similarity. Overall, these three parameters do not affect the validity of similarity and AIM bias attribution. Users can adjust these values (or use other expert-specified similarity metrics if available) based on the application scenario.

B More Results and Discussions

Additional Results In Figure 2, we report the results with only ROC on 4 FairML tasks due to space limitation. Here we provide additional results showing the utility-fairness trade-off between 2 utility metrics (ROC, AP) and 2 unfairness metrics (DP, EO) on 7 real-world FairML tasks from different domains. Other protocols are identical to Figure 2. We report the additional results in Figure 6. It can be observed that the additional results are consistent with the conclusions derived in the paper from Figure 2. Across all 28 (2 utility metrics x 2 unfairness metrics x 7 FairML tasks) settings, AIM achieves the optimal trade-off compared to other group fairness baselines: it either outperforms or matches the best baseline in terms of utility-fairness trade-off (close to the upper-left corner). This further validates the effectiveness and generality of AIM in mitigating unfairness.

Complexity Analysis The complexity of estimating bias/credibility through Eq. (3)/Eq. (6) is $O(N^2)$. However, we note that both bias and credibility computations can be performed in matrix form, and thus can be efficiently accelerated with the parallel computing capabilities of modern GPUs. The time complexity of bias computation

[1] https://github.com/Trusted-AI/AIF360
[2] https://github.com/fairlearn/fairlearn
[3] https://github.com/IBM/inFairness
[4] https://github.com/ahxt/fair_fairness_benchmark
[5] https://github.com/iosifidisvasileios/AdaFair

Table 3: Dataset statistics. We report the quantity of samples with positive/negative labels, number of continuous features and discrete features (one-hot encoded), sensitive attribute used, as well as size and positive ratio of the privileged/protected group.

Dataset	Domain	#Samples (Positive/Negative)	Positive Ratio	#Features (Cont./One-hot)	Sensitive Attribute	Group Size (Privileged/Protected)	Positive Ratio (Privileged/Protected)
Adult	census	11,208 / 34,014	24.8%	5 / 93	gender	30,527 / 14,695	31.2% / 11.4%
					race	38,903 / 6,319	26.2% / 15.8%
Compas	criminological	2,737 / 3,138	46.6%	5 / 6	sex	4,714 / 1,161	49.2% / 36.2%
					race	3,528 / 2,347	51.3% / 39.5%
LSA	educational	15,482 / 39,966	27.9%	5 / 4	race	40,989 / 14,459	29.6% / 23.1%
MEPS	medical	2,721 / 13,118	17.2%	5 / 120	sex	8,255 / 7,584	20.8% / 13.3%
					race	5,659 / 10,180	25.6% / 12.5%

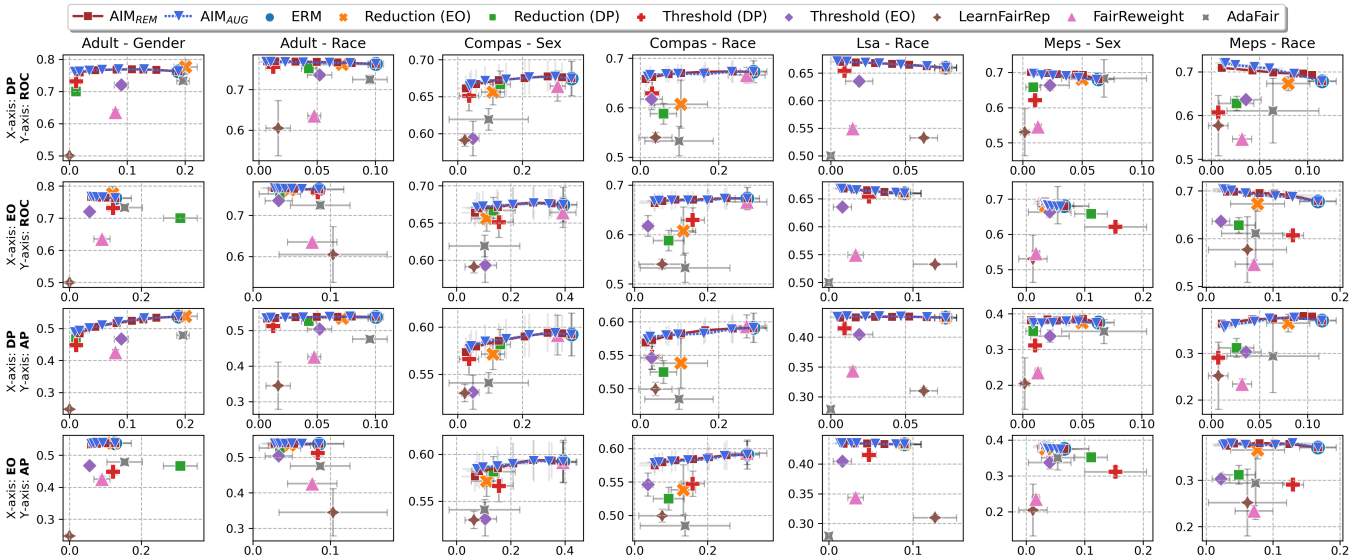

Figure 6: Compare AIM_{REM} and AIM_{AUG} with group fairness baselines. We show the utility-fairness trade-off between 2 utility metrics (x-axis) and 2 unfairness metrics (y-axis) on 7 real-world FairML tasks. Results close to the *upper-left corner have better trade-offs, i.e., with low unfairness (x-axis) and high utility (y-axis).* Each column corresponds to a FairML task, and each row corresponds to a utility-unfairness metric pair. As AIM's utility-unfairness trade-off can be controlled by the sample removal/augmentation budget, we show its performance with line plots. We show error bars for both utility and unfairness.

can be reduced to $O(\frac{N^2}{C})$, where C is the number of available computing units. The complexity of computing similarity mainly arises from the matrix inversion step in RWR. There are many techniques can be employed to accelerate the solution of RWR, such as Fast Random Walk with Restart [55] that use low-rank approximation and Sherman–Morrison Lemma [52] to approximate $(1 - p\tilde{\mathbf{W}})^{-1}$. In practice, if the dataset is large enough, and there are sufficient comparable samples to support the bias attribution and explanation for each sample, one can also consider directly using sample comparability as similarity to avoid additional computational costs.

Limitation and Future Works At the end of the paper, we discuss the limitations of AIM and possible future directions to address them. One potential limitation is the static data assumption of AIM: as society evolves and relevant laws improve, the distribution of observed data also changes [24, 28]. It may not be reasonable to assess and interpret bias in a new sample using data collected a decade ago. A possible solution is to introduce a reasonable time-discounting factor into the definitions of bias and credibility to account for concept drift [23]. Additionally, in streaming data scenarios, the current definition requires recalculating similarity and sample bias/credibility when new data arrives. Possible directions include maintaining a core matrix based on matrix low-rankness [56] to estimate data similarity after incorporating new data, thereby avoiding the need for re-computing RWR similarity each time. Exploring how to extend AIM to online scenarios to detect bias in newly arrived data in real-time, and its impact on existing data, would be a valuable future direction. We also note that our class-imbalance-aware design significantly contributes to AIM's advantage in predictive utility. And the class imbalance problem (classifier's uneven attention to different classes) [37–39] is closely related to unfairness, particularly group fairness. Exploring the relationship between these issues and developing joint solutions will be a promising direction for future research. Finally, we note that the comparable graph in this work can be seen as a simple nearest-neighbor-based graph. In many practical applications, there are complex networks of relationships between data points, forming graphs with intricate topologies [40, 59, 63]. FairML on graph data has recently gained significant attention, and extending AIM form artificial nearest-neighbor-based graph to natural complex graph data would also be an interesting direction.