MIFair: A Mutual-Information Framework for Intersectionality and Multiclass Fairness

Jeanne Monnier[1,2], Thomas George[1], Christèle Tarnec[1], Frédéric Guyard[1], and Marios Kountouris[2]

***Abstract*— Fairness in machine learning remains challenging due to its ethical complexity, the lack of a universal definition, and the need for context-specific bias metrics. Existing methods also remain limited in handling intersectionality, multiclass settings, and broader flexibility and generality. To address these limitations, we introduce *MIFair*, a unified framework for bias assessment and mitigation based on mutual information. MIFair provides a flexible metric template and an in-processing mitigation method inspired by the Prejudice Remover, defining group fairness as statistical independence between prediction-derived variables and sensitive attributes, while establishing equivalences with widely used notions such as independence and separation. MIFair naturally supports intersectionality, complex subgroup structures, and multiclass classification and employs regularization-based training to reduce bias according to the selected metric. Its key advantage is its versatility: it consolidates diverse fairness requirements into a single coherent framework, enabling consistent benchmarking and facilitating practical adoption. Experiments on real-world tabular and image datasets show that MIFair effectively reduces bias, including in previously unaddressed multi-attribute scenarios, while maintaining strong predictive performance across all evaluated settings.**

I. Introduction

Fairness in AI has received increasing attention because biased models can cause harm in high-stakes applications. However, fairness in machine learning remains a broad and contested concept, which has led to multiple, often incompatible, fairness notions [1], [2]. Moreover, AI systems can reproduce or amplify biases present in the training data, and such biases may arise throughout the development pipeline, from data collection to model deployment and user feedback [3]. As AI systems become more influential, mitigating these biases is increasingly important. Yet most fairness metrics remain tied to specific formulations, limiting their generality and their ability to reflect the diversity of real-world settings.

Fairness in AI/ML is commonly studied through individual, group, and causal perspectives; here, we focus on group fairness. Group fairness includes multiple criteria [1], each capturing a different notion of fairness across demographic groups [4], [5]. Beyond standard notions such as Statistical Parity, we also consider Conditional Fairness [6], which defines fairness as the conditional independence of model outcomes and sensitive attributes. Building on this view, MIFair unifies a broader set of fairness criteria within a mutual-information framework while supporting intersectional and multiclass settings, enabling more comprehensive and realistic fairness evaluation.

Information-theoretic approaches to measuring and enforcing fairness have gained prominence, building on foundational work such as [7], which we extend in several directions. Many conditional-fairness methods likewise rely on information-theoretic tools, including Shannon entropy and conditional mutual information [8], [9], as well as Jensen–Shannon divergence [10].

Our work builds on in-processing mitigation methods [8], [10]–[17] by extending regularization-based approaches to address biases arising from both the data and the training process. The closest prior works are [7] and [9]. Prejudice Remover [7] uses mutual information to enforce Statistical Parity, but it neither goes beyond this criterion nor explicitly evaluates intersectionality. Cho et al. [9] extend this line to Equalized Odds, yet still focus only on these two criteria and do not experimentally address intersectionality. Both methods also differ from ours in how mutual information is approximated and integrated into training, and neither considers multiclass classification.

In this paper, we introduce MIFair (pronounced "Am I Fair?"), an information-theoretic framework for unified fairness evaluation and mitigation in AI/ML. MIFair reformulates group fairness through mutual information, extending the Prejudice Remover [7] into a flexible metric template that captures multiple fairness notions, together with a corresponding regularization-based in-processing method. A key contribution of MIFair is its explicit support for two important yet underexplored challenges: *intersectionality* and *multiclass classification*. While most prior work focuses on single-attribute or binary settings [18], and existing tools rarely support intersectionality in practice despite its importance for capturing structural discrimination [19], MIFair enables fairness assessment and mitigation across complex subgroup structures and multiclass prediction tasks [20], [21]. Experiments on Adult and CelebA show that MIFair effectively reduces bias in intersectional and multiclass settings while providing a unified information-theoretic framework for comparing fairness notions. By reducing the fragmentation of existing fairness methods, MIFair offers a more accessible, adaptable, and principled approach to fairness in machine learning.

II. Problem Setting

Data and sensitive attributes. Let $D = \{(\mathbf{x}_d, \mathbf{a}_d, y_d)\}_{d=1}^{|D|}$ be a dataset drawn from an underlying distribution over $(\mathbf{X}, \mathbf{A}, Y)$, where each sample represents an individual. Here, $\mathbf{x}_d = (x_{d,1}, \ldots, x_{d,m}) \in \mathscr{X}$ denotes the vector of non-sensitive

[1]Orange Research, Châtillon, France
[2]EURECOM, Sophia Antipolis, France

attributes, $\mathbf{a}_d = (a_{d,1}, \ldots, a_{d,n-m}) \in \mathscr{A}$ denotes the vector of sensitive attributes, and $y_d \in \mathscr{Y}$ denotes the prediction target. The sensitive attributes are collected in the random vector $\mathbf{A} = (A_1, \ldots, A_{n-m})$, with sensitive-attribute space $\mathscr{A} = \prod_{i=1}^{n-m} \mathscr{A}_i$; these are the features with respect to which discrimination should be avoided, and their selection is context- and application-dependent. The remaining features are collected in the random vector $\mathbf{X} = (X_1, \ldots, X_m)$, with feature space $\mathscr{X} = \prod_{i=1}^{m} \mathscr{X}_i$, and are referred to as non-sensitive attributes. The number of sensitive attributes is $n-m$. Although these variables may in principle be discrete or continuous, this paper focuses on discrete classification settings; accordingly, the mutual-information expressions are written in discrete form.

Demographic groups. Given the sensitive attributes, we define *demographic groups* as sets of individuals sharing the same joint sensitive-attribute value $\mathbf{a} \in \mathscr{A}$. Let G denote the set of such groups, and let $|G|$ denote its cardinality. We distinguish two settings:

1) $|G| = 2$, corresponding to the binary case, which is the setting most commonly studied in the literature;
2) $|G| > 2$, corresponding to the non-binary case, which is more realistic in practice but far less explored.

Intersectionality. Intersectionality acknowledges that discrimination often arises from the combination of multiple sensitive attributes rather than from any single one considered in isolation. When multiple sensitive attributes are considered jointly, the setting typically becomes non-binary and may yield $|G| > 2$ demographic subgroups. Real-world applications commonly involve such intersectional structures, requiring fairness metrics capable of handling vectors of sensitive attributes.

Fairness metrics. Fairness is an ethical principle that cannot be captured by a single universal rule, which has led to the development of multiple *fairness notions*. Mathematical formalizations in AI/ML therefore yield diverse, and often incompatible, criteria [2]. Selecting an appropriate notion is therefore crucial to ensure alignment with the data and application context and to avoid unintended or amplified biases. An extensive overview of fairness notions is provided in [1], with a subset shown in Table I. These criteria are typically defined in a simplified setting in which both the sensitive attribute $\mathbf{A}$ and the model output $\hat{Y}$ are binary, although some notions extend to multiclass classification or regression. Most definitions rely on the joint distributions of $\mathbf{A}$, Y, and $\hat{Y}$. Following this line of work, we focus on three well-known fairness notions and two relaxations introduced below (see Section III-B), covering a broad portion of the existing literature.

Model training and optimization problem. Our goal is to learn a fair model by modifying the training objective to reduce bias while preserving the task's essential structure. Let $\mathbf{w}$ denote the model parameters. Standard learning seeks to approximate the ground-truth distribution $P(Y|\mathbf{X},\mathbf{A})$ through Empirical Risk Minimization (ERM):

$$\min_{w \in \mathbb{R}^h} L(D,w) = \min_{w \in \mathbb{R}^h} \sum_{d=1}^{|D|} \ell(\mathbf{x}_d, \mathbf{a}_d, y_d; w), \tag{1}$$

where L is the empirical risk and $\ell(\cdot)$ denotes the loss incurred on instance d. The ERM fits the patterns present in the data; therefore, if the data are biased, the trained model will tend to reproduce these biases, making the standard ERM insufficient for learning fair models.

III. MIFair: A Versatile Framework for Fairness Metrics

We now introduce the bias-assessment component of MIFair: a versatile metric template that unifies fairness evaluation and accommodates the diversity of existing methods. This template broadens the applicability of group fairness metrics by providing a standardized and extensible formulation[1] for the major fairness notions. We present the definitions used to measure bias under these notions and show their equivalence to the formulations listed in Table I.

A close examination of group fairness metrics reveals a common underlying structure: each metric specifies a particular variable, defined as a function of the model's predictions, that is directly tied to what is considered a non-prejudice (fair) treatment under the corresponding fairness notion. This variable can be interpreted as quantifying the *benefit* (or, conversely, the detriment) of belonging to a particular demographic subgroup. It therefore provides a basis for comparing the model's effects across subgroups. In the ideal case, a fully non-discriminatory model is obtained when the benefit variable is statistically independent of the sensitive attributes.

Measuring model bias can thus be framed as quantifying the mutual dependence between the sensitive attributes and the benefit variable derived from the model's predictions. Building on this insight, as well as on the Prejudice Remover method of [7], we introduce an information-theoretic fairness metric template grounded in Mutual Information. This template accommodates a broad range of group fairness notions and naturally supports both intersectional analyses and multiclass classification settings, thanks to the flexibility of mutual information in handling random variables of different dimensions.

A. A Metric Based on Mutual Information

Building on the preceding observation, we quantify the *mutual dependence* between the sensitive-attribute vector $\mathbf{A}$ and the benefit $B = f_b(\hat{Y}, Y)$, whose definition depends on the fairness notion under consideration (Section III-B).

[1]Throughout this work, we use raw mutual information $I(A;B)$ as the fairness criterion and optimization quantity. When cross-notion comparison on a common numeric range is desired, a normalized variant such as $I(A;B)/H(B)$, or its conditional analogue $I(A;B \mid C=c)/H(B \mid C=c)$, may also be required whenever the denominator is nonzero. Accordingly, MIFair should be interpreted primarily as a unified information-theoretic formulation; raw MI values across different fairness notions are not, in general, directly comparable on a strictly common scale.

Fairness Notion	Formulation	Ref.	Classification
Statistical parity	$P(\hat{Y} \mid A=0) = P(\hat{Y} \mid A=1)$	[22]	Independence
Equalized odds	$P(\hat{Y}=1 \mid Y=y, A=0) = P(\hat{Y}=1 \mid Y=y, A=1) \quad \forall y \in \{0,1\}$	[23]	Separation
Equal opportunity	$P(\hat{Y}=1 \mid Y=1, A=0) = P(\hat{Y}=1 \mid Y=1, A=1)$		
Predictive equality	$P(\hat{Y}=1 \mid Y=0, A=0) = P(\hat{Y}=1 \mid Y=0, A=1)$	[24]	
Overall accuracy equality	$P(\hat{Y}=Y \mid A=0) = P(\hat{Y}=Y \mid A=1)$	[25]	Other metrics

TABLE I: Classification of state-of-the-art fairness notions [1]

Using mutual information allows us to overcome two key limitations of existing fairness metrics. First, *intersectionality*: most common metrics compare probabilities across only two groups, restricting their applicability to binary protected attributes. Although extensions exist [26], they require numerous pairwise comparisons, which scale poorly as $|G|$ increases. Second, *multiclass classification*: real-world tasks frequently involve more than two outcome classes, requiring fairness metrics that go beyond binary positive/negative predictions. Our formulation naturally addresses both challenges.

We leverage tools from information theory, namely Mutual Information (MI), as the foundation of our approach. MI is non-negative and equals zero if and only if the variables are statistically independent; lower values therefore indicate weaker statistical dependence.

We define our metric template as

$$\iota_{\text{fairness}} = I(\mathbf{A};B) = \sum_{\mathbf{a}\in\mathscr{A}} \sum_{b\in\mathscr{B}} P_{\mathbf{A},B}(\mathbf{a},b) \log\left(\frac{P_{\mathbf{A},B}(\mathbf{a},b)}{P_{\mathbf{A}}(\mathbf{a})\,P_B(b)}\right), \tag{2}$$

where $P_{\mathbf{A},B}$ denotes the joint distribution of $\mathbf{A}$ and B, and $P_{\mathbf{A}}$ and P_B denote their corresponding marginal distributions.

$I(\mathbf{A};B)$ quantifies the amount of information about the sensitive attributes that is encoded in B. A value of 0 indicates complete fairness under the chosen notion, whereas larger values reflect stronger dependence and, therefore, greater bias. In practice, the true distributions are unknown and are replaced by their empirical estimates, $\widetilde{P}_{\mathbf{A},B}$, $\widetilde{P}_{\mathbf{A}}$, and $\widetilde{P}_B$, which enables the computation of ι_{fairness} in real-world scenarios (Section IV).

B. Unifying Existing Fairness Notions by Explicitly Defining the Corresponding Benefit

We examine the five notions of fairness listed in Table I and, for each of them, provide an explicit instantiation of our metric template ι_{fairness} by specifying the corresponding benefit B. We also show that fairness under our metric implies fairness under the corresponding classical fairness metric.

Among these notions, Overall Accuracy naturally extends to multiclass classification. Independence- and Separation-based notions, originally defined for positive predictions in binary settings, can also be generalized either by focusing on a particular class or by requiring the distribution of the considered event ($\hat{Y}$ or $\hat{Y}|Y$) to be equal across all classes. Under this extension, Statistical Parity, Equalized Odds, and its relaxations, Equal Opportunity and Predictive Equality, also apply to multiclass scenarios. Our equivalence and implication results remain valid in this setting, illustrating the flexibility of the framework and its potential to define new categories of fair models.

Statistical Parity. For Statistical Parity, the benefit B is defined as the model prediction $\hat{Y}$, i.e., $B := \hat{Y}$. The resulting fairness metric is

$$\iota_{SP} = I(\mathbf{A};\hat{Y}). \tag{3}$$

Equal Opportunity. Let $P_{Y=1} = P(\cdot \mid Y=1)$ denote the conditional distribution. Under Equal Opportunity, we measure the dependence between $\mathbf{A}$ and $\hat{Y}$ conditioned on $Y=1$, that is, we define $B := \hat{Y} \mid Y=1$:

$$\iota_{EO} = I_{Y=1}(\mathbf{A};\hat{Y}) = I(\mathbf{A};\hat{Y} \mid Y=1) \tag{4}$$

$$= \sum_{\mathbf{a}\in\mathscr{A}} \sum_{\hat{y}\in\mathscr{Y}} P_{Y=1}(\mathbf{a},\hat{y}) \log\left(\frac{P_{Y=1}(\mathbf{a},\hat{y})}{P_{Y=1}(\mathbf{a})\,P_{Y=1}(\hat{y})}\right). \tag{5}$$

Predictive Equality. Let $P_{Y=0} = P(\cdot \mid Y=0)$ denote the conditional distribution. Under Predictive Equality, we measure the dependence between $\mathbf{A}$ and $\hat{Y}$ conditioned on $Y=0$, that is, we define the benefit variable as $B := \hat{Y} \mid Y=0$:

$$\iota_{PE} = I_{Y=0}(\mathbf{A};\hat{Y}) = I(\mathbf{A};\hat{Y} \mid Y=0). \tag{6}$$

Equalized Odds. Let $P_{Y=y} = P(\cdot \mid Y=y)$ denote the conditional distribution for $y \in \{0,1\}$. For Equalized Odds, we combine the conditional dependencies associated with both outcomes:

$$\iota_{EOdds} = \lambda_0 I_{Y=0}(\mathbf{A};\hat{Y}) + \lambda_1 I_{Y=1}(\mathbf{A};\hat{Y}), \qquad \lambda_0,\lambda_1 > 0. \tag{7}$$

where $\lambda_0,\lambda_1 > 0$ are weighting coefficients.

Overall Accuracy Equality. For Overall Accuracy Equality, we define $B := \mathbb{1}\{\hat{Y}=Y\}$ and the corresponding metric becomes

$$\iota_{OAE} = I\big(\mathbf{A};\mathbb{1}\{\hat{Y}=Y\}\big). \tag{8}$$

IV. MIFair: A Comprehensive Framework for Bias Mitigation

We introduce a bias mitigation framework grounded in our fairness metric ι_{fairness}. MIFair operates as an in-processing method, addressing bias directly during model training. Among in-processing techniques, an effective way to improve fairness is to regularize the ERM objective in (1). Inspired by the Prejudice Remover [7], which our method generalizes, we incorporate our flexible metric ι_{fairness} as a tailored regularization term added to the loss function. Standard ERM optimizes only for accuracy and therefore

reproduces the biases embedded in the training data. Introducing a fairness regularizer shifts the objective toward approximating a modified distribution that remains close to the original distribution while exhibiting reduced dependence according to the selected fairness notion. This shift generally entails some accuracy loss on the training data, which is acceptable, or even desirable, when the data itself is biased. Nonetheless, maintaining adequate predictive performance remains essential. The trade-off between fidelity to the data and fairness is governed by the regularization hyperparameter η. As discussed in Section III-A, minimizing mutual information reduces the dependence between two variables. Since our metric provides an empirical approximation of this quantity for the variables relevant to a chosen fairness notion, it serves naturally as a regularizer in the training objective. The initial optimization problem (1) thus becomes

$$\min_{w\in\mathbb{R}^h} L(D,w) + \eta\, \iota_{\text{fairness}}(D,w). \tag{9}$$

The proposed framework consists of the following steps:

1) **Identify sensitive attributes**: determine which features in the dataset should be treated as sensitive and define the associated demographic subgroups.
2) **Select a fairness definition**: choose the appropriate notion for the task and define ι_{fairness} accordingly.
3) **Set up the optimization problem**: incorporate the chosen ι_{fairness} into the training objective.
4) **Train the model**: optimize the regularized objective.

In practice, the value of ι_{fairness} required in (9) is computed by estimating mutual information from minibatch samples. The joint and marginal distributions $P_{\mathbf{A},B}$, $P_{\mathbf{A}}$, and P_B are approximated by their empirical counterparts, denoted by $\widetilde{P}$, which are derived from the training data and the model output probabilities $p_{\mathbf{w}}$. Let $\mathscr{M}$ denote the current minibatch, with cardinality $|\mathscr{M}|$. The fairness regularizer is estimated from minibatch samples as

$$\hat{\iota}_{\text{fairness}} = \sum_{\mathbf{a}\in\mathscr{A}} \sum_{b\in\mathscr{B}} \widetilde{P}_{\mathbf{A},B}(\mathbf{a},b) \log\left(\frac{\widetilde{P}_{\mathbf{A},B}(\mathbf{a},b)}{\widetilde{P}_{\mathbf{A}}(\mathbf{a})\,\widetilde{P}_B(b)}\right), \tag{10}$$

where

$$\widetilde{P}_{\mathbf{A},B}(\mathbf{a},b) = \frac{1}{|\mathscr{M}|}\sum_{d\in\mathscr{M}} \mathbb{1}\{\mathbf{a}_d=\mathbf{a}\}\, p_w(B_d=b),$$

$$\widetilde{P}_{\mathbf{A}}(\mathbf{a}) = \frac{1}{|\mathscr{M}|}\sum_{d\in\mathscr{M}} \mathbb{1}\{\mathbf{a}_d=\mathbf{a}\}, \qquad \widetilde{P}_B(b) = \frac{1}{|\mathscr{M}|}\sum_{d\in\mathscr{M}} p_w(B_d=b).$$

Adequate minibatch coverage is necessary to obtain a reliable approximation of the distribution of $\mathbf{A}$ over $\mathscr{A}$.

We adopt a count-based approximation rather than more computationally intensive adversarial estimation methods [9] or variational estimators (e.g., MINE [27]), to avoid adversarial training, reduce computational cost, and prevent the underestimation and instability that may arise from the use of lower bounds. To ensure reliable probability estimates, the batch size must be sufficiently large and must include representative samples from all demographic subgroups. Such coverage is necessary to approximate the distribution of $\mathbf{A}$ over its full domain.

V. Experiments

We evaluate MIFair across multiple scenarios to assess its fairness performance. The results show that MIFair consistently achieves strong fairness performance under all tested notions, including highly intersectional settings and multiclass classification. The induced accuracy trade-offs remain modest, indicating that the method effectively adjusts the original distribution without excessive distortion. In our experiments, we implement three of the five fairness notions introduced in Section III-B: Statistical Parity (SP), Equal Opportunity (EO), and Overall Accuracy Equality (OAE).

A. Setup

*a) **Adult Dataset**:* We evaluate MIFair on the UCI Adult dataset [28], a standard fairness benchmark. The dataset includes 14 binary, categorical, or continuous features and a binary income label indicating whether an individual earns more than \$50K. We use a two-layer fully connected neural network with a softmax output, trained using cross-entropy loss. We consider three sensitive attributes, *race*, *sex*, and *relationship status*, each treated as binary, with values $\{White, Non\text{-}White\}$, $\{Male, Female\}$, and $\{Not\text{-}in\text{-}Family, In\text{-}Family\}$, respectively.

*b) **CelebA Dataset**:* We further evaluate MIFair on the large-scale CelebA vision dataset [29] using a ResNet18 model. We consider both a binary (Task 1) and a multiclass (Task 2) prediction task, namely, determining whether a person is *smiling* (Task 1) and whether the person has *blond*, *brown*, or *black* hair (Task 2), while enforcing fairness with respect to the sensitive attributes *Male*/*Female* and *Chubby*/*Non-Chubby*. The dataset is highly imbalanced, for example, the $\{Female, Non\text{-}Chubby\}$ subgroup contains hundreds of times as many samples as $\{Female, Chubby\}$. Unlike the Adult dataset, these sensitive attributes are implicit in the input (image pixels) rather than explicitly provided, demonstrating MIFair's ability to mitigate bias in unstructured data.

*c) **Evaluation**:* We evaluate the performance of the trained models based on their predictions using the following two accuracy metrics:

$$\text{ACC}_{\text{mean}} = \frac{1}{|D|}\sum_{d=1}^{|D|} \mathbb{1}\{\hat{y}_d = y_d\}.$$

$$\text{ACC}_{\text{weighted}} = \frac{1}{|G|}\sum_{g\in G}\frac{1}{N_g}\sum_{d:\,\mathbf{a}_d=g} \mathbb{1}\{\hat{y}_d = y_d\},$$

where $N_g = |\{d\in\{1,\dots,|D|\} : \mathbf{a}_d = g\}|$. $\text{ACC}_{\text{weighted}}$ assigns equal importance to each demographic subgroup's accuracy, regardless of its size, providing a clearer view of balance across subgroups. In contrast, ACC_{mean} maintains representation biases, as more frequent subgroups in the training data have a greater impact on the overall evaluation, potentially masking such biases.

Second, we use the following **MIFair fairness metrics**

for fairness evaluation:

$$\iota_{\text{SP}} = I(\mathbf{A};\hat{Y})$$
$$\iota_{\text{EO}} = I(\mathbf{A};\hat{Y} \mid Y=1)$$
$$\iota_{\text{OAE}} = I\big(\mathbf{A};\mathbb{1}\{\hat{Y}=Y\}\big).$$

For comparison, we also employ the following **baseline metrics** from the literature. Let $(\mathbf{a},\tilde{\mathbf{a}}) \in \mathscr{A}^2$, with $\mathbf{a} \neq \tilde{\mathbf{a}}$.

$$\text{SPD}(\mathbf{a},\tilde{\mathbf{a}}) = P(\hat{Y}=1 \mid \mathbf{A}=\mathbf{a}) - P(\hat{Y}=1 \mid \mathbf{A}=\tilde{\mathbf{a}})$$
$$\text{EOD}(\mathbf{a},\tilde{\mathbf{a}}) = P(\hat{Y}=1 \mid Y=1,\mathbf{A}=\mathbf{a}) - P(\hat{Y}=1 \mid Y=1,\mathbf{A}=\tilde{\mathbf{a}})$$
$$\text{OAE}(\mathbf{a},\tilde{\mathbf{a}}) = P(\hat{Y}=Y \mid \mathbf{A}=\mathbf{a}) - P(\hat{Y}=Y \mid \mathbf{A}=\tilde{\mathbf{a}}).$$

These metrics take values in $[-1,1]$, as they represent differences between probabilities. Their values may be either positive or negative, depending on the ordering of the subgroups. A positive value indicates that group $\mathbf{a}$ is the privileged group, whereas a negative value indicates that $\tilde{\mathbf{a}}$ is the privileged group. A value of zero corresponds to a perfectly fair predictor under the given fairness criterion. In practice, values in the range $[-0.2, 0.2]$ are often considered acceptable [30].

*d) **Experiments**:* We consider an intersectional setting with $n-m=3$ (*sex*, *race*, *relationship*) and $|G|=8$ for Adult, and $n-m=2$ (*sex*, *chubby*) and $|G|=4$ for CelebA. Three experiments are conducted by training a model with the MIFair metric instantiated using *SP*, *EO*, and *OAE*, respectively. For CelebA, only Experiment 1 was conducted.

$$\text{Exp. 1:} \quad \min_{w\in\mathbb{R}^h} L(D,w) + \eta\, \iota_{SP}(D,w),$$
$$\text{Exp. 2:} \quad \min_{w\in\mathbb{R}^h} L(D,w) + \eta\, \iota_{EO}(D,w),$$
$$\text{Exp. 3:} \quad \min_{w\in\mathbb{R}^h} L(D,w) + \eta\, \iota_{OAE}(D,w).$$

For each experiment, we verify whether the corresponding ι_{fairness} metric is minimized and whether the associated baseline metric (*SPD*, *EOD*, or *OAE*) is also reduced across all subgroup pairs $(\mathbf{a},\tilde{\mathbf{a}}) \in \mathscr{A}^2$, as expected from the minimization of ι_{fairness} (Section III-B). For each configuration, we run five independent trials with distinct random seeds, using identical data splits across methods, and report the mean performance across runs.

*e) **Varying regularization strength**:* We analyze the influence of the regularization strength by varying the value of η in (9) and identify threshold values that yield satisfactory fairness results while maintaining a favorable fairness-accuracy trade-off. $\eta = 0$ (no regularization) is denoted as the "vanilla" setting.

B. Results

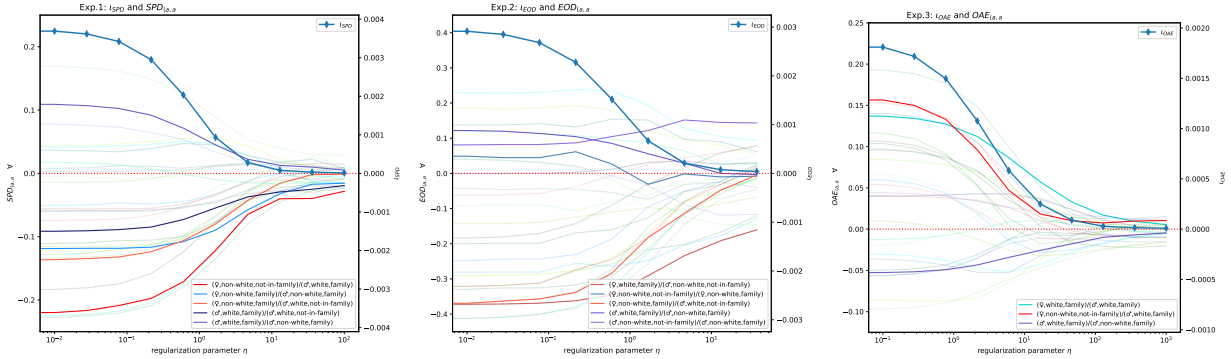
*a) **Increased regularization strength minimizes the corresponding metric**:* As a sanity check, we verify that adding the regularization term effectively reduces the corresponding MIFair metric across fairness notions. Fig. 1 shows that ι_{SP}, ι_{EO}, or ι_{OAE} decrease sharply as η increases in Exp. 1, 2 or 3 respectively, confirming the expected effect of stronger regularization. We also observe that minimizing the MIFair term consistently reduces the corresponding classical metrics (SPD, EOD, and OAE). For all three notions, the empirical metrics remain bounded by a decreasing function of η, demonstrating progressive bias mitigation. Using a threshold $s = 0.2$ across subgroups, we find that achieving $|SPD| \leq s$ requires $\eta \geq 10^{-0.8}$, and $|EOD| \leq s$ requires $\eta \geq 10^{1.1}$. The vanilla model already satisfies $|OAE| \leq s$. Fig. 2 shows the same behavior on the CelebA dataset, confirming that MIFair performs consistently across the evaluated data distributions.

*b) **MIFair handles intersectionality**:* In Fig. 1, the baseline metrics (SPD, EOD, OAE) are evaluated across all subgroup pairs $(\mathbf{a},\tilde{\mathbf{a}}) \in \mathscr{A}^2, \mathbf{a} \neq \tilde{\mathbf{a}}$. Most values converge toward zero, which corresponds to complete fairness, as regularization increases, with a few exceptions discussed below. Fig. 1a highlights five representative curves and shows that MIFair effectively mitigates intersectional biases across diverse subgroup differences. The two curves in shades of red-orange correspond to subgroup pairs that do not share any sensitive-attribute values, that is, to the most distinct and therefore the most biased intersectional groups. These pairs exhibit the highest disparities at $\eta = 0$. As regularization increases, their gaps steadily shrink and eventually converge toward zero, illustrating MIFair's ability to reduce intersectional bias. All subgroup pairs ultimately lie within the commonly accepted fairness band $[-0.2, 0.2]$. Similar behavior is observed for the red-orange curves in Figs. 1b and 1c, further confirming MIFair's robustness in intersectional settings.

*c) **MIFair works in multiclass classification**:* Fig. 3 shows the results of MIFair-regularized training for a multiclass classification task under the *SP* criterion. As observed for binary tasks, accuracy losses remain limited, with a maximum loss of 9.1% under the strongest regularization setting, $\eta = 10^2$, whereas the fairness gains are substantial, with a 96% decrease in the ι_{SP} value. Moreover, the loss in accuracy can be reduced further while maintaining a substantial level of fairness through an appropriate choice of η, thereby improving the fairness-accuracy trade-off. These results show that MIFair can effectively mitigate bias in multiclass classification tasks under the selected fairness definition.

*d) **MIFair maintains performance on biases arising from binary subgroup comparisons**:* The blue and purple curves in Fig. 1 correspond to subgroup pairs differing in only one sensitive attribute, thereby effectively reducing the comparison to a binary setting ($|G| = 2$, with a single varying sensitive attribute). These curves also converge toward zero as η increases, showing that MIFair mitigates biases in binary scenarios just as effectively as in intersectional ones.

*e) **MIFair has limitations with scarce data**:* Results in Fig. 1b are slightly weaker than those in Figs. 1a and 1c. As a separation-based notion, *Equal Opportunity* relies only on positive-label samples, reducing the effective data available. The Adult dataset is strongly imbalanced, with negative labels outnumbering positives by roughly 3:1, which limits the accuracy of both the distribution estimates and the regularization term. This results in suboptimal behavior for



(a) Exp.1: $SPD(\mathbf{a},\tilde{\mathbf{a}})$ & ι_{SP} (b) Exp.2: $EOD(\mathbf{a},\tilde{\mathbf{a}})$ & ι_{EO} (c) Exp.3: $OAE(\mathbf{a},\tilde{\mathbf{a}})$ & ι_{OAE}

Fig. 1: (Adult dataset.) MIFair metric ι_{fairness} (right y-axis) and the corresponding fairness metric over all subgroups (left y-axis) as η varies, for three fairness definitions: *SP* (Fig. 1a), *EO* (Fig. 1b) and *OAE* (Fig. 1c). The highlighted curves are discussed further in Section V-B.

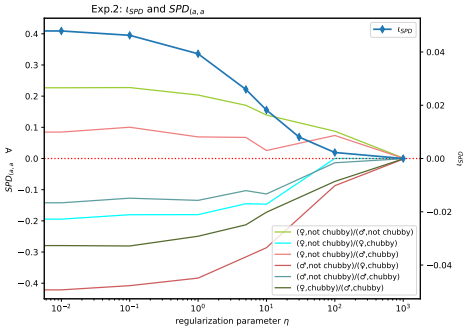


Fig. 2: (CelebA dataset.) ι_{SP} (right y-axis) and the corresponding fairness metrics from the literature over all subgroups (left y-axis) for the binary prediction task, as η varies (x-axis) in Experiment 1 on Task 1.

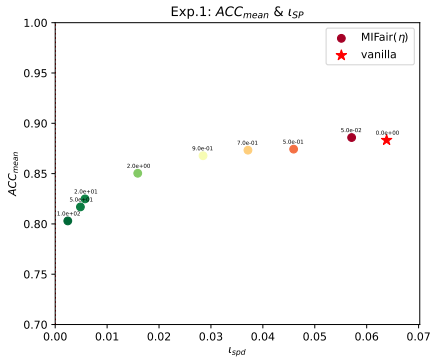
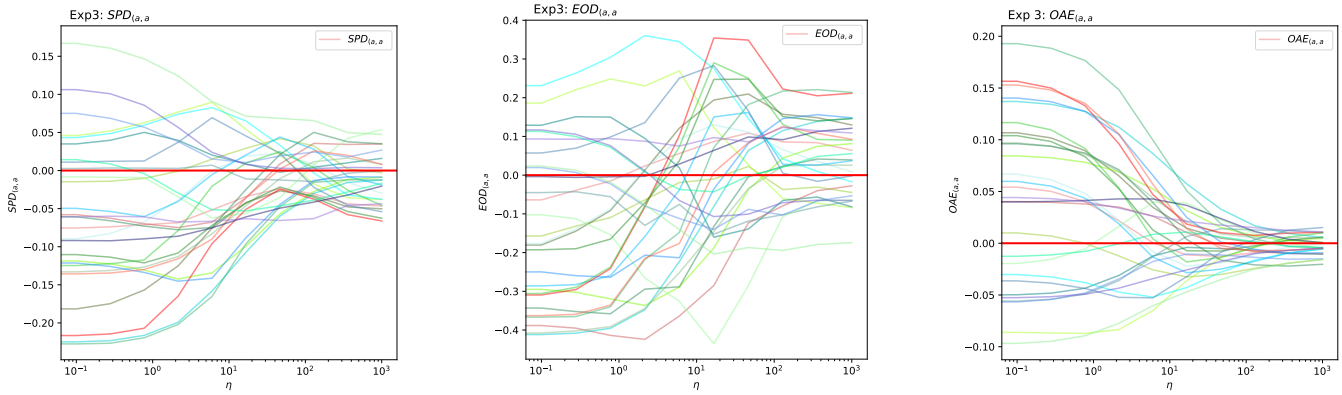


Fig. 3: (CelebA dataset.) Fairness-accuracy trade-off for the multiclass classification task (Exp.1 on Task 2).

underrepresented subgroups, particularly those with $A_{\text{race}} =$ Non-White, where probability estimates become less reliable, restricting MIFair's ability to fully mitigate bias. Nevertheless, the fairness condition $|EOD| \leq s$ is still satisfied, indicating that acceptable fairness performance is achieved even for these disadvantaged groups.

f) ***Fairness notions incompatibility****:* A key advantage of MIFair is its ability to unify diverse fairness definitions within a single framework. To illustrate the consequences of selecting an unsuitable fairness notion, Fig. 4 shows the evolution of three metrics in Experiment 3 on Adult, where only ι_{OAE} is minimized, targeting *Overall Accuracy Equality*. While Fig. 4c shows that *OAE* effectively converges to zero across all subgroup pairs as η increases, Figs. 4a and 4b reveal that *SPD* and *EOD* are not minimized under this regularization, unlike in Experiments 1 and 2 (Fig. 1). *SPD* shows only minor improvement, far below the gains obtained in Experiment 1, and *EOD* is substantially degraded, offering no fairness benefit and in some cases worsening disparities. These findings highlight the inherent incompatibility between fairness notions and underscore the importance of selecting the correct metric for the fairness objective. MIFair mitigates this risk by providing a unified, accessible framework that guides practitioners toward appropriate bias mitigation choices.

g) ***Fairness-accuracy trade-off****:* We analyze how accuracy evolves under increasing regularization. Accuracy reflects how much information the model extracts from the data; if it collapses, training becomes meaningless. However, when fairness is prioritized, the relevance of accuracy, especially ACC_{mean}, depends on the fairness notion being enforced. Achieving fairness often requires modifying the original data distribution; therefore, some decrease in accuracy is to be expected and may, in certain cases, even be desirable. In Section I, we noted that representation bias causes underrepresented subgroups to contribute less to the objective, making ACC_{mean} a poor indicator of subgroup-level performance. In contrast, $\text{ACC}_{\text{weighted}}$ better captures accuracy across subgroups. When representation biases are reduced, $\text{ACC}_{\text{weighted}}$ may increase, or decline more slowly, even as ACC_{mean} decreases. Fig. 6 illustrates this behavior in Experiment 1, where regularization targets ι_{SP}. As regularization increases, $\text{ACC}_{\text{weighted}}$ drops less sharply than ACC_{mean}, indicating a partial correction of representation bias.

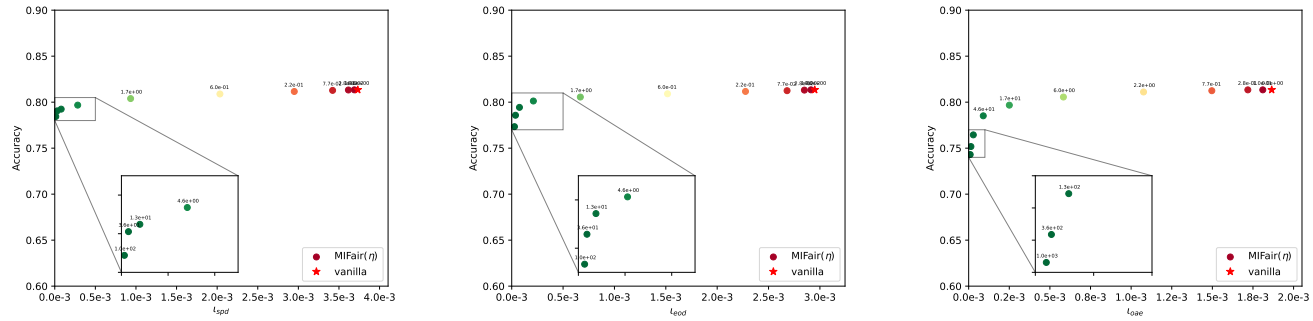


(a) *SPD* across subgroup pairs when ι_{OAE} is minimized

(b) *EOD* across subgroup pairs when ι_{OAE} is minimized

(c) *OAE* across subgroup pairs when ι_{OAE} is minimized

Fig. 4: (Adult dataset.) **Incompatibility between fairness notions**. In Exp. 3, MIFair is configured for *Overall Accuracy Equality*, successfully minimizing ι_{OAE} and *OAE* (Fig. 4c). However, this improvement comes at the cost of degraded performance under other fairness notions, such as *Statistical Parity* (Fig. 4a) and *Equal Opportunity* (Fig. 4b), illustrating the inherent incompatibility between fairness criteria. Each curve corresponds to a pair of subgroups.



(a) Exp.1: ACC_{mean} & ι_{SP}

(b) Exp.2: ACC_{mean} & ι_{EO}

(c) Exp.3: ACC_{mean} & ι_{OAE}

Fig. 5: (Adult dataset.) **Fairness-accuracy trade-off**: Stronger MIFair regularization reduces the corresponding MIFair metric (x-axis) and shifts the predictions away from the original empirical distribution of the test data, as reflected by a lower value of ACC_{mean} (y-axis).

The fairness-accuracy trade-off and its evolution with η reflect how closely the model follows the original, potentially biased, data distribution and how much accuracy is sacrificed to improve fairness. Figs. 5 and 3 show the corresponding gains in fairness and losses in accuracy. Across all experiments, ι_{fairness} decreases substantially, whereas the drop in accuracy at the largest values of η remains limited relative to the vanilla model. Although we intentionally consider large values of η to strongly reduce the fairness metrics, such extreme regularization is generally unnecessary in practice; moderate values of η typically provide a more favorable fairness-accuracy trade-off.

*h) **Baselines**:* Comparing our method with prior work is nontrivial because most baselines neither support nor evaluate intersectional settings or multiclass outputs. Existing code and protocols typically assume one binary sensitive attribute and binary classification, so a like-for-like comparison would either narrow our study and weaken our main contribution or require substantial third-party extensions that would confound the comparison. Furthermore, MIFair adopts an information-theoretic evaluation framework that quantifies the dependence between the sensitive-attribute vector $\mathbf{A}$ and a notion-specific variable B using mutual information, rather than pairwise subgroup disparity metrics such as SPD, EOD, and OAE. Although both approaches coincide at zero bias, they capture different forms of residual dependence; evaluating competing methods under a single framework would therefore systematically favor that framework. Accordingly, in Table II, we compare MIFair with the KDE-based framework proposed by [31] on two fairness criteria. This framework aims to reduce subgroup deviations from the overall positive prediction rate with respect to demographic parity:

$$DDP := \sum_{\mathbf{a}\in\mathscr{A}} \left|\Pr(\hat{Y}=1 \mid \mathbf{A}=\mathbf{a}) - \Pr(\hat{Y}=1)\right|. \tag{11}$$

For the comparison with [31], we use the preprocessing and subgroup definition provided in the authors' code, which

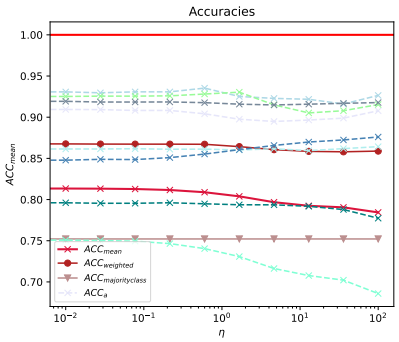


Fig. 6: (Adult dataset.) Exp. 1: Accuracy decreases slightly as η increases.

η/η_{max}	MIFair			Baseline [31]		
	Acc.	DDP	MI	Acc.	DDP	MI
0	**0.846**	1.10	$2.83 \cdot 10^{-2}$	**0.846**	1.10	$2.83 \cdot 10^{-2}$
0.5	**0.834**	0.610	0.029	0.760	0.021	0.32
0.75	0.827	0.541	0.012	0.676	0.086	0.035
0.9	0.788	0.191	0.0045	0.602	**0.081**	0.0054

TABLE II: (Adult dataset.) Comparative evaluation of MIFair and the baseline [31] across regularization strengths η in Experiment 1.

differs from the binary race encoding used in our main Adult experiments and yields $|G| = 20$. We report both criteria, ι_{SP} and DDP, across all experiments. As expected, MIFair achieves a lower ι_{SP}, whereas the KDE-based baseline yields a smaller DDP. However, MIFair exhibits substantially better accuracy while still tending toward a virtually unbiased model. Beyond performance, this comparison highlights that mutual information and pairwise distributional metrics capture bias differently. More broadly, especially in intersectional settings, it shows that aggregating subgroup-level measures into a single fairness score can materially affect conclusions about residual bias.

VI. Conclusion

We introduced MIFair, a unified mutual-information-based framework for assessing and mitigating bias across various fairness notions originally defined in binary settings, thereby extending their applicability to intersectional and multiclass scenarios. The resulting regularization-based in-processing method supports multiple sensitive-attribute configurations and prediction settings. Experiments on neural-network models for tabular data and on deep models for image classification demonstrate that MIFair provides a robust and adaptable fairness mechanism that substantially reduces bias while incurring only limited accuracy loss. Future work includes exploring alternative mutual-information estimation methods and extending MIFair to continuous features and outputs through continuous mutual-information formulations.

References

[1] K. Makhlouf, S. Zhioua, and C. Palamidessi, "Machine learning fairness notions: Bridging the gap with real-world applications," *Information Processing & Management*, 2021.

[2] S. Mitchell, E. Potash, S. Barocas, A. D'Amour, and K. Lum, "Prediction-based decisions and fairness: A catalogue of choices, assumptions, and definitions," *arXiv:1811.07867*, 2018.

[3] N. Mehrabi, F. Morstatter, N. Saxena, K. Lerman, and A. Galstyan, "A survey on bias and fairness in machine learning," *ACM Computing Surveys (CSUR)*, 2021.

[4] B. Ruf and M. Detyniecki, "Towards the right kind of fairness in AI," *arXiv preprint arXiv:2102.08453*, 2021.

[5] S. Barocas, M. Hardt, and A. Narayanan, *Fairness and machine learning: Limitations and opportunities*. MIT Press, 2023.

[6] R. Xu, P. Cui, K. Kuang, B. Li, L. Zhou, Z. Shen, and W. Cui, "Algorithmic decision making with conditional fairness," in *Proceedings of the 26th ACM SIGKDD*, 2020.

[7] T. Kamishima, S. Akaho, H. Asoh, and J. Sakuma, "Fairness-aware classifier with prejudice remover regularizer," in *ECML PKDD*. Springer, 2012.

[8] A. Ghassami, S. Khodadadian, and N. Kiyavash, "Fairness in supervised learning: An information theoretic approach," in *IEEE ISIT*, 2018.

[9] J. Cho, G. Hwang, and C. Suh, "A fair classifier using mutual information," in *IEEE ISIT*, 2020.

[10] J. Hwa, Q. Zhao, A. Lahiri, A. Masood, B. Salimi, and E. Adeli, "Enforcing conditional independence for fair representation learning and causal image generation," in *IEEE/CVF*, 2024.

[11] B. H. Zhang, B. Lemoine, and M. Mitchell, "Mitigating unwanted biases with adversarial learning," in *AAAI/ACM AIES*, 2018.

[12] D. Madras, T. Pitassi, and R. Zemel, "Predict responsibly: improving fairness and accuracy by learning to defer," *NeurIPS*, 2018.

[13] V. Iosifidis and E. Ntoutsi, "Adafair: Cumulative fairness adaptive boosting," in *ACM CIKM*, 2019.

[14] E. Krasanakis, E. Spyromitros-Xioufis, S. Papadopoulos, and Y. Kompatsiaris, "Adaptive sensitive reweighting to mitigate bias in fairness-aware classification," in *WWW*, 2018.

[15] H. Jiang and O. Nachum, "Identifying and correcting label bias in machine learning," in *AISTATS*, 2020.

[16] R. Jiang, A. Pacchiano, T. Stepleton, H. Jiang, and S. Chiappa, "Wasserstein fair classification," in *UAI*. PMLR, 2020.

[17] H. Do, P. Putzel, A. S. Martin, P. Smyth, and J. Zhong, "Fair generalized linear models with a convex penalty," in *ICML*. PMLR, 2022.

[18] F. Yang, M. Cisse, and S. Koyejo, "Fairness with overlapping groups; a probabilistic perspective," *NeurIPS*, 2020.

[19] J. Buolamwini and T. Gebru, "Gender shades: Intersectional accuracy disparities in commercial gender classification," in *Conference on fairness, accountability and transparency*. PMLR, 2018.

[20] L. E. Celis and V. Keswani, "Improved adversarial learning for fair classification," *arXiv preprint arXiv:1901.10443*, 2019.

[21] F. Kamiran and T. Calders, "Data preprocessing techniques for classification without discrimination," *KAIS*, 2012.

[22] 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.

[23] M. Hardt, E. Price, and N. Srebro, "Equality of opportunity in supervised learning," *NeurIPS*, 2016.

[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*, 2017.

[25] 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*, 2021.

[26] M. Kearns, S. Neel, A. Roth, and Z. S. Wu, "Preventing fairness gerrymandering: Auditing and learning for subgroup fairness," in *ICML*, 2018.

[27] M. I. Belghazi, A. Baratin, S. Rajeshwar, S. Ozair, Y. Bengio, A. Courville, and D. Hjelm, "Mutual information neural estimation," in *ICML*. PMLR, 2018.

[28] B. Becker and R. Kohavi, "Adult," *UCI Mach. Learning Repository*, 1996.

[29] Z. Liu, P. Luo, X. Wang, and X. Tang, "Deep learning face attributes in the wild," in *IEEE ICCV*, 2015.

[30] P. Saleiro, B. Kuester, L. Hinkson, J. London, A. Stevens, A. Anisfeld, K. T. Rodolfa, and R. Ghani, "Aequitas: A bias and fairness audit toolkit," 2019. [Online]. Available: https://arxiv.org/abs/1811.05577

[31] J. Cho, G. Hwang, and C. Suh, "A fair classifier using kernel density estimation," *Advances in NeurIPS*, 2020.