Multi-group Fairness Measures for Classification

Massimiliano Mancini[a], Donatella Merlini[b] and Maria Cecilia Verri[c]
Dipartimento di Statistica, Informatica, Applicazioni, Università degli Studi di Firenze, Viale Morgagni 65, Florence, Italy
{*massimiliano.mancini, donatella.merlini, mariacecilia.verri*}*@unifi.it*

Keywords: fairness measures, classification, calibration, machine learning, multiple groups, non-discrimination criteria

Abstract: This paper proposes a robust, reliable, and reproducible methodology for evaluating group fairness in classification algorithms. Building upon established theoretical definitions of non-discrimination criteria - Independence, Separation, and Sufficiency- we present a comprehensive approach to quantifying fairness. Notably, our methodology extends beyond binary group comparisons to accommodate scenarios with multiple sensitive groups. Furthermore, we provide a methodological approach for assessing fairness within the Sufficiency criterion by operationalizing Calibration, elucidating critical issues and conceptual subtleties that appear to have been overlooked in existing literature. We assess the reliability and robustness of our model through applications to one real dataset.

1 INTRODUCTION

Artificial intelligence increasingly supports decision-making in domains such as finance, recruitment, healthcare, and education (Purificato et al., 2023; Mukerjee et al., 2002; Chen et al., 2024; Reddy et al., 2020; Deo, 2015; Cortez and Silva, 2008; Holmes and Porayska-Pomsta, 2023; Zlatkin-Troitschanskaia et al., 2019). Since these applications often rely on classification algorithms, ensuring fair outcomes is essential. In this work, we focus on fairness, viewed as an *unobservable theoretical construct* inferred from observable properties (Jacobs and Wallach, 2021). Building on established non-discrimination criteria (Barocas et al., 2023), we propose a reproducible framework for assessing group fairness through Independence, Separation, and Sufficiency. In particular, we introduce an operational formulation of Calibration (Chouldechova, 2017), addressing the lack of practical implementations for Sufficiency-based measures. While fairness evaluation is traditionally framed as a comparison between two groups, many real-world applications involve multiple sensitive groups simultaneously. Our work extends this conventional setting by providing practical measures and interpretable representations for multi-group fairness analysis. Research on fairness in machine learning has grown substantially in recent years, leading to a wide range of definitions and metrics often organized into taxonomies to clarify their relationships (Makhlouf et al., 2021; Caton and Haas, 2024; Pessach and Shmueli, 2022; Mehrabi et al., 2021; Loureiro et al., 2025; Soremekun et al., 2025). Complementary efforts have focused on bias mitigation through pre-, in-, and post-processing techniques (Zhang et al., 2020; Lohia et al., 2019; Calmon et al., 2017), as well as on the development of practical frameworks such as FairLearn (Bird et al., 2020), IBM Fairness AI 360 (Zhang et al., 2020), and Aequitas (Saleiro et al., 2018). However, these tools generally do not operationalize Sufficiency-based measures such as Calibration, and existing studies on Calibration (Mancini et al., 2026a; Mancini et al., 2026b) are largely limited to binary group settings. Despite this progress, operational approaches that jointly account for Independence, Separation, and Sufficiency remain limited. While these criteria are generally incompatible (Barocas et al., 2023), we argue that they can still be jointly analyzed to provide complementary perspectives on fairness. In this work, we aim to bridge the gap between theoretical definitions and practical measurement by proposing a unified framework for multi-group fairness evaluation. In particular, we introduce a practical method for computing Calibration, for which few operational implementations exist in the literature and software libraries. Under suitable conditions, our approach also enables comparisons across measures derived from different non-discrimination criteria, supporting a more com-

[a] https://orcid.org/0009-0007-2569-3033
[b] https://orcid.org/0000-0001-8109-4077
[c] https://orcid.org/0000-0003-0444-0778

prehensive analysis of demographic disparities.

2 Preliminaries and Notation

We consider a supervised learning setting on a dataset S containing attributes related to individuals, partitioned into k groups defined by a sensitive attribute $G \in \{g_1, \ldots, g_k\}$. The learning task involves a binary notion of outcome, where $Y \in \{-1, 1\}$ denotes the true success label and $\hat{Y} \in \{-1, 1\}$ the predicted label. For measures based on Independence and Separation, fairness is evaluated using $G, Y, \hat{Y}$. For Sufficiency-based measures, we additionally assume the presence of a score variable $R \in \mathcal{R}$ and its prediction $\hat{R} \in \mathcal{R}$, where $\mathcal{R}$ is a generic (possibly discrete or continuous) outcome space. These scores provide a finer-grained representation of outcomes than binary labels. To connect score-based and binary formulations, we define a sign function $s : \mathcal{R} \to \{-1, 1\}$ such that $Y = s(R)$ and $\hat{Y} = s(\hat{R})$. This function encodes a success criterion over the score space. Table 1 summarizes the main variables used in the framework. We

Variable	Description
G	Sensitive attribute
R	Observed score
Y	True binary outcome
$\hat{R}$	Predicted score
$\hat{Y}$	Predicted binary outcome

Table 1: Model variables

use standard probabilistic notation, where $\perp$ denotes statistical independence and $\mid$ conditional independence. Given a dataset S, we use the selection operator $\sigma_F(S)$ to denote the subset satisfying predicate F. For group-specific subsets, we write $S^i = \sigma_{G=g_i}(S)$. The cardinality of a set is denoted by $|\cdot|$. Fairness measures are later represented in matrix form, where capital letters denote matrices and lowercase indexed entries denote values for each pair of groups.

3 Methodology

We propose a unified fairness evaluation framework that adapts to different numbers of sensitive groups. We distinguish three scenarios of increasing generality: (i) a binary setting, (ii) a one versus all multi-group setting, and (iii) a full multi-group setting with aggregated representations. This progression allows a consistent transition from simple scalar evaluations to richer inter-group structures. In the binary case ($k = 2$), fairness is expressed as a single scalar in $[-1, 1]$. Positive values indicate advantage for the first group, negative values for the second, and zero corresponds to perfect fairness. For $k > 2$, fairness can be evaluated with respect to a reference group, say g_1, by comparing it against the aggregation of all other groups. All measures are defined in probabilistic terms, ensuring independence from group size. In a full multi-group setting with $k > 2$ groups, we define a pairwise fairness matrix $E \in \mathbb{R}^{k \times k}$, where $e_{ij} \in [-1, 1]$ measures the fairness gap between groups g_i and g_j. The matrix satisfies $e_{ii} = 0$ and $e_{ij} = -e_{ji}$. To obtain a compact representation suitable for interpretation and learning, we summarize E using an aggregate magnitude measure. Since we are interested in deviation from fairness rather than direction, we consider absolute values and compute statistics over the upper triangular part of the matrix:

$$\mathrm{EP} = \frac{2}{k(k-1)} \sum_{i=1}^{k} \sum_{j=i+1}^{k} |e_{ij}| \tag{1}$$

$$\mathrm{DM} = \frac{2}{k(k-1)} \sum_{i=1}^{k} \sum_{j=i+1}^{k} (|e_{ij}| - \mathrm{EP})^2. \tag{2}$$

Equilibrium Point (EP) captures the average level of disparity, while Dispersion Measure (DM) measures its heterogeneity across group pairs.

4 Non-discrimination criteria

We consider three standard non-discrimination criteria, Independence, Separation, and Sufficiency, and associate each of them with a representative difference-based measure to ensure comparability across criteria. We define three matrices P, A, and C, computed on the same dataset S, corresponding respectively to Statistical Parity, Total Accuracy, and Calibration.

Independence requires statistical independence between predictions and group membership, i.e., $\hat{Y} \perp G$. We operationalize this criterion using Statistical Parity, defined as:

$$p_{ij} = \frac{|\sigma_{\hat{Y}=1}(S^i)|}{|S^i|} - \frac{|\sigma_{\hat{Y}=1}(S^j)|}{|S^j|}. \tag{3}$$

This measures differences in positive prediction rates between groups. The same formulation can be applied to the ground truth Y to quantify pre-existing disparities in the data.

Separation requires independence between predictions and sensitive attributes conditioned on the true label, i.e., $\hat{Y} \perp G \mid Y$. We adopt Total Accuracy as a representative confusion-matrix-based measure:

$$a_{ij} = \frac{|\sigma_{Y=\hat{Y}}(S^i)|}{|S^i|} - \frac{|\sigma_{Y=\hat{Y}}(S^j)|}{|S^j|}. \tag{4}$$

This captures differences in classification accuracy between groups.

Sufficiency criterion considers all the possible values of the predicted target variable $\hat{R}$. The Sufficiency non-discrimination criterion states that given a certain value of $\hat{R}$, the ground truth success variable Y must be independent of the sensible variable G, as defined in (5):

$$Y \perp G \mid \hat{R}. \tag{5}$$

In the next section, we illustrate how this criterion can be implemented.

5 Calibration

Calibration operationalizes the Sufficiency criterion by comparing group-specific success rates across values of the predicted score $\hat{R}$. Since $\hat{R}$ may take multiple values, we aggregate contributions over all outcomes using a weighted formulation that accounts for their empirical frequency. To ensure a coherent comparison across groups and score levels, we define group differences in a consistent order using the sign function $s : \mathcal{R} \to \{-1, 1\}$. This guarantees that positive and negative values are interpreted consistently across all $\hat{R}$ levels. The Calibration measure is then defined as:

$$w_{ij}^{\hat{r}} = \frac{|\sigma_{\hat{R}=\hat{r}}(S^i \cup S^j)|}{|S^i \cup S^j|} \tag{6}$$

$$n_i^{\hat{r}} = \begin{cases} \frac{|\sigma_{\hat{R}=\hat{r} \wedge Y=1}(S^i)|}{|\sigma_{\hat{R}=\hat{r}}(S^i)|} & \text{if } |\sigma_{\hat{R}=\hat{r}}(S^i)| \neq 0 \\ 0 & \text{otherwise} \end{cases} \tag{7}$$

$$c_{ij}^{\hat{r}} = s(\hat{r})\, w_{ij}^{\hat{r}} (n_i^{\hat{r}} - n_j^{\hat{r}}) \tag{8}$$

$$c_{ij} = \sum_{\hat{r} \in \hat{R}} c_{ij}^{\hat{r}}. \tag{9}$$

This formulation aggregates calibration disparities across all score values while preserving their relative contribution according to their frequency in the data. When computing $c_{ij}^{\hat{r}}$, different values of $\hat{R}$ may result in distinct scenarios depending on whether items from neither, one, or both groups are present. The most straightforward case occurs when the classifier assigns no items to a given value $\hat{R} = \hat{r}$, that is, when $|\sigma_{\hat{R}=\hat{r}}(S^i \cup S^j)| = 0$. In such a case, $w_{ij}^{\hat{r}} = 0$, implying that this specific value of $\hat{R}$ does not contribute further to the overall computation. A similarly simple situation arises when the classifier assigns items to both groups, i.e., $|\sigma_{\hat{R}=\hat{r}}(S^i)| \neq 0$ and $|\sigma_{\hat{R}=\hat{r}}(S^j)| \neq 0$. In this scenario, $c_{ij}^{\hat{r}}$ can be directly calculated, and its value contributes to the final value of c_{ij}. A more nuanced situation emerges when the classifier assigns the value $\hat{R} = \hat{r}$ exclusively to one group, for instance, when $|\sigma_{\hat{R}=\hat{r}}(S^i)| = 0$ and $|\sigma_{\hat{R}=\hat{r}}(S^j)| \neq 0$, or the opposite. This case requires additional consideration. Two main strategies can be followed: (i) treat the group with no assigned items as contributing zero, while retaining the value of $n_i^{\hat{r}}$ for the other group, or (ii) exclude all items corresponding to that particular value of $\hat{R}$, from both groups. Each option carries specific consequences. The first option introduces a degree of ambiguity, as two distinct cases become indistinguishable: (i) $n_i^{\hat{r}} = 0$ due to the absence of items labeled with $\hat{R} = \hat{r}$ for a given group ($|\sigma_{\hat{R}=\hat{r}}(S^i)| = 0$), or (ii) $n_i^{\hat{r}} = 0$ because no item within the group corresponds to a positive ground-truth outcome ($|\sigma_{\hat{R}=\hat{r} \wedge Y=1}(S^i)| = 0$). Furthermore, if fairness evaluation is intended to be performed separately for each value of $\hat{R}$, the absence of one group may be regarded as a violation of the method's assumptions preventing fairness computation. This interpretation is adopted in (Mancini et al., 2026a). However, removing all items linked to a specific value of $\hat{R}$ whenever one group is missing, may lead to significant information loss, especially with small datasets, thus introducing potential bias in the analysis. More importantly, it is unnecessary to treat each value of $\hat{R}$ as an isolated instance, since the aggregation function, implemented as a weighted mean, naturally reintroduces the contribution of items from the group present at that value of $\hat{R}$, thereby mitigating information loss. Consequently, as expressed in Equation (9), we advocate for adopting this latter approach.

In classification settings where the target variable exhibits high cardinality, that is, when the domain $\mathcal{R}$ contains a large number of distinct values, training a model to predict the exact outcome may lead to a substantial reduction in accuracy. As the number of possible classes increases, the prediction task becomes inherently more difficult, and misclassification errors are more likely to occur. This deterioration in predictive performance may, in turn, affect the reliability of subsequent fairness assessments, since Calibration is sensitive to classification errors and class imbalance. For this reason, when the cardinality $|\mathcal{R}|$ is large, it can be methodologically appropriate to replace the original target variable with a discretized version defined over a smaller number of ordered classes. Such discretization reduces the complexity of the prediction task, stabilizes accuracy estimates, and allows Calibration to be evaluated on more robust and interpretable groupings of outcomes, without altering the ordinal structure of the original variable.

6 Experiment on the Law School dataset

We evaluate the proposed framework on the Law School dataset (Wightman, 1998), which contains 22,027 admission records collected by the Law School Admissions Council (LSAC) across 163 U.S. law schools. The dataset has been widely used in fairness studies (Chzhen et al., 2020; Kearns et al., 2019; Kusner et al., 2017; Russell et al., 2017; Lahoti et al., 2020; Yang et al., 2020; Ruoss et al., 2020). We consider a preprocessed version with eight attributes (Table 2) and define the protected attribute as *race1*, with five groups: white (18425), black (1304), hispanic (1009), asian (886), and other (403). To model a realistic yet simplified law school admission scenario, we base the decision process on the LSAT scores reported in Table 2. The continuous score variable is discretized into six ordinal levels, defining the score domain $\mathcal{R}$, and a binary outcome $Y = s(R)$ is derived by applying a threshold at 29: levels 1–3, corresponding to LSAT intervals $[11,17)$, $[17,23)$, $[23,29]$, map to failure, while levels 4–6, corresponding to $(29,35.34)$, $[35.34,41.67)$, $[41.67,48]$, map to success. The experimental pipeline includes standard preprocessing (removal of missing and redundant data) and two classifiers with different inductive biases: eXtreme Gradient Boosting (XGBoost) and a MultiLayer Perceptron (MLP). Both models are evaluated using 5-fold cross-validation, and predictions are generated for both $\hat{R}$ and $\hat{Y}$. Fairness measures are computed in matrix form for both scenario 2 (1-vs-all) and scenario 3 (full multi-group setting), as defined in Section 3. Heatmaps are used to support interpretation, while EP and DM are reported for each matrix as in Equations (1) and (2). For clarity, only representative results are shown, while all computed results are used for analysis. We evaluate performance using the already introduced metrics: accuracy on $\hat{Y}$ (binary classification) and accuracy on $\hat{R}$ (6-class ordinal prediction). Since $\hat{R}$ has six classes, random accuracy is approximately 16.7%, making observed results substantially above this threshold meaningful in terms of predictive signal. Before presenting the classifiers, we report in Figure 1 the measure of Statistical Parity computed on the observed Y. This measure can be used as a reference level of the bias present in the dataset, a reference against which other measures can be compared to evaluate whether the classifier amplifies or mitigates the initial bias.

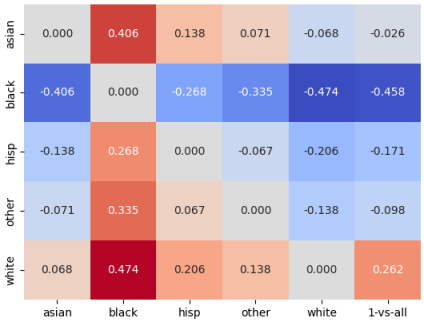


Figure 1: Statistical Parity computed on Y

6.1 XGBoost classifier

The first algorithm we used is XGBoost, namely the Python implementation (Chen and Guestrin, 2016)[1]. The prediction achieved an accuracy of 51.24% on $\hat{R}$ and a consequent accuracy of 94.00% on $\hat{Y}$. In Tables 3 and 4 are reported the measures of Statistical Parity and Total Accuracy, respectively. For an accessible read of results, heatmaps are also reported in Figures 2 and 3. Even at first glance, it is noticeable that for all the measures, the disadvantaged group is *black*. More specifically, the Statistical Parity matrix in Table 3 highlights a marked imbalance in the rate of positive predictions, with the *black* group consistently receiving substantially lower scores when compared to all other groups, both in pairwise comparisons and in the 1-vs-all setting. This suggests that, despite the high overall predictive performance, the classifier tends to favor *non-black* individuals in terms of selection outcomes. A similar pattern emerges from the Total Accuracy matrix in Table 4. Although differences are generally attenuated with respect to Statistical Parity, the *black* group still exhibits the lowest relative accuracy, indicating that the classifier performs worse on this subgroup. Overall, the joint analysis of Statistical Parity, Total Accuracy, and Calibration reveals that high aggregate accuracy does not prevent substantial fairness violations, and that the *black* group is systematically disadvantaged under all three fairness criteria. The comparison between true Statistical Parity (Figure 1) and predicted Statistical Parity (Table 3, Figure 2) shows that the algorithm preserves the relative ordering of advantaged and dis-

[1]Parameters for XGBoost:
`objective='multi:softprob', num_class=6, eval_metric='mlogloss', max_depth=4, learning_rate=0.05, n_estimators=300, subsample=0.8, colsample_bytree=0.8, min_child_weight=5, gamma=0, reg_lambda=1, random_state=42`

Attribute	Type	Values	Description
sex	Categorical	{1, 2}	female / male
DOB_yr	Numerical	[1910 - 1971]	year of birth
fulltime	Categorical	{1, 2}	work experience indicator
fam_inc	Categorical	{1,...,5}	family income category
parttime	Categorical	{0, 1}	part-time intention
race1	Categorical	{asian, black, hisp, white, other}	protected attribute
lsat	Numerical	[11 - 48]	LSAT score

Table 2: Law School Dataset attributes

	asian	black	hisp	other	white	1-vs-all
asian	0.000	0.577	0.141	0.089	-0.020	0.027
black	-0.577	0.000	-0.435	-0.488	-0.597	-0.586
hisp	-0.141	0.435	0.000	-0.053	-0.161	-0.121
other	-0.089	0.488	0.053	0.000	-0.108	-0.064
white	0.020	0.597	0.161	0.108	0.000	0.279
	EP: 0.267, DM: 0.047					

Table 3: Classifier: XGBoost; measure: Statistical Parity

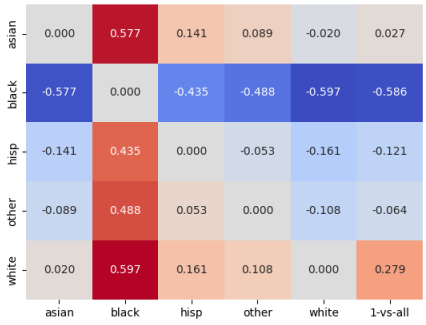


Figure 2: Classifier XGBoost; measure: Statistical Parity

advantaged groups, indicating that it does not alter which groups are favored. Nevertheless, the magnitude of the pairwise differences increases after classification, suggesting that the algorithm amplifies the pre-existing disparities present in the data rather than attenuating them.

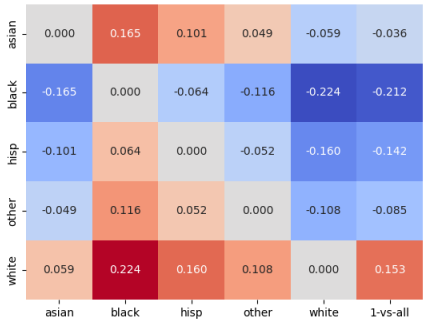


Figure 3: Classifier: XGBoost; measure: Total Accuracy

6.2 MultiLayer Perceptron classifier

In order to further investigate the relationship between predictive performance and fairness, we evaluated a neural network–based classifier. We used the Python MLPClassifier from the Sci-Kit Learn library [2]. This classifier achieves an accuracy of 50.79% on $\hat{R}$ leading to an accuracy of 92.63% on $\hat{Y}$, similar to XGBoost. The fairness analysis reveals patterns that are consistent with those already observed. Tables 5, and 6 report the matrices for Total Accuracy and Calibration, respectively. The Total Accuracy matrix in Table 5 reveals disparities across sensitive groups. In particular, the *black* group consistently exhibits negative values both in pairwise comparisons and in the 1-vs-all setting, indicating a substantially lower rate of positive predictions compared to all other groups. The relative accuracy for the *black* group is lower than that of the other groups but EP and DM denote a lower variability and a lower magnitude of the bias. In Figures 4 and 5 we present two heatmaps that give a visual interpretation of the discussed results. The Calibration results reported in Table 6 exhibit the most light fairness violations. Compared to XGBoost, the magnitude of calibration disparities is lower, highlighting a not particularly critical limitation of the neural classifier in terms of reliability across sensitive groups. Overall, the MLP classifier confirms

[2]Parameters for MLPClassifier:
`hidden_layer_sizes=(256, 128, 64), activation='relu', solver='adam', alpha=5e5, batch_size=128, learning_rate='adaptive', learning_rate_init=5e4, max_iter=500, tol=1e5, early_stopping=True, n_iter_no_change=20, random_state=42, verbose=False`

	asian	black	hisp	other	white	1-vs-all
asian	0.000	0.165	0.101	0.049	-0.059	-0.036
black	-0.165	0.000	-0.064	-0.116	-0.224	-0.212
hisp	-0.101	0.064	0.000	-0.052	-0.160	-0.142
other	-0.049	0.116	0.052	0.000	-0.108	-0.085
white	0.059	0.224	0.160	0.108	0.000	0.153
	EP: 0.110, DM: 0.003					

Table 4: Classifier: XGBoost; measure: Total Accuracy

	asian	black	hisp	other	white	1-vs-all
asian	0.000	0.185	0.103	0.061	-0.057	-0.032
black	-0.185	0.000	-0.081	-0.124	-0.242	-0.229
hisp	-0.103	0.081	0.000	-0.043	-0.160	-0.141
other	-0.061	0.124	0.043	0.000	-0.118	-0.093
white	0.057	0.242	0.160	0.118	0.000	0.160
	EP: 0.117, DM: 0.004					

Table 5: Classifier: MLP; measure: Total Accuracy.

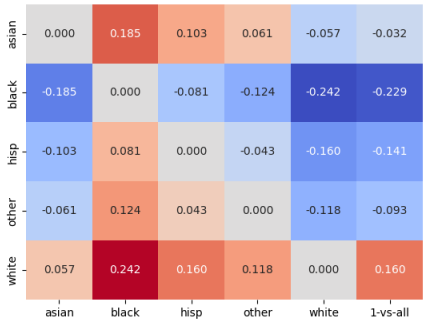


Figure 4: Classifier: MLP; measure: Total Accuracy

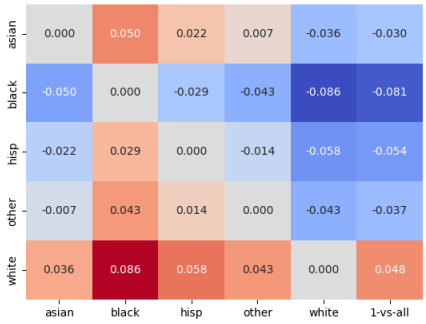


Figure 5: Classifier: MLP; measure: Calibration

the central finding of this study: predictive accuracy does not necessarily correspond to more biased outcomes. Even under stable and well-converged training conditions, the neural model reproduces and, in some cases, exacerbates the disparities observed in the data, as shown in the Statistical Parity measure.

6.3 Considerations

Several patterns emerge from the joint analysis. XGBoost achieves the highest predictive accuracy (51.24% on $\hat{R}$ and 94.00% on $\hat{Y}$), but it does not yield a significant fairness improvement: its Statistical Parity EP corresponds to 0.267 which is close to the EP of the MLP classifier, that is 0.271. The MLP exhibits the highest dispersion in Statistical Parity ($DM = 0.066$), suggesting more heterogeneous group disparities, yet it achieves the lowest Calibration with $EP = 0.039$ and $DM = 0.000$, indicating comparatively balanced reliability across groups. In general, higher predictive accuracy does not systematically correspond to improved fairness; in some cases, the most accurate models are also those with larger group disparities under certain criteria. Table 7 provides a compact comparative overview of the two classifiers and fairness criteria considered in the study. For each model, it reports EP and DM for all three fairness criteria, predictive accuracy, and a baseline Statistical Parity on Y as reference for disparity amplification or attenuation. The proposed framework could also be intended as a decision-support tool rather than a single fairness verdict. By presenting EP, DM, and accuracy side by side, practitioners can explicitly evaluate trade-offs between predictive performance and fairness constraints. Instead of selecting a model solely on aggregate accuracy, decision-makers can identify the configuration that best fits their context — for example, prioritizing lower Statistical Parity disparities or stronger Calibration across groups. In this sense, the method enables an informed and transparent selection process, where accuracy and fairness are jointly balanced according to the operational and eth-

	asian	black	hisp	other	white	1-vs-all
asian	0.000	0.050	0.022	0.007	-0.036	-0.030
black	-0.050	0.000	-0.029	-0.043	-0.086	-0.081
hisp	-0.022	0.029	0.000	-0.014	-0.058	-0.054
other	-0.007	0.043	0.014	0.000	-0.043	-0.037
white	0.036	0.086	0.058	0.043	0.000	0.048
	EP: 0.039, DM: 0.000					

Table 6: Classifier: MLP; measure: Calibration.

Fairness Measure	**EP**	**DM**	**Acc.** $\hat{R}$	**Acc.** $\hat{Y}$
Statistical Parity computed on Y	0.217	0.020		
XGBoost			**51.24%**	**94.00%**
Statistical Parity	0.267	0.047		
Total Accuracy	0.110	0.003		
Calibration	0.082	0.004		
MLPClassifier			**50.79%**	**92.63%**
Statistical Parity	0.271	0.066		
Total Accuracy	0.117	0.004		
Calibration	0.039	0.000		

Table 7: Fairness measures and accuracy comparison across models

ical requirements of the application domain.

7 Conclusions

Fairness in machine learning remains an open and context-dependent problem, requiring the translation of abstract ethical principles into measurable and operational quantities. This work proposed a unified framework for evaluating fairness across multiple sensitive groups, grounded in three non-discrimination criteria: Independence, Separation, and Sufficiency. The framework yields consistent and comparable measurements across groups while preserving interpretability and computational tractability. The experimental results on the Law School dataset show that the proposed measures are stable across different classifiers and remain informative in multi-group settings. This supports their applicability in realistic decision-making scenarios and highlights the importance of evaluating fairness beyond binary group comparisons. A specific contribution concerns the operationalization of Calibration under the Sufficiency criterion. We introduced a merit-based stratification approach that evaluates group disparities across score levels. Overall, the framework enables structured comparisons across fairness criteria, showing that Independence, Separation, and Sufficiency provide complementary perspectives rather than interchangeable ones. The results further confirm that fairness cannot be reduced to a single scalar metric, but should be treated as a multi-dimensional evaluation problem. Future work will focus on integrating these fairness measures directly into learning algorithms, particularly tree-based and boosting models, moving from post-hoc evaluation to fairness-aware optimization.

REFERENCES

Barocas, S., Hardt, M., and Narayanan, A. (2023). *Fairness and Machine Learning: Limitations and Opportunities*. MIT Press, Cambridge, MA.

Bird, S., Dudík, M., Edgar, R., Horn, B., Lutz, R., Milan, V., Sameki, M., Wallach, H., and Walker, K. (2020). Fairlearn: A toolkit for assessing and improving fairness in ai. *Microsoft, Tech. Rep. MSR-TR-2020-32*.

Calmon, F., Wei, D., Vinzamuri, B., Ramamurthy, K. N., and Varshney, K. R. (2017). Optimized pre-processing for discrimination prevention. *Advances in neural information processing systems*, 30.

Caton, S. and Haas, C. (2024). Fairness in machine learning: A survey. *ACM Computing Surveys*, 56(7):1–38.

Chen, T. and Guestrin, C. (2016). Xgboost: a scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 785–794, New York, NY, USA. Association for Computing Machinery.

Chen, Y., Giudici, P., Liu, K., and Raffinetti, E. (2024). Measuring fairness in credit ratings. *Expert Systems with Applications*, 258:125184.

Chouldechova, A. (2017). Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big data*, 5(2).

Chzhen, E., Denis, C., Hebiri, M., Oneto, L., and Pontil, M. (2020). Fair regression via plug-in estimator

and recalibration with statistical guarantees. *Advances in Neural Information Processing Systems*, 33:19137–19148.

Cortez, P. and Silva, A. M. G. (2008). Using data mining to predict secondary school student performance. In *Proceedings of 5th Annual Future Business Technology Conference, Porto, 2008*, pages 5–12, Ostend, Belgium. EUROSIS-ETI.

Deo, R. C. (2015). Machine learning in medicine. *Circulation*, 132(20):1920–1930.

Holmes, W. and Porayska-Pomsta, K. (2023). *The ethics of artificial intelligence in education*. Routledge Taylor, London.

Jacobs, A. Z. and Wallach, H. (2021). Measurement and fairness. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pages 375–385, New York, NY, USA. Association for Computing Machinery.

Kearns, M., Neel, S., Roth, A., and Wu, Z. S. (2019). An empirical study of rich subgroup fairness for machine learning. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, FAT* '19, pages 100–109, New York, NY, USA. Association for Computing Machinery.

Kusner, M. J., Loftus, J., Russell, C., and Silva, R. (2017). Counterfactual fairness. In *Advances in Neural Information Processing Systems*, volume 30, pages 4069–4079, Red Hook, NY, USA. Curran Associates, Inc.

Lahoti, P., Beutel, A., Chen, J., Lee, K., Prost, F., Thain, N., Wang, X., and Chi, E. (2020). Fairness without demographics through adversarially reweighted learning. In *Advances in Neural Information Processing Systems*, volume 33, pages 728–740, Red Hook, NY, USA. Curran Associates, Inc.

Lohia, P. K., Ramamurthy, K. N., Bhide, M., Saha, D., Varshney, K. R., and Puri, R. (2019). Bias mitigation post-processing for individual and group fairness. In *Icassp 2019-2019 Ieee International Conference on Acoustics, Speech and Signal Processing (Icassp)*, pages 2847–2851.

Loureiro, R. B., Pagano, T. P., Lisboa, F. V. N., Nascimento, L. F. S., Oliveira, E. L. S., Winkler, I., and Sperandio Nascimento, E. G. (2025). Correlation-based methods for representative fairness metric selection: An empirical study on efficiency and caveats in model evaluation. *Expert Systems with Applications*, 268:126344.

Makhlouf, K., Zhioua, S., and Palamidessi, C. (2021). On the applicability of machine learning fairness notions. *SIGKDD Explor. Newsl.*, 23(1):14–23.

Mancini, M., Merlini, D., and Verri, M. C. (2026a). Fairness measures for educational datasets. In Cerrato, M., Kalinauskaitė, D., Lukoševičius, M., Pechenizkiy, M., and Šutienė, K., editors, *Machine Learning and Principles and Practice of Knowledge Discovery in Databases*, pages 435–452, Cham. Springer Nature Switzerland.

Mancini, M., Merlini, D., and Verri, M. C. (2026b). Operationalizing calibration for fair educational artificial intelligence. In *Artificial Intelligence with and for Learning Sciences. Past, Present, and Future Horizons*, pages 183–198, Cham, Switzerland. Springer.

Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., and Galstyan, A. (2021). A survey on bias and fairness in machine learning. *ACM Computing Surveys*, 54(6).

Mukerjee, A., Biswas, R., Deb, K., and Mathur, A. P. (2002). Multi–objective evolutionary algorithms for the risk–return trade–off in bank loan management. *International Transactions in Operational Research*, 9(5):583–597.

Pessach, D. and Shmueli, E. (2022). A review on fairness in machine learning. *ACM Computing Surveys*, 55(3).

Purificato, E., Lorenzo, F., Fallucchi, F., and De Luca, E. W. (2023). The use of responsible artificial intelligence techniques in the context of loan approval processes. *International Journal of Human–Computer Interaction*, 39(7):1543–1562.

Reddy, D. J. M., Regella, S., and Seelam, S. R. (2020). Recruitment prediction using machine learning. In *2020 5th International Conference on Computing, Communication and Security (ICCC)*, pages 1–4.

Ruoss, A., Balunovic, M., Fischer, M., and Vechev, M. (2020). Learning certified individually fair representations. In *Advances in Neural Information Processing Systems*, volume 33, pages 7584–7596, Red Hook, NY, USA. Curran Associates, Inc.

Russell, C., Kusner, M. J., Loftus, J., and Silva, R. (2017). When worlds collide: Integrating different counterfactual assumptions in fairness. In *Advances in Neural Information Processing Systems*, volume 30, pages 6414–6423., Red Hook, NY, USA. Curran Associates, Inc.

Saleiro, P., Kuester, B., Hinkson, L., London, J., Stevens, A., Anisfeld, A., Rodolfa, K. T., and Ghani, R. (2018). Aequitas: A bias and fairness audit toolkit. arXiv preprint arXiv:1811.05577.

Soremekun, E., Papadakis, M., Cordy, M., and Le Traon, Y. (2025). Software fairness: An analysis and survey. *ACM Comput. Surv.*, 58(3).

Wightman, L. F. (1998). Lsac national longitudinal bar passage study. lsac research report series. Technical report, Law School Admission Council, Newtown, PA.

Yang, F., Cisse, M., and Koyejo, S. (2020). Fairness with overlapping groups; a probabilistic perspective. In *Advances in Neural Information Processing Systems*, volume 33, pages 4067–4078, Red Hook, NY, USA. Curran Associates, Inc.

Zhang, Y., Bellamy, R. K. E., Singh, M., and Liao, Q. V. (2020). Introduction to ai fairness. In *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems*, pages 1–4.

Zlatkin-Troitschanskaia, O., Schlax, J., Jitomirski, J., Happ, R., Kühling-Thees, C., Brückner, S., and Pant, H. A. (2019). Ethics and fairness in assessing learning outcomes in higher education. *Higher Education Policy*, 32:537–556.