The Impact of Model Selection Metrics during Hyperparameter Tuning on Algorithmic Fairness: An Empirical Study

Bianca Matos de Barros[1], Diego Dimer Rodrigues[1], Gabriela Bellardinelli Oliveira[1], Mariana Recamonde-Mendoza[1,2]

[1]Instituto de Informática – Universidade Federal do Rio Grande do Sul (UFRGS)
Caixa Postal 15.064 – 91501-970 – Porto Alegre – RS – Brasil

[2]Núcleo de Bioinformática – Hospital de Clínicas de Porto Alegre (HCPA)
Av. Protásio Alves, 211, Santa Cecília – 90035-903 – Porto Alegre – RS – Brasil

{bmbarros, ddrodrigues, gboliveira, mrmendoza}@inf.ufrgs.br

Abstract. *The growing use of machine learning in high-stakes domains raises concerns about fairness. The role of optimization metrics in shaping these outcomes remains underexplored. Using a controlled setup, this study investigates how seven performance metrics used for hyperparameter tuning and model selection affect fairness outcomes across five benchmark datasets. Results show that metrics are not neutral: recall-based optimization yields higher disparities, while precision and specificity lead to more balanced outcomes, with PR-AUC showing intermediate behavior. Overall, metric choice influences fairness, but outcomes are largely driven by dataset characteristics, with optimization redistributing errors rather than eliminating bias.*

1. Introduction

Machine learning (ML) models are increasingly deployed in high-stakes decision-making domains such as finance, criminal justice, education, and healthcare. In these contexts, predictive systems are often used to support or automate decisions such as credit approval, risk assessment, admission, or access to specialized care, thereby directly influencing the allocation of resources, opportunities, and services [O'Neil 2017, Mehrabi et al. 2021]. However, real-world data is often imbalanced and reflects historical and structural inequalities, which can cause predictive models to achieve systematically worse performance for underrepresented or historically marginalized populations.

When such performance disparities affect decisions, they may result in unequal access to benefits or increased exposure to harm for certain groups [Castelnovo et al. 2022]. In healthcare, such disparities violate health equity principles. A notable example is a widely-used algorithm that encoded structural inequalities by using historical costs as a proxy for medical need [Obermeyer et al. 2019]. Because Black patients face unequal access to care and consequently lower healthcare expenditures, the model systematically underestimated their medical needs, limiting their access to critical care programs.

A fundamental yet often overlooked aspect of this problem lies in the objective functions used during training. As Meredith Broussard observes, when a machine learns, it means that the machine has become more accurate at performing a single specific task

according to a specific metric that a person has defined [Broussard 2018]. This highlights the fact that ML models optimize toward explicitly defined metrics, which may prioritize either overall performance or specific types of errors (e.g., false positives or false negatives). Because these errors may be unevenly distributed across subpopulations, it is plausible to hypothesize that the choice of optimization metric can shape model behavior in ways that systematically advantage or disadvantage certain groups.

Despite extensive research on fairness-aware ML and bias mitigation strategies [Rabonato and Berton 2025], comparatively less attention has been given to how standard optimization objectives, such as accuracy, sensitivity or precision, interact with dataset characteristics (and any pre-existing biases) and influence fairness outcomes. There is limited systematic evidence on whether the choice of performance metric during model training amplifies, attenuates, or masks disparities across groups. This raises a fundamental question: to what extent are fairness disparities driven by data characteristics versus by modeling decisions, particularly the choice of optimization metric?

Motivated by this gap, this work systematically investigates the role of metric choice in shaping fair outcomes in supervised predictive models. Specifically, we address four research questions: **(RQ1)** Does the choice of optimization metric influence disparities in predictive performance across sociodemographic subgroups? **(RQ2)** Which optimization metrics are more likely to produce inequitable predictions? **(RQ3)** Can optimization metrics amplify or mitigate pre-existing biases? **(RQ4)** What is the trade-off between predictive performance and fairness? By empirically analyzing these questions across five datasets with distinct characteristics and application domains, this study provides a systematic perspective on how evaluation metrics function as design choices that influence fairness outcomes in ML. The remainder of this paper is organized as follows. Section 2 presents the background and related work, Section 3 describes the methodology, Section 4 discusses the results, and Section 5 concludes the paper.

2. Related work

The formalization of fairness in ML has largely evolved along two perspectives: individual and group-based parity. Seminal work by [Dwork et al. 2012] introduced the principle of "treating similar individuals similarly," while [Hardt et al. 2016] popularized group fairness definitions such as Equalized Odds. Importantly, [Chouldechova 2017] showed that when base rates differ across groups, it is mathematically impossible to satisfy multiple fairness criteria at once. This result highlights that fairness is not a single notion, but a set of competing constraints that often require explicit trade-offs.

The literature on the origins of bias has moved beyond a sole focus on data to consider the entire modeling pipeline. [Suresh and Guttag 2021] provides a comprehensive view of bias sources, including historical, representation, measurement, and aggregation biases. More recent work shows that even when datasets are balanced, the inductive bias of learning algorithms and the choice of objective functions can still amplify disparities [Mehrabi et al. 2021].

Approaches for bias mitigation are commonly categorized into pre-, in-, and post-processing methods. For example, [Kamiran and Calders 2012] demonstrated that transforming data prior to model training can reduce bias, whereas [Agarwal et al. 2018] proposed incorporating fairness constraints directly into the optimization process. More

recently, fairness-aware hyperparameter optimization approaches have emerged, either through intelligent hyperparameter search strategies or by explicitly incorporating fairness as an optimization objective during model tuning. These methods have shown promising results in improving fairness while maintaining predictive performance [Islam et al. 2021, Nadeem 2025]. Furthermore, more advanced multi-objective Bayesian optimization approaches have been proposed [Almasi et al. 2026], prioritizing hyperparameters according to their joint influence on fairness and predictive performance.

Collectively, these studies demonstrate that hyperparameter selection represents an important leverage point for improving fairness and that the optimization process itself can substantially affect model outcomes. However, existing approaches generally assume a fixed optimization objective and focus on mitigating observed disparities, rather than investigating whether the choice of optimization metric itself systematically contributes to fairness disparities, independently of algorithmic or hyperparameter choices.

The choice of evaluation metric becomes particularly relevant in settings with imbalanced data. [Buolamwini and Gebru 2018] demonstrated in the "Gender Shades" study that high overall accuracy can hide substantial performance gaps across subpopulations. While work on imbalanced learning recommends metrics such as the PR-AUC to better reflect performance [Gaudreault et al. 2021], there is still limited understanding of how these metrics influence fairness outcomes when used as optimization objectives.

In summary, existing work largely focuses on adding constraints or post-hoc adjustments, rather than examining whether disparities may stem from the mathematical properties of the metrics used during training. Thus, the effect of the optimization metric over model bias remains less explored. This paper addresses this gap by analyzing how the choice of performance metric shapes the distribution of errors across groups, framing metric selection as a key design decision in fair ML.

3. Methodology

This section presents the experimental methodology adopted in this study, including the datasets, model training procedure, hyperparameter optimization strategy, and fairness evaluation protocol. The experimental design was defined to ensure a controlled analysis of how optimization objectives influence predictive performance and equity outcomes.

3.1. Datasets and problem definition

We evaluate our research questions using five benchmark tabular datasets commonly employed in fairness ML studies, covering diverse high-stakes domains: COMPAS (criminal recidivism), Adult (income prediction), Heart (clinical diagnosis), Dropout (student retention), and Intersectional Bias (simulated healthcare access). The latter is a synthetic benchmark comprising demographic and clinical features, originally developed to explicitly assess intersectional disparities in mental health diagnoses. Each dataset contains a binary target variable and at least one protected attribute (e.g., race, sex, or gender), which correspond to sensitive characteristics along which disparities may arise. In general, protected attribute define subpopulations for fairness assessment and are commonly used to evaluate whether model performance differs across groups.

Favorable outcomes were defined according to domain conventions (e.g., predicting "no recidivism" or "no disease" as the positive class), enabling the computation of

group-based fairness metrics. Descriptive statistics and pre-training bias indicators are summarized in Table 1, including the number of instances (N), the class imbalance ratio (CI Ratio) and the Disparate Impact (DI) value for the original dataset (pre-training DI). Disparate Impact is used as a group-based measure to assess potential disparities between protected and unprotected groups, as discussed in Section 3.4. Intuitively, it captures whether favorable outcomes are distributed proportionally across groups defined by a protected attribute. In this table, DI is computed prior to model training, using the observed outcomes rather than model predictions, thus serving as a baseline indicator of pre-existing disparities in the data.

Table 1. Features and pre-training bias evaluation. The table reports the target and protected attributes, dataset size (N), class distribution, class imbalance (CI) ratio, and Disparate Impact (DI) prior to model training.

Dataset	Target Attribute	Protected Attribute	N	Class Dist (%)	CI Ratio	DI Pre-train.
COMPAS	Recidivism	Race	11,757	62.4 / 37.6	1.66	1.31
Heart	Heart Disease	Sex	302	45.7 / 54.3	1.19	1.68
Adult	Income	Race	32,537	24.1 / 75.9	3.15	0.60
Dropout	Academic Dropout	Gender	4,424	49.9 / 50.1	1.00	0.65
Intersectional	Clinical Diagnosis	Sex	11,000	48.9 / 51.1	1.05	0.69

3.2. Experimental setup and data splitting

All experiments were conducted using nested cross-validation over the complete datasets, without reserving a fixed holdout set. This choice ensures that all available data contribute to both training and evaluation while providing unbiased estimates of generalization performance. We employ a stratified nested cross-validation scheme, with a 3-fold stratified cross-validation in the inner loop for hyperparameter optimization and a 3-fold stratified cross-validation in the outer loop for performance estimation. Stratification is performed not only with respect to the target variable, but also considering the protected attribute(s), preserving the original class and group proportions across folds. The same random seed is used across all experiments to ensure reproducibility and fair comparison.

3.3. Models training and hyperparameter optimization

We evaluate two widely used learning algorithms for tabular data: Random Forests (RF) and Gradient Boosting Machines (GBM). These models are included to ensure that our findings are not specific to a single learning algorithm; however, our goal is not to compare their predictive performance, but rather to analyze how different optimization objectives influence fairness outcomes across models. For each algorithm, a predefined hyperparameter search space is specified and kept fixed across all experiments to ensure comparability. We varied the maximum tree depth (RF and GBM), the number of trees (RF), and the number of boosting iterations and learning rate (GBM). Hyperparameter tuning is performed using randomized search, evaluating 30 candidate configurations per model.

Model selection is conducted via cross-validation, where the best-performing configuration is chosen according to a specified optimization metric. The optimization objective is varied across experiments, while all other aspects of the training procedure remain unchanged. The set of optimization metrics includes Accuracy, Precision, Recall (Sensitivity), Specificity, ROC-AUC, PR-AUC, and Matthews Correlation Coefficient (MCC).

For each metric, hyperparameter tuning and model selection are performed independently, resulting in distinct models optimized for different performance criteria.

3.4. Performance evaluation and bias assessment

Model performance is evaluated in the outer loop of the nested cross-validation. Global predictive performance is reported using the same metric employed during optimization, along with complementary metrics for descriptive purposes. After model training, disparities in model behavior are quantified using three complementary fairness metrics: *Disparate Impact (DI)*, the *Sensitivity Gap* (also known as equal opportunity difference), and the *Average Absolute Odds Difference (AAOD)*.

In this stage, DI is computed based on model predictions, capturing the ratio of favorable predicted outcomes between unprivileged and privileged groups. The widely used 80% rule is adopted as a reference threshold for identifying potential fairness violations.

$$DI = \frac{P(\hat{Y} = 1 \mid S = u)}{P(\hat{Y} = 1 \mid S = p)} \tag{1}$$

The *Sensitivity Gap* measures the absolute difference in recall (true positive rate or TPR) between groups:

$$\text{Sensitivity Gap} = |\text{TPR}_u - \text{TPR}_p| \tag{2}$$

where TPR_u and TPR_p denote the true positive rates for the unprivileged and privileged groups, respectively. Thus, negative values indicate that model is capable of detecting more positives for the privileged group than for the unprivileged one. This metric captures disparities in the model's ability to correctly identify positive instances across groups, which is particularly relevant in high-stakes domains such as healthcare.

Average Absolute Odds Difference (AAOD) evaluates fairness by measuring the average absolute difference in False Positive Rates (FPR) and True Positive Rates (TPR) across groups. To maintain domain relevance, we define TPR and FPR relative to the favorable class ($\hat{Y} = favorable$):

$$AAOD = \frac{1}{2}\left[|FPR_u - FPR_p| + |TPR_u - TPR_p|\right] \tag{3}$$

where TPR represents the probability of a favorable prediction given a favorable ground truth, and FPR represents the probability of a favorable prediction given an unfavorable ground truth. A value of $AAOD = 0$ represents perfect equality of odds, where error distributions are identical for both protected and non-protected individuals. This metric captures disparities in both false positive and true positive rates across groups, providing a comprehensive view of error asymmetry.

4. Results and Discussion

This section presents the empirical findings guided by the research questions, starting with overall predictive performance and progressively examining disparities, bias dynam-

ics, and trade-offs between performance and fairness. Additional data is available in our online repository[1].

4.1. Overall Predictive Performance

Across all datasets and optimization strategies, both GBM and RF models achieved consistently high predictive performance, with relatively limited variation in global metrics such as accuracy and recall (Figure 1). Model choice (GBM vs. RF) had a limited impact on global performance, with both algorithms exhibiting similar central tendencies and dispersion across evaluation metrics. Moreover, in general, the choice of optimization metric did not lead to substantial differences in global performance rankings. Although some variation is observed, mainly for datasets Adult and Heart, models optimized for accuracy, MCC, PR-AUC, precision, recall, ROC-AUC, or specificity yielded comparable distributions of overall performance within each dataset. Although we illustrate this behavior using accuracy and recall, similar patterns were consistently observed across all evaluated metrics.

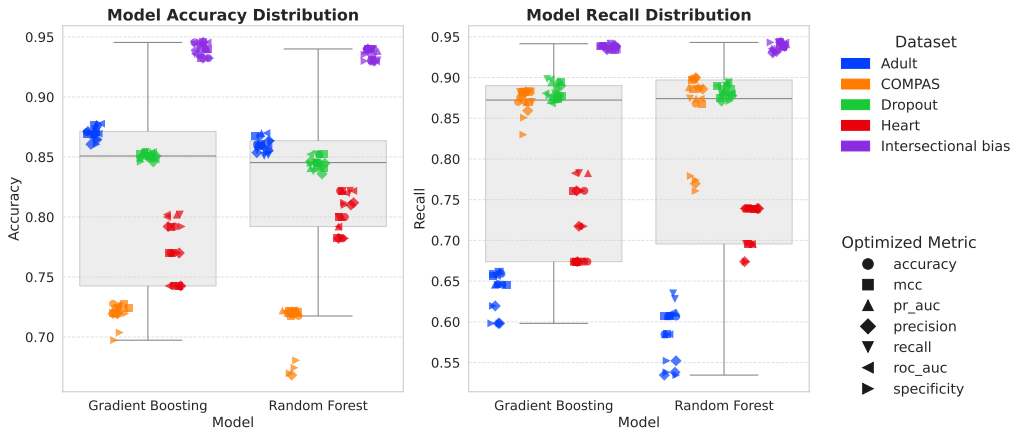

Figure 1. Overall model performance in terms of accuracy and recall. Colors represent datasets, and marker shapes indicate the optimization metric.

However, a closer inspection reveals that global performance aggregates conceal substantial heterogeneity across datasets. For example, datasets such as COMPAS consistently exhibit lower performance across multiple metrics, whereas the intersectional bias dataset shows near-ceiling performance. Moreover, variation across models optimized for different metrics also depend on the dataset being analyzed. This suggests that, at the aggregate level, performance is more strongly influenced by dataset characteristics than by either the learning algorithm or the optimization objective. Importantly, this aggregated view does not capture how predictive performance is distributed across sociodemographic subgroups. As we show in the following sections, optimization choices that appear equivalent at the global level can lead to markedly different, and sometimes inequitable, outcomes when examined at a finer granularity.

4.2. RQ1 - Does the choice of optimization metric influence disparities in predictive performance across sociodemographic subgroups?

The empirical results of this study reveal that the choice of optimization metric influences disparities in predictive performance across sociodemographic subgroups, although its

[1] https://github.com/mlab-inf-ufrgs/semish2026_OptimizationMetrics-Bias

Table 2. Fairness metrics for distinct datasets and optimization metrics. Mean and standard deviation consider results for both algorithms, GBM and RF.

Dataset	Fairness Metric	Accuracy	MCC	PR AUC	Precision	Recall	ROC AUC	Specificity
Adult	AAOD	0.045 ± 0.019	0.047 ± 0.020	0.047 ± 0.017	0.030 ± 0.013	0.053 ± 0.015	0.048 ± 0.019	0.030 ± 0.013
	Disparate Impact	0.558 ± 0.038	0.552 ± 0.044	0.554 ± 0.034	0.589 ± 0.029	0.540 ± 0.033	0.548 ± 0.040	0.589 ± 0.029
	Sensitivity Gap	-0.063 ± 0.032	-0.065 ± 0.034	-0.066 ± 0.029	-0.043 ± 0.022	-0.068 ± 0.024	-0.068 ± 0.031	-0.043 ± 0.022
COMPAS	AAOD	0.126 ± 0.013	0.127 ± 0.011	0.123 ± 0.011	0.124 ± 0.012	0.125 ± 0.009	0.125 ± 0.009	0.107 ± 0.010
	Disparate Impact	0.809 ± 0.017	0.808 ± 0.014	0.813 ± 0.014	0.808 ± 0.018	0.811 ± 0.011	0.812 ± 0.011	0.816 ± 0.011
	Sensitivity Gap	-0.123 ± 0.013	-0.123 ± 0.013	-0.121 ± 0.010	-0.128 ± 0.017	-0.121 ± 0.009	-0.121 ± 0.010	-0.124 ± 0.014
Dropout	AAOD	0.091 ± 0.026	0.091 ± 0.026	0.088 ± 0.021	0.089 ± 0.022	0.087 ± 0.023	0.091 ± 0.024	0.091 ± 0.025
	Disparate Impact	1.657 ± 0.094	1.657 ± 0.094	1.642 ± 0.073	1.647 ± 0.075	1.638 ± 0.083	1.658 ± 0.086	1.662 ± 0.082
	Sensitivity Gap	0.086 ± 0.029	0.086 ± 0.029	0.085 ± 0.024	0.087 ± 0.029	0.080 ± 0.024	0.089 ± 0.025	0.089 ± 0.031
Heart	AAOD	0.160 ± 0.060	0.160 ± 0.060	0.137 ± 0.046	0.152 ± 0.063	0.158 ± 0.063	0.120 ± 0.036	0.152 ± 0.063
	Disparate Impact	0.464 ± 0.228	0.464 ± 0.228	0.495 ± 0.194	0.472 ± 0.217	0.460 ± 0.221	0.510 ± 0.181	0.472 ± 0.217
	Sensitivity Gap	-0.007 ± 0.153	-0.007 ± 0.153	-0.011 ± 0.148	0.039 ± 0.118	-0.011 ± 0.148	0.031 ± 0.112	0.039 ± 0.118
Intersectional Bias	AAOD	0.045 ± 0.007	0.045 ± 0.007	0.045 ± 0.008	0.044 ± 0.006	0.047 ± 0.005	0.045 ± 0.008	0.044 ± 0.006
	Disparate Impact	1.553 ± 0.027	1.553 ± 0.027	1.555 ± 0.030	1.557 ± 0.026	1.551 ± 0.028	1.554 ± 0.029	1.557 ± 0.026
	Sensitivity Gap	-0.007 ± 0.014	-0.007 ± 0.014	-0.006 ± 0.015	-0.004 ± 0.013	-0.009 ± 0.016	-0.005 ± 0.015	-0.004 ± 0.013

impact is more pronounced in the variability and distribution of disparities than in their average magnitude, as shown in Table 2. Across all configurations, subgroup performance gaps are consistently observed, suggesting that inequities persist regardless of the optimization objective. This is expected, since experiments do not implement any strategy for bias mitigation. Our analysis demonstrates that no single optimization metric is capable of overcoming the pre-existing structural biases.

4.3. RQ2: Which optimization metrics are more likely to produce inequitable predictions?

Across all fairness metrics, we observe that certain optimization objectives are consistently associated with higher levels of disparity. In particular, recall-based optimization leads to the largest deviations in fairness, yielding higher AAOD values, greater sensitivity gaps, and DI values further from parity. Metrics such as accuracy, MCC, and ROC-AUC also exhibit relatively high levels of unfairness, though to a lesser extent. Optimization based on PR-AUC presents an intermediate behavior, generally resulting in moderate disparity levels without consistently ranking among either the best or worst-performing strategies. In contrast, optimization based on precision and specificity tends to produce lower disparity levels across datasets. These findings indicate that optimization objectives do not merely influence predictive performance, but can systematically impact disparities depending on the type of error they prioritize.

A plausible explanation for these findings lies in the asymmetric way in which optimization objectives redistribute prediction errors across subgroups. Metrics such as recall prioritize maximizing the true positive rate, encouraging the model to reduce false negatives globally. However, due to differences in group representation, class distribution, and feature separability, these improvements are not uniform across subgroups, leading to increased disparities. Optimization based on PR-AUC, which balances precision and recall, results in a more moderated redistribution of errors, typically avoiding the extreme disparities associated with recall while not enforcing the same level of constraint as precision- or specificity-based objectives.

In contrast, optimization objectives such as precision and specificity impose stricter constraints on predictions, particularly by limiting false positives, which tends to result in more balanced error distributions across groups. Additionally, the observed

differences reflect an interaction between the optimization objective and dataset characteristics, suggesting that fairness outcomes are shaped by both the type of error being optimized and the underlying structure of the data.

To facilitate comparison across optimization metrics, we rank them for each dataset and fairness metric based on their level of disparity, as shown in Figure 2. Lower rank values mean lower disparity. This ranking highlights consistent patterns in how optimization objectives affect fairness, while also revealing dataset-specific variations that are less apparent when analyzing raw metric values. When aggregating fairness metrics across datasets, models, and folds, we observe that the choice of optimization metric leads to modest but consistent differences in disparity levels. Optimization based on recall, accuracy, MCC, and ROC-AUC tends to produce higher rankings for AAOD and sensitivity gap values, indicating increased disparities across subgroups. In contrast, specificity, precision, and PR-AUC are associated with lower ranking, meaning lower disparity levels on average. Differences in DI are comparatively small across optimization strategies, suggesting that this metric is less sensitive to the choice of objective function. The complete statistical tables are available in our repository.

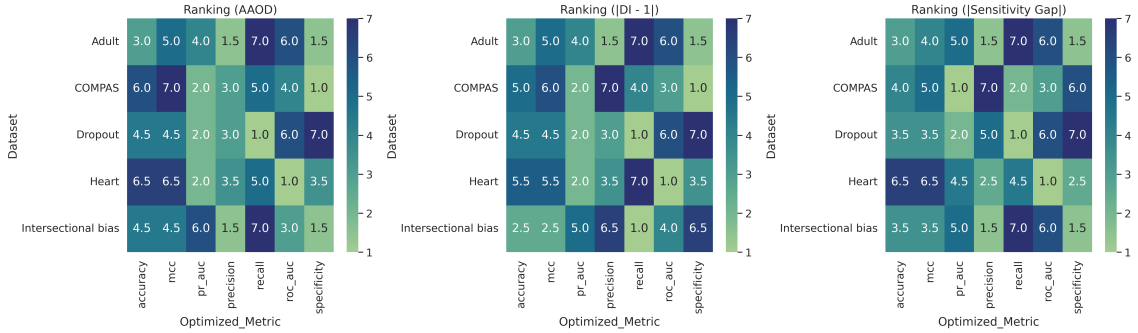

Figure 2. Ranking of optimization metrics by fairness level across datasets and metrics. Lower ranks indicate lower disparity.

Our empirical results extend the findings of [Buolamwini and Gebru 2018]. While their work demonstrated how aggregate performance metrics can obscure disparities in post-hoc evaluation, our study shifts the focus upstream into the model development lifecycle. We show the quantifiable impact of optimization metric selection during hyperparameter tuning, proving that equity distortions are actively shaped before deployment. Furthermore, the chosen optimization objective must be explicitly documented and justified within standard transparency mechanisms, allowing stakeholders to fully understand how prediction errors were structurally distributed across subgroups prior to deployment.

4.4. RQ3: Can optimization metrics amplify or mitigate pre-existing biases?

To investigate the effect of optimization metrics on bias, we analyze DI as an indicator of changes in selection rates across subgroups. The dumbbell analysis for recall-based optimization (Figure 3) reveals substantial shifts in DI across all datasets, often moving predictions further away from parity (DI = 1). Depending on the dataset, this shift can either amplify existing biases or even reverse their direction, indicating that optimization does not preserve initial bias patterns and tends to increase disparity. Similar qualitative patterns were observed across other optimization metrics. Notably, none of the tested

metrics succeeded in bringing the Disparate Impact within 80% rule threshold ($[0.8, 1.25]$) across all datasets, reinforcing the conclusion that standard optimization acts primarily as a mechanism of error redistribution rather than a reliable bias mitigation strategy.

Despite these shifts, the heatmap in Figure 4 shows that the direction of bias change is largely determined by the dataset rather than the optimization metric. Within each dataset, all optimization objectives produce shifts of similar sign and magnitude, indicating consistent amplification or mitigation patterns. For instance, Dropout and Intersectional Bias exhibit strong amplification, whereas COMPAS and Heart show systematic mitigation. Differences between metrics are comparatively small, suggesting that optimization primarily modulates, rather than determines, bias direction.

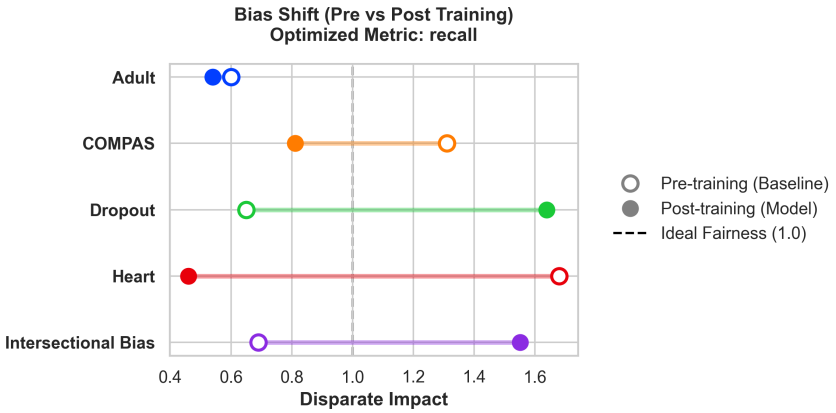

Figure 3. Disparate Impact (DI) before and after model training under recall-based optimization.

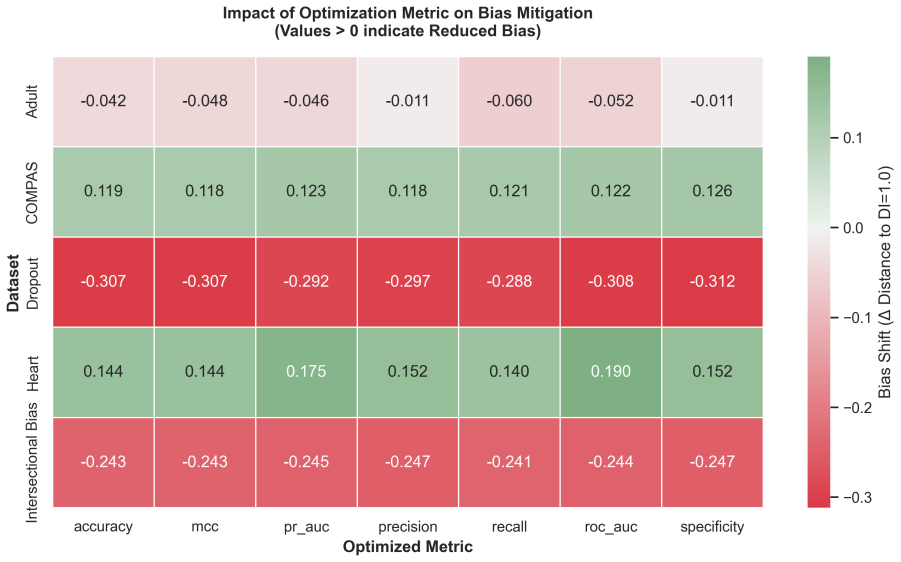

Figure 4. Analysis of the direction of bias change, in terms of Disparate Impact (DI), across datasets and optimization metrics.

Overall, these results indicate that optimization acts as a mechanism of bias redistribution constrained by the data distribution. While objectives such as recall can induce larger shifts, fairness outcomes are mainly driven by dataset characteristics. DI captures

the direction and magnitude of bias but remains relatively insensitive to optimization choices. Therefore, it should be interpreted alongside error-based metrics (e.g., sensitivity gap and AAOD), which better reflect how optimization redistributes errors across groups. These metrics show that recall-based optimization tends to amplify disparities in error rates across subgroups, while optimization based on precision and specificity leads to more balanced error distributions across groups. PR-AUC provides a compromise, reducing extreme disparities but not consistently achieving the lowest unfairness.

4.5. RQ4: What is the trade-off between predictive performance and fairness?

Finally, we analyze whether the expected trade-off between predictive performance and fairness holds for models optimized using different objective functions. The Figure 5 illustrates the trade-off between predictive performance (based on accuracy, for simplicity) and fairness, showing that higher accuracy does not consistently correspond to lower disparity. Instead, the points form a Pareto-like frontier in which improvements in accuracy are generally associated with similar or increased levels of AAOD. While some configurations achieve relatively high accuracy with moderate disparity, none simultaneously reach both high accuracy and low disparity, indicating an inherent tension between these objectives. Moreover, the dispersion of points along the frontier suggests that the choice of optimization metric primarily determines where a model lies within this trade-off space, rather than overcoming it. Overall, the results reinforce that improving predictive performance often comes at the cost of fairness, and that optimization strategies mainly redistribute this trade-off rather than eliminate it.

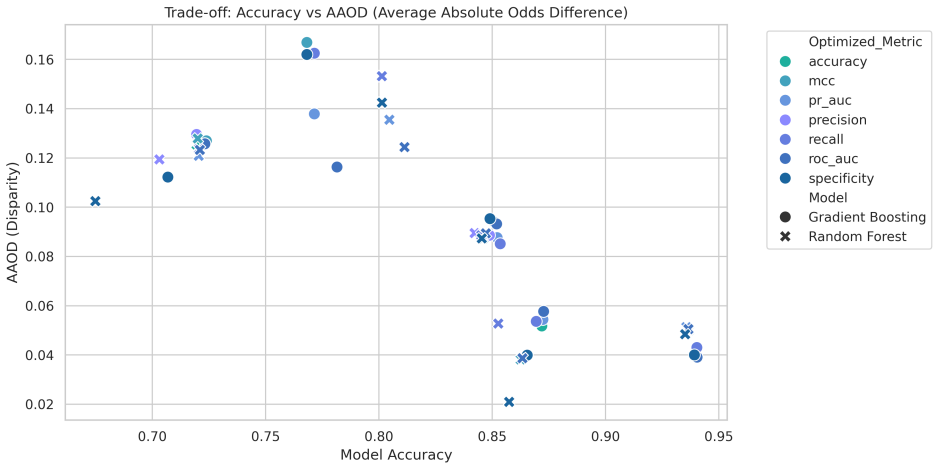

Figure 5. Pareto frontiers mapping global Accuracy against the Average Absolute Odds Difference (AAOD).

5. Conclusion

This study provides a systematic and empirical investigation of a question that is often assumed but rarely demonstrated: how the choice of optimization metric shapes algorithmic results, not only in terms of performance, but also of model fairness. Our work offers direct empirical evidence, across five distinct high-stakes domains, that optimization objectives play a measurable and non-trivial role in redistributing disparities. Our results show that different objective functions (e.g., Accuracy, PR-AUC, Recall) lead to distinct fairness outcomes across subgroups. In particular, recall-based optimization is consistently

associated with higher disparity levels, while precision- and specificity-based objectives tend to produce more balanced error distributions.

Furthermore, our analysis of Pareto frontiers demonstrates that optimization metrics do not resolve the fundamental trade-off between predictive performance and fairness. Instead, they determine where models lie along an inherently constrained frontier, largely shaped by the underlying data distribution. This reinforces that fairness outcomes cannot be decoupled from data characteristics and that optimization acts primarily as a mechanism of error redistribution rather than bias removal.

From a practical standpoint, our results suggest several recommendations. First, optimization metrics should not be treated as neutral defaults, but as design decisions with fairness implications. Second, reliance on recall as a sole optimization objective should be carefully considered in high-stakes settings, given its association with increased disparities. Third, practitioners should evaluate fairness using multiple complementary metrics, combining selection-based and error-based perspectives. Fourth, fairness evaluation should be integrated throughout the model development process rather than applied only post hoc. Fifth, comparing multiple optimization objectives under the same data conditions provides a more informative view of fairness trade-offs. Sixth, dataset characteristics, such as class imbalance and subgroup representation, should be explicitly analyzed, as they play a dominant role in shaping outcomes. Finally, when fairness requirements are critical, metric selection alone is insufficient and should be complemented with fairness-aware modeling strategies.

We also observe that standard optimization procedures do not consistently reduce disparities and may, in some cases, increase them. This reinforces that fairness is not an emergent property of performance optimization, but a result of deliberate modeling and evaluation choices. Overall, our findings contribute empirical evidence that the impact of optimization metrics on fairness is real, measurable, and context-dependent, and should be explicitly considered in the design and deployment of ML systems.

While this study is limited to tabular data, binary protected attributes, and two tree-based models, our findings highlight the need for future work exploring neural networks and cost-sensitive learning to natively penalize discriminatory outcomes. By assigning asymmetric, heavier penalties to prediction errors (false positives or false negatives) that disproportionately affect marginalized groups, the optimization process could theoretically be forced to prioritize equity over aggregate performance. Furthermore, future research should investigate multi-objective optimization strategies that directly embed fairness constraints into the loss function, comparing these in-processing interventions against the baseline error redistribution patterns identified in this study to determine the absolute limits of algorithmic fairness in structurally biased environments.

Acknowledgments

This study was partially funded by CAPES - Finance Code 001, FAPERGS [Proj. No. 22/2551-0000390-7] and CNPq [Proj. No. 308075/2021-8].

References

Agarwal, A., Beygelzimer, A., Dudík, M., Langford, J., and Wallach, H. (2018). A reductions approach to fair classification. In *International Conference on Machine*

Learning (ICML), pages 60–69. PMLR.

Almasi, M., Nezami, N., Di_Carlo, F., Asudeh, A., and Anahideh, H. (2026). Adaptive pareto exploration (apex) for fairness-aware hyperparameter optimization in fairpilot. *Information and Software Technology*, 189(C).

Broussard, M. (2018). *Artificial unintelligence: How computers misunderstand the world*. mit Press.

Buolamwini, J. and Gebru, T. (2018). Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on Fairness, Accountability and Transparency (FAccT)*, pages 77–91. PMLR.

Castelnovo, A., Crupi, R., Greco, G., Regoli, D., Penco, I. G., and Cosentini, A. C. (2022). A clarification of the nuances in the fairness metrics landscape. *Scientific reports*, 12(1):4209.

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

Dwork, C., Hardt, M., Pitassi, T., Reingold, O., and Zemel, R. (2012). Fairness through awareness. In *Proceedings of the 3rd innovations in theoretical computer science conference*, pages 214–226.

Gaudreault, J.-G., Branco, P., and Gama, J. (2021). An analysis of performance metrics for imbalanced classification. In *International Conference on Discovery Science*, pages 67–77. Springer.

Hardt, M., Price, E., and Srebro, N. (2016). Equality of opportunity in supervised learning. In *Advances in neural information processing systems (NIPS)*, pages 3315–3323.

Islam, R., Pan, S., and Foulds, J. R. (2021). Can we obtain fairness for free? In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, pages 586–596.

Kamiran, F. and Calders, T. (2012). Data preprocessing techniques for classification without discrimination. *Knowledge and Information Systems*, 33(1):1–33.

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):1–35.

Nadeem, M. A. (2025). Optimizing fairness in machine learning: A hyperparameter tuning approach. In *2025 IEEE International Conference on Omni-layer Intelligent Systems (COINS)*, pages 1–5. IEEE.

Obermeyer, Z., Powers, B., Vogeli, C., and Mullainathan, S. (2019). Dissecting racial bias in an algorithm used to manage the health of populations. *Science*, 366:447–453.

O'Neil, C. (2017). *Weapons of math destruction: How big data increases inequality and threatens democracy*. Crown.

Rabonato, R. T. and Berton, L. (2025). A systematic review of fairness in machine learning. *AI and Ethics*, 5(3):1943–1954.

Suresh, H. and Guttag, J. (2021). A framework for understanding sources of harm throughout the machine learning life cycle. In *Proceedings of the 1st ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization*, pages 1–9.