Applied Soft Computing Journal 181 (2025) 113426

Contents lists available at ScienceDirect

Applied Soft Computing

journal homepage: www.elsevier.com/locate/asoc

Optimizing fairness and utility in healthcare machine learning models☆

Maíra Blumer Fatoretto [a,*], Gökhan Özbulak [b,c], André Anjos [b], Lilian Berton [a]

[a] Institute of Science and Technology (ICT) - Federal University of São Paulo, Cesare Monsueto Giulio Lattes Avenue, 1201 - Eugênio de Melo, São José dos Campos, 12247-014, São Paulo, Brazil

[b] Idiap Research Institute, Rue Marconi 19, 1920, Martigny, Switzerland

[c] École Polytechnique Fédérale de Lausanne (EPFL), Rte Cantonale, 1015, Lausanne, Switzerland

ARTICLE INFO

Keywords:
Fairness
Machine learning
Metrics
Optimization

ABSTRACT

Demographic fairness or equity is a crucial aspect of machine learning models, particularly in critical domains such as healthcare, where errors can have severe consequences. A fair model avoids making distinctions between groups with different sensitive or protected attributes. Although several metrics are available to measure fairness and ensure equity between groups, proposed solutions must uphold fairness without compromising the utility of these models. Optimization problems that incorporate both fairness and utility information can help find the best machine learning model. Mathematical programming emerges as a valuable tool in this context. One approach is to use optimization functions with constraints, imposing a maximum difference between groups. In this sense, considering multiple constraints that encompass various sensitive attributes present in the dataset when adjusting the models is essential to ensure intersectional fairness, minimize hidden biases, and promote equitable decisions in diverse contexts. In this work, we propose to constrain the minimization of the loss function with multiple fairness-related metrics, ensuring that fairness metrics do not exceed a maximum limit concerning the impartiality of the decision boundary. We use metrics derived from Pareto fronts, a method used in multi-objective optimization, adapting it for single-objective optimization and incorporating fairness characteristics into the constraints. The points observed in this graph use different fairness thresholds. We compare our proposed model with existing literature and demonstrate the convergence of our model to logistic regression with simulated data. Furthermore, we apply this strategy to health-related datasets and other domains present in most fairness and optimization articles. As a result, we found that, using the proposed metrics, our model performs better, even with imbalanced data concerning sensitive attributes and smaller datasets.

1. Introduction

The increasing use of Machine Learning (ML) in society brings transformative opportunities and conveniences but raises critical concerns about fairness and equity. Fairness in ML refers to the ethical principle of ensuring that algorithms and models do not discriminate or exhibit bias against individuals or groups based on certain characteristics, such as ethnicity, gender, age, or other protected attributes [1]. For example, biases present in training data can lead to biased models. Algorithms can introduce biases based on the chosen optimization objectives, hyperparameters, or inherent characteristics. If features correlated with sensitive attributes (e.g., gender, ethnicity) are included, the model might learn to make predictions based on these attributes, resulting in unfair outcomes. Additionally, when data collection or sampling is done in a skewed manner, it can affect the training process and lead to distorted predictions [2].

Researchers and practitioners are actively developing techniques for mitigating bias and promoting fairness in ML. This includes methods for debiasing data, designing fair algorithms, and evaluating models concerning fairness metrics [3–5]. Common metrics used to quantify fairness in ML include disparate impact, equalized odds, and statistical parity [6,7]. However, it is a challenge to consider and balance multiple metrics simultaneously. Different fairness metrics may sometimes be at odds with each other. For example, optimizing for demographic parity might lead to discrepancies in predictive performance across different groups. Measuring fairness in machine learning is complex due to the various definitions of fairness appropriate for different contexts.

Research on fairness criteria requires precise definitions to guide their application in both mathematical models and real-world scenarios. Therefore, finding a balance requires a careful and nuanced approach, often involving trade-offs that must be tailored to the specific

☆ This document is the results of the research project funded by the São Paulo Research Foundation Grants grants 21/14725-3 and 23/12468-9.

* Corresponding author.
E-mail addresses: mairafatoretto@gmail.com (M.B. Fatoretto), lberton@unifesp.br (L. Berton).

https://doi.org/10.1016/j.asoc.2025.113426
Received 9 October 2024; Received in revised form 19 May 2025; Accepted 2 June 2025
Available online 28 June 2025
1568-4946/© 2025 Published by Elsevier B.V.

context and goals of the application. It also highlights the need for multi-objective optimization techniques, where the model can be fine-tuned to achieve an acceptable balance between competing fairness metrics while maintaining overall performance.

Previous work has employed optimization methods to incorporate one or more fairness functions, with or without constraints [8–10], aiming to enhance model fairness. However, some challenges are the inherent conflict between different fairness metrics, maintaining a high-performing model while also meeting fairness criteria, besides compare different models and methods used to achieve fairness across various metrics.

This work aims to integrate constrained optimization into machine learning models that account for fairness, with the goal of developing more comprehensive and sophisticated approaches to addressing equity constraints. The main contributions of the proposed approach are as follows:

- Incorporating a range of fairness constraints into the optimization process, with the aim of achieving more nuanced and balanced results in terms of equity.
- Exploring metrics associated with the Pareto frontier, which represents the set of optimal solutions balancing trade-offs between accuracy and fairness objectives.
- Investigating indicators that simultaneously consider fairness and model performance, allowing for comparison between different approaches and proposed models.
- Applying the methodology in both simulated and real-world scenarios, with a particular focus on healthcare, where fairness considerations involve complex interactions between different demographic groups.

Based on the above items, we seek to evaluate whether including constraints related to various sensitive attributes present in the datasets results in significant improvements in the model, both in terms of fairness and utility.

The remainder of this article is organized as follows. Section 2 presents other works that explored multi-objective functions for fairness. Section 4 presents the proposed methodology. Section 3 presents the datasets employed in this work, and Section 5 results achieved in the experiments. Section 6 concludes with final remarks.

2. Related work

Various authors have delved into diverse strategies to address the optimization challenge in the pursuit of fairness within machine learning models. Following, we present the most recent approaches.

Gong and Guo [8] introduced the Fair Influence Maximization Problem (FIMP) for fairness in social networks, solved using the FIMMOGA algorithm. Dandl et al. [9] developed a method for counterfactual fairness audits using a genetic algorithm for multi-objective optimization. Wu et al. [10] proposed the Multi-FR framework to balance accuracy and fairness in recommendation systems. Padh et al. [11] used a differentiable approach to approximate fairness, integrated into a multi-objective framework. Liu and Vicente [12] defined accuracy-fairness trade-offs through stochastic multi-objective optimization. Hao et al. [13] addressed biases in collaborative filtering with a constrained Multi-Objective Optimization approach for balanced subgroup performance.

Quantitative comparison of the performance of different algorithms in multi-objective optimization is challenging because the result is often an approximation of the Pareto-optimal set. With the large number of metrics proposed over the past decades, there is no consensus on which to use to evaluate these approximations. A recent study for the quantitative analysis of utility-fairness trade-off systems is proposed in a healthcare setting by Özbulak et al. [14]. Nevertheless, this emerging research direction requires broader attention to enrich evaluation methodologies within multi-objective optimization.

In single-objective minimization problems, the quality of a solution is easily quantifiable. However, for approximations of the Pareto set, the evaluation is more complex. According to Zitzler et al. [15], a good approximation should minimize the distance between the Pareto front and its representation, ensure a good distribution of points in the objective space, and cover a wide range of values for each objective. Many performance indicators have been developed to score these approximations and assist in comparing algorithms. Riquelme et al. [16] and Audet et al. [17] present a review of these proposed indicators, discussing their advantages and disadvantages. They give special attention to the hypervolume indicator, proposed by Zitzler [18], as it has good mathematical properties, can capture dominance and distribution properties, and does not require knowledge of the Pareto front.

Some authors take a distinct approach to simultaneously optimize fairness and accuracy, considering constrained optimization. This has been presented as an efficient heuristic for the problem. The objective is to compel a model to produce fair outputs by integrating fairness constraints into both the learning mechanism and parameter estimations [19].

Zafar et al. [20] proposed methods that incorporate fairness directly into the training process by introducing constraints based on a novel measure of decision boundary fairness. Building on this idea, Cotter et al. [21] framed the problem as a two-player game, decoupling model training from constraint enforcement to improve generalization. Later, Komiyama et al. [22] extended this framework to non-convex settings, showing that exact solutions are attainable even when fairness constraints involve both numerical and categorical sensitive attributes.

Despite these advances, many existing approaches still face important limitations. Techniques that rely on preprocessing or sample reweighting, for example, attempt to mitigate bias by modifying the input data. However, such modifications can introduce distortions that undermine the clinical interpretability of the results. In contrast, our method operates during the training phase, preserving the original data distributions and ensuring that outputs remain interpretable in healthcare contexts.

Another common limitation is the focus on optimizing a single fairness metric – such as demographic parity – without considering the trade-offs with other critical performance indicators like accuracy or sensitivity. Our approach addresses this by simultaneously incorporating multiple fairness criteria, including demographic parity and equal opportunity. This allows us to generate a Pareto front that makes trade-offs between fairness and predictive performance explicit, supporting more transparent and informed decision-making.

Finally, many fairness-driven models in the literature lack validation in real-world clinical settings. In this study, we use the MIMIC-III dataset and focus on clinically relevant conditions, such as diabetic ketoacidosis, to ensure that our findings are not only theoretically sound but also applicable in practice.

Related works show promising results using mathematical optimization methods to achieve fairness in machine learning models. Through optimization algorithms and functions that measure fairness among different sensitive groups, the aim is to find the optimal balance between accuracy and equity. In this study, we compare our proposed approach to that of Zafar et al. [20] and propose incorporating various constraints for fairness metrics into a logistic regression model. The model parameters are estimated through maximum likelihood under constraints on the differences between sensitive groups, limiting these differences to predefined thresholds. In other words, we determine the decision boundary parameters by optimizing the corresponding utility function over the training set under fairness constraints. However, we introduce an approach where we can demonstrate, through performance indicators, that it is possible to compare various algorithms, even in a single-objective optimization scenario, considering different fairness measures.

Table 1
Datasets summary.

Dataset	Instances	Protected attributes
Arrhythmia	452	Gender and Age
MIMIC III — Diabetes	1635	Ethnicity and Age and Gender
MIMIC III — Intracranial hemorrhage	526	Ethnicity and Age and Gender

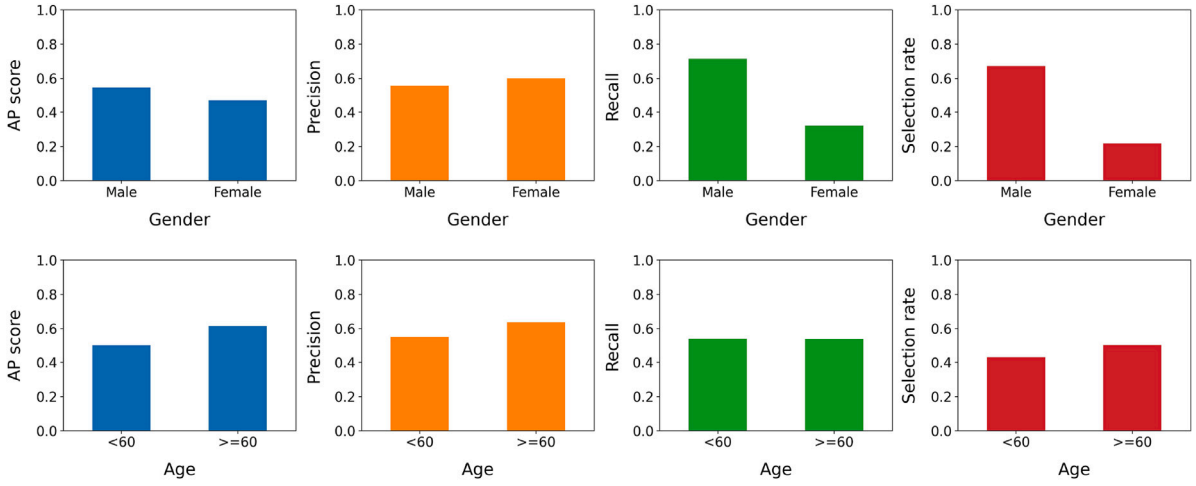


Fig. 1. Metrics for comparing classical logistic regression model results for the Arrhythmia dataset by gender and age.

3. Datasets

This section provides a comprehensive overview of the datasets utilized in our experiments and details the methodologies employed to derive the experimental results.

In our experiments, we utilize two real datasets focusing on healthcare (Table 1). These contexts inherently display discriminatory biases that may persist in machine learning models. These datasets are frequently employed in studies addressing fairness.

Arrhythmia [23]: The Arrhythmia dataset encompasses data from 452 patients, featuring a range of physical measurements and heart rate characteristics essential for diagnosing an arrhythmia. The primary goal of this analysis is to distinguish between the presence and absence of cardiac arrhythmia. In our examination, we observed a demographic parity difference of 0.4543 for gender and 0.0702 for age when applying a classical logistic regression model without accounting for fairness across different groups. As part of our initial analysis of the dataset, we present the measures of AP score, recall, precision, and selection rate, as outlined in the previous sections, for the logistic regression model adjusted for fairness considerations. We then examine and report the results for each group based on the sensitive attribute (Fig. 1).

MIMIC III [24]: The Medical Information Mart for Intensive Care III (MIMIC-III) is a publicly accessible database containing data from patients admitted to intensive care units (ICUs) at Beth Israel Deaconess Medical Center between the years 2001 and 2012. This dataset comprises a wide range of data types, including demographics (such as age, sex, and ethnicity), diagnoses coded with the ICD-9 (International Classification of Diseases, Ninth Revision) system, medication records, laboratory test results, medical interventions, and clinical notes written by healthcare providers.

- **MIMIC III — Diabetes Analysis:** Given the extensive size of the MIMIC-III database, our first analysis focuses on patients with specific diagnoses: Diabetic Ketoacidosis. These conditions were selected because they are directly relevant to our study of diabetes and its complications. We analyze the impact of these diabetes diagnoses in conjunction with other conditions such as Coronary Artery Disease and Congestive Heart Failure, which are often comorbid with diabetes. Additionally, we examine how these diagnoses relate to patient outcomes, including mortality during hospitalization. In assessing fairness within this cohort, we evaluated demographic parity differences using a classical logistic regression model. The analysis revealed a demographic parity difference of 0.0110 for ethnicity (comparing White and Black patients), 0.0156 for age groups where we can observe some more significant differences related to the participants' age (Fig. 2) and 0.0251 for gender. These relatively low disparity values indicate minor differences in model predictions across these sensitive attributes. Consequently, for this analysis, we designated ethnicity and age as the sensitive attributes to further investigate their impact on model fairness.
- **MIMIC III — Intracranial Hemorrhage Analysis:** Our second analysis within the MIMIC-III database focuses on Intracranial Hemorrhage. This condition was chosen due to its smaller sample size, which facilitates a comparative analysis with our previous study. When evaluating demographic parity within this cohort using a classical logistic regression model, we found a difference of 0.0356 for gender and 0.0864 for ethnicity, and 0.2656 for age. Considering these results and the metrics presented in Fig. 3, we do not observe significant differences across the groups for the sensitive variables

Simulation dataset: In addition to real-world datasets, we also considered a simulated dataset. Synthetic data was generated with sample sizes of 300, 800, and 3,000, encompassing 20 features drawn from a normal distribution. Binary labels were created using a logistic function applied to a linear combination of these features. This process involves calculating a linear combination of features and transforming it into probabilities using the logistic function, which is then employed to generate binary outcomes. Sensitive attributes, namely "age" and "gender" were introduced and assigned randomly to each sample. To simulate real-world imbalances, the distributions of the age and gender variables were adjusted according to the class label, resulting in imbalances where certain age groups and genders are more prevalent in specific classes. The distribution of these sensitive attributes within the response variable was examined across five distinct scenarios (Fig. S1).

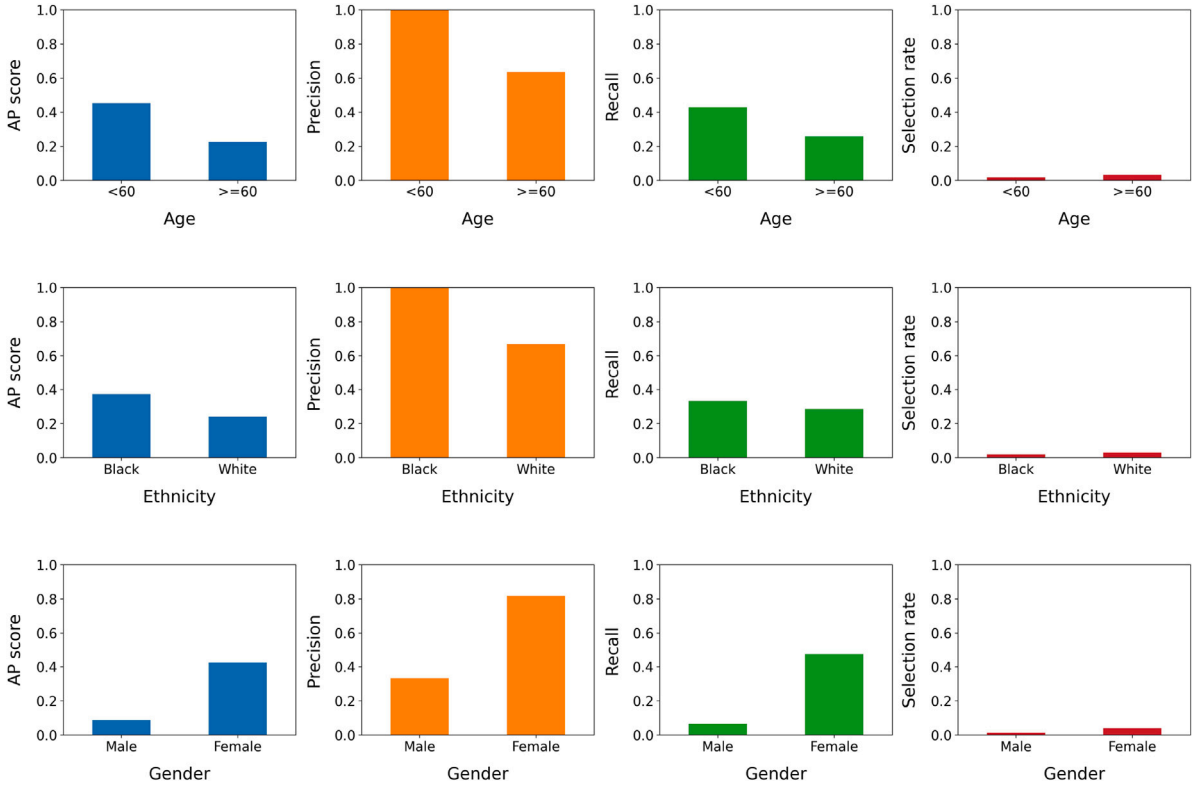


Fig. 2. Metrics for comparing classical logistic regression model results for the MIMIC III — Diabetes dataset by age ethnicity and gender.

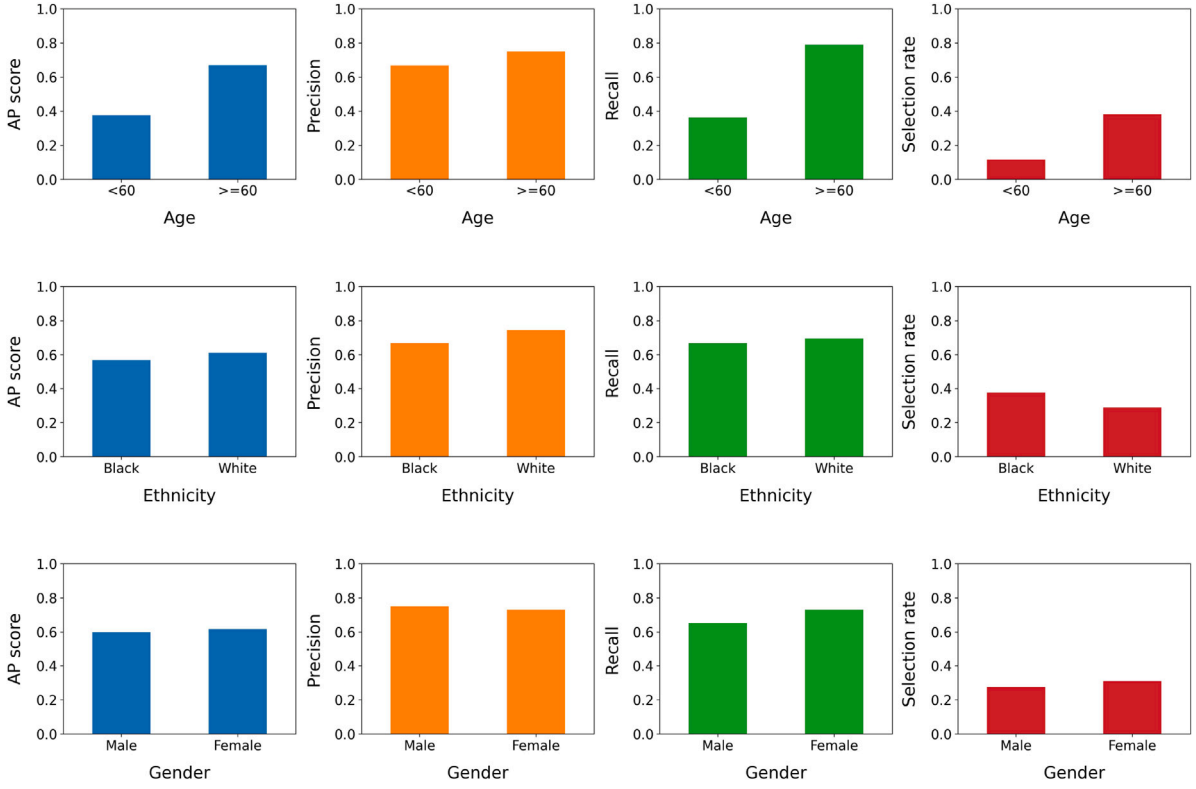


Fig. 3. Metrics for comparing classical logistic regression model results for the MIMIC III — Intracranial hemorrhage dataset by age, ethnicity and gender.

4. Methodology

In this work, we aim to address group fairness. As a principle, Group Fairness involves ensuring that distinct groups or categories of individuals are treated fairly and equitably. In this context, the goal is to prevent discrimination or systematic inequality among groups based on ethnicity, gender, ethnic origin, and religion, among others. Some metrics can be employed to measure fairness in algorithms. The following section introduces commonly used metrics for measuring algorithmic fairness.

4.1. Fairness metrics

Demographic Parity: is a metric of algorithmic fairness in group equity, where the probability of a positive outcome is the same for all demographic groups, regardless of their characteristics [1]. Demographic parity can be expressed by:

$$P(\hat{Y} = 1|Z = a) = P(\hat{Y} = 1|Z = b), \quad \forall a, b \in Z$$

where Z represents a demographic attribute, such as ethnicity or gender. Demographic parity is achieved when the probability of predicting $\hat{Y} = 1$ (a positive outcome) is the same for both groups [25].

A measurement commonly used for assessing disparate impact is known as the CV score [26] or Demographic Parity Difference, the measure of the absolute difference in positive probability between the groups, is defined as follows:

$$DPD(\hat{Y}) = |P(\hat{Y} = 1|Z = a) - P(\hat{Y} = 1|Z = b)| \tag{1}$$

For margin-based classifiers, these metrics can be approximated by the empirical covariance between the sensitive attributes Z and the best hyperplane, making use of the decision boundary covariance that achieves the greatest classification accuracy [20]:

$$\text{Cov}(\boldsymbol{z}, d_\theta(\boldsymbol{x})) \approx \frac{1}{N}\sum_{i=1}^{N}(\boldsymbol{z_i} - \bar{\boldsymbol{z}})d_\theta(\boldsymbol{x_i}) \tag{2}$$

where $d_\theta(\boldsymbol{x})$ is the signed distance from the users' feature vectors to the decision boundary. According to Zafar et al. [20], (Eq. (2)) is a convex function concerning the decision boundary parameters $\boldsymbol{\theta}$, as $_\theta(\boldsymbol{x_i})$ is convex with respect to θ for all linear and convex margin-based classifiers. Therefore, it can be easily incorporated into the formulation of these classifiers without increasing the complexity of training.

The metrics detailed above can be classified under the criterion of independence, where to achieve independence, we require that the scores produced by the model are independent of the protected attribute Z.

In addition to fairness metrics, we also calculated model fitted metrics for each group defined by the sensitive attribute. In this work, we initially fitted a classic logistic regression model, using 30% of the data for training. After fitting the model, we calculated several metrics to compare the observed outcomes across different categories using the `fairlearn` library. The obtained metrics include:

- **AP Score (Average Precision)**:

  $$\text{AP Score} = \sum_n (R_n - R_{n-1})P_n$$

  where R_n and P_n are the recall and precision at point n on the precision–recall curve, respectively. The AP Score approximates the area under the precision–recall curve, indicating the model's ability to handle imbalanced datasets by focusing on positive class prediction.
- **Precision**:

  $$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}$$

  where TP are True Positives and FP are False Positives. Precision measures the proportion of correctly predicted positive observations among all predicted positives, showing how many of the predicted positives were actually correct.
- **Recall**:

  $$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

  where TP are True Positives and FN are False Negatives. Recall measures the proportion of correctly predicted positive observations among all actual positives, reflecting the model's ability to identify true positive cases.
- **Selection Rate**:

  $$\text{Selection Rate} = P(\hat{Y} = 1 \mid Z)$$

  where $\hat{Y}$ is the predicted outcome and the Selection Rate represents the probability that an individual from a given group is predicted as positive (i.e., $\hat{y} = 1$). This metric provides insight into how frequently each category is selected by the model.

4.2. Maximizing under fairness constraints

In this section, we develop classifiers that optimize accuracy while adhering to fairness constraints. To achieve this, we determine the decision boundary parameters by minimizing the associated loss function of logistic regression. In logistic regression considering feature vectors $\mathbf{x_i}$ to the class labels y_i by means of a probability distribution:

$$p(y_i = 1|\mathbf{x_i}, \boldsymbol{\theta}) = \frac{1}{1 + \exp(-\boldsymbol{\theta}^T \mathbf{x_i})}$$

where $\boldsymbol{\theta}$ is estimated solving the $\boldsymbol{\theta}^* = \text{argmin}_\theta - \sum_{i=1}^{N} \log p(y_i|\mathbf{x_i}, \theta)$, being the corresponding loss function $-\sum_{i=1}^{N} \log p(y_i|\boldsymbol{x_i}, \boldsymbol{\theta})$.

To maximize Accuracy Under Fairness Constraints, Zafar et al. [20], Vincent [27] proposed to optimize the decision boundary parameters by minimizing the loss function over the training set with fairness constraints

$$\begin{aligned} \min \quad & -\sum_{i=1}^{N} \log p(y_i|\boldsymbol{x_i}, \boldsymbol{\theta}) \quad -\sum_{i=1}^{N} \log p(y_i|\boldsymbol{x_i}, \boldsymbol{\theta}) \\ \text{s.t.} \quad & \frac{1}{N}\sum_{i=1}^{N}(\boldsymbol{z_i} - \bar{\boldsymbol{z}})d_\theta(\boldsymbol{x_i}) \leq \mathbf{c} \\ & \frac{1}{N}\sum_{i=1}^{N}(\boldsymbol{z_i} - \bar{\boldsymbol{z}})d_\theta(\boldsymbol{x_i}) \geq -\mathbf{c} \end{aligned} \tag{3}$$

We propose to estimate the parameters of the logistic model while taking into account two or more sensible attributes. To achieve this, we employed the *scikit-lego* package [28]. This is a library in Python that extends the capabilities of scikit-learn, particularly focusing on adding functionalities for building and assessing composite models. Regarding fairness, *scikit-lego* provides modules related to fairness in machine learning. For example, it includes constraint optimization for creating and evaluating models with an emphasis on fairness considerations. The clarity and extensibility of the functions created in the package are positive aspects. This allowed us to introduce multiple fairness constraints in the optimization and estimation of the parameters of the models under study. In the course of our investigations, a uniform threshold was applied to both constraints; nevertheless, this parameter can be readily customized to align with the researcher's specific preferences. Furthermore, it is noteworthy that alternative loss functions for diverse classification models, including but not limited to Support Vector Machines, may similarly undergo optimization while being subject to

fairness constraints.

$$
\begin{aligned}
\min \quad & -\sum_{i=1}^{N} \log p(y_i|\boldsymbol{x_i},\boldsymbol{\theta}) \quad -\sum_{i=1}^{N} \log p(y_i|\boldsymbol{x_i},\boldsymbol{\theta}) \\
\text{s.t.} \quad & \frac{1}{N}\sum_{i=1}^{N}(\boldsymbol{z_{1i}}-\bar{\boldsymbol{z}}_{\boldsymbol{1}})d_\theta(\boldsymbol{x_i}) \le \mathbf{c} \\
& \frac{1}{N}\sum_{i=1}^{N}(\boldsymbol{z_{1i}}-\bar{\boldsymbol{z}}_{\boldsymbol{1}})d_\theta(\boldsymbol{x_i}) \ge -\mathbf{c} \\
& \frac{1}{N}\sum_{i=1}^{N}(\boldsymbol{z_{2i}}-\bar{\boldsymbol{z}}_{\boldsymbol{2}})d_\theta(\boldsymbol{x_i}) \le \mathbf{c} \\
& \frac{1}{N}\sum_{i=1}^{N}(\boldsymbol{z_{2i}}-\bar{\boldsymbol{z}}_{\boldsymbol{2}})d_\theta(\boldsymbol{x_i}) \ge -\mathbf{c} \\
& \vdots \qquad\qquad\qquad\qquad \vdots \\
& \frac{1}{N}\sum_{i=1}^{N}(\boldsymbol{z_{ji}}-\bar{\boldsymbol{z}}_{\boldsymbol{j}})d_\theta(\boldsymbol{x_i}) \le \mathbf{c} \\
& \frac{1}{N}\sum_{i=1}^{N}(\boldsymbol{z_{ji}}-\bar{\boldsymbol{z}}_{\boldsymbol{j}})d_\theta(\boldsymbol{x_i}) \ge -\mathbf{c}
\end{aligned}
\tag{4}
$$

where j represents the number of sensitive attributes included as constraints. For each sensitive attribute j, two constraints are imposed: one ensuring that the weighted deviation from the mean, calculated with $d_\theta(\boldsymbol{x_i})$, does not exceed an upper bound $\mathbf{c}$, and another ensuring it does not fall below a lower bound $-\mathbf{c}$. The number of constraints grows proportionally with the number of sensitive attributes, such as age, gender, or ethnicity.

4.3. Performance indicators in optimization

To facilitate a comprehensive comparison of these different constraints previously presented, this paper introduces the use of metrics commonly employed in multi-objective optimization (MOO) and adapted in the context of optimization with constraints. By applying these metrics, we aim to evaluate the performance of models under various fairness constraints, providing deeper insights into their efficacy and fairness.

In MOO, the objective functions can be denoted as $f_1(\boldsymbol{x}), f_2(\boldsymbol{x}), \ldots, f_K(\boldsymbol{x})$, where K is the number of objective functions in the optimization problem and $\mathbf{x} = [x_1, x_2, \ldots x_n]^T$ is the n decision variables that are the numerical quantities for which values are to be chosen in an optimization problem [29].

The notion of Pareto Dominance is crucial in this context. Considering two vectors $\mathbf{x}, \mathbf{x}' \in \Omega$ (an n-dimensional decision space), the vector $\mathbf{x}$ is said to dominate $\mathbf{x}'$, denoted as $\mathbf{x} \succ \mathbf{x}'$, if $\mathbf{x}$ is not worse than $\mathbf{x}'$ in any objective function and is strictly better in at least one objective function [30]. In this way, the Pareto optimal set (P^*) encompasses all solutions that are non-dominated with respect to Ω, defined as:

$$
P^* := \mathbf{x} \in \Omega | \neg \exists \mathbf{x}' \in \Omega \quad \text{such that} \quad \mathbf{x}' \succ \mathbf{x} \tag{5}
$$

Pareto optimal solutions are known as the Pareto set, and the visual representation of this set is termed the Pareto front [17,31].

In real-world applications, calculating the exact Pareto optimal front is often infeasible. Instead, a reference set is employed as an approximation. Generally, a well-approximated Pareto front suffices for practical use. Let $A \subseteq Z$ be a collection of objective vectors, with $\mathbf{z} = (z_1, z_2, \ldots z_n) \in Z$ and $\mathbf{z} = \mathbf{f}(\mathbf{x}) = (f_1(\mathbf{x}), f_2(\mathbf{x}), \ldots, f_m(\mathbf{x}))$ [32]. An approximation set A is defined such that no element in A dominates or is equal to any other objective vector within A. This definition aims to exclude solutions dominated by the optimization algorithm. Importantly, this definition lacks an explicit quality metric; the intention is for the algorithm to produce a high-quality approximation set for the Pareto optimal set while recognizing that completely achieving this is often impractical [32].

In the design of metrics for MOO, four primary performance criteria are typically considered to analyze the Pareto Front optimality: capacity (or cardinality), convergence, spread, and distribution [17, 31,33]. The cardinality of set A denotes the quantity of solutions it contains. A greater number of solutions is preferred to characterize more extensively the trade-offs on the Pareto front. On the other hand, convergence measures how far A deviates from the theoretical Pareto optimal front. Distribution relates to the relative spacing among solutions within A, while spread pertains to the extent of the value range covered by the solutions in A.

4.3.1. The hyperarea/hypervolume (HV)

The most widely used performance metric according to [17,31] is the hypervolume (HV) in three dimensions or hyperarea (HA) in two dimensions. This unary metric measures the size of the objective space covered by an approximation set, relying on a reference point for calculation. Hypervolumes take into account distribution, spreading, and convergence at the same time, making these unique in this regard. Recognized for their distinctive properties, hypervolumes are Pareto-compliant, ensuring that any approximation set achieving maximum quality for an MOO contains all Pareto optimal solutions. The reference point can be simply attained by constructing a vector of worst objective function values.

To compute the HV, we initially assess solutions while balancing utility (AP) and fairness independence metric (Demographic Parity Difference). These diverse solutions are derived by varying covariance thresholds to encompass all conceivable scenarios.

Within the framework of Multi-Objective Optimization (MOO), our objectives include maximizing AP while minimizing disparities in Demographic Parity. The Pareto Front curve thus delineates the optimal configurations within this trade-off space. Consequently, we establish the worst-case scenario (0 for AP and 1 for fairness metrics) as a baseline and calculate the HV for the corresponding region. The HV enables a comparative evaluation of different optimization models, pinpointing those that exhibit superior performance.

The functions compute the hypervolume (HV) for a given set of Pareto front points (covariance thresholds) in any dimensional space, using a specified reference point or worst scenario (0,1,1, ...,1) to delineate the region of interest. The Pareto front points are sorted, and prisms (or hyperprisms) are computed between each point and the reference point. The cumulative volume of these prisms determines the hypervolume. This approach applies to 2D, 3D, or higher-dimensional cases, allowing for the visualization and comparative analysis of dominated and non-dominated points in a plot. This facilitates the comparison of multi-objective optimization models based on their effectiveness in balancing objectives across different dimensional spaces. Below, we present an example of how this calculation would be performed for the 3D cases (Fig. 4).

The HV could be expressed by

$$
HV(PF) = VOL\left(\bigcup_{x \in PF} [x, r]\right) \tag{6}
$$

where x is the non-dominated points and r is the reference point. Each pair of $[x, r]$ defines a prism-shaped hypervolume in metric space and the union of all these hypervolume constructs the total hypervolume for the PF.

In our research, we propose a comparison of models derived from equations (Eq. (3) and (4)) to investigate how this metric responds when multiple sensible attributes are integrated.

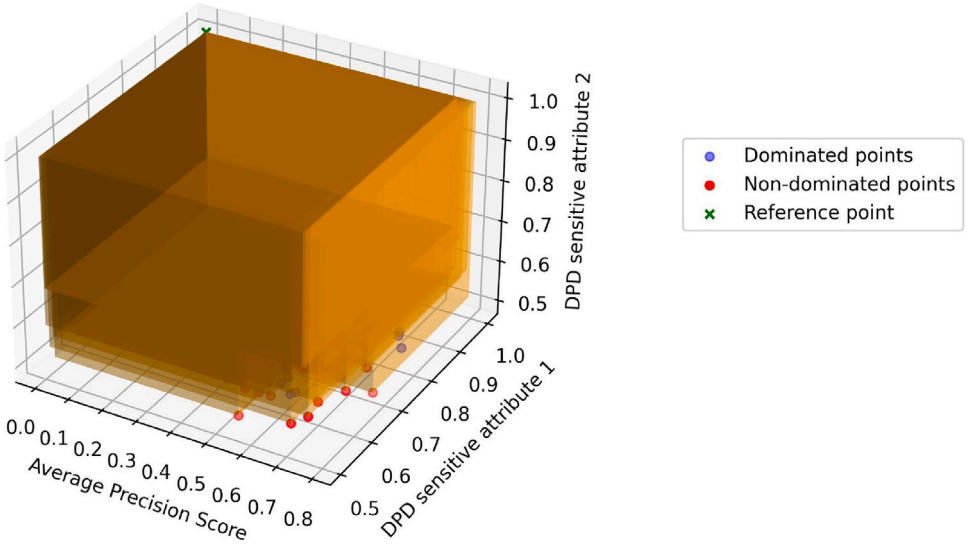


Fig. 4. Plot of the example of how hypervolume is calculated in 3D dimensions.

4.3.2. The Spacing indicator (SP)

The Spacing Indicator (SP), introduced by Schott [34], serves as an important tool for assessing the distribution of solutions along the Pareto front. This metric quantifies the variation in the distance between each solution and its nearest neighboring solution. Essentially, it computes the standard deviation of these distances, thereby offering a measure of the uniformity of solution distribution.

The foremost objective of the SP is to gauge how evenly distributed the solutions are along the Pareto front. A more uniform distribution signifies a higher quality front, as it implies that the solutions are consistently spaced. A lower SP value is desirable, as it indicates a more uniform distribution of solutions, meaning they are equidistantly dispersed along the Pareto front.

The calculation of the SP involves determining the distance between each solution and its closest neighboring solution. Subsequently, the standard deviation of these distances is calculated. The SP metric represents the variation in these distances, providing insight into the uniformity of the solution distribution.

$$SP = \sqrt{\frac{1}{NPF - 1} \sum_{i=1}^{NPF} (d_i - \bar{d})^2}, \tag{7}$$

where d_i is defined by

$$d_i = \min\left(\sum_{m=1}^{M} |f_m^i(x) - f_m^j(x)|\right) \tag{8}$$

and where NPF represents the number of solutions in the solution set RR, belonging to the Pareto front obtained; $\bar{d}$ is the average of all values of d_i; M is the number of objective functions, i and j are the solutions in the Pareto front.

The SP is a vital measure of the uniformity of the solution distribution on a Pareto front. Lower SP values are preferable, as they denote a more uniform distribution of solutions, reflecting a higher quality Pareto front (Fig. 5). An SP value of zero indicates that all members of the solution set are spaced equidistantly based on Manhattan distance. However, it is important to note that SP only assesses the distribution in the immediate "neighborhood" of solutions. Even when used alongside MS (Mean Squared Distance), SP alone cannot fully capture the diversity quality of the solution set.

4.3.3. Non-dominated proportion (NDP)

The non-dominated ratio is a metric used in multi-objective optimization to quantify the quality of a set of solutions relative to a reference set. It provides insight into how many of the solutions are in the Pareto front (non-dominated solutions) compared to the total number of solutions available (non-dominated plus dominated) and the "efficacy" of a particular method in generating non-dominated solutions.

This can be expressed as:

$$NDP = \frac{NPF}{NS} \tag{9}$$

where NPF and NS means the number of non-dominated solutions and the number of solutions, respectively.

This ratio gives an indication of the quality and diversity of the solutions set. A higher ratio suggests a larger proportion of non-dominated solutions.

4.4. Methods and algorithms developed

To begin the performance comparison between the model with a single constraint, which limits the Demographic Parity Difference (DPD) to a single sensitive attribute (model Eq. (3)), and the model that incorporates multiple constraints for various sensitive attributes (model Eq. (4)), we use the `DemographicParityClassifier` class from the `sklego` library and a newly proposed class that combines more than one constraint. These classes fit a logistic regression model and inherit methods from the `FairClassifier` class, which uses the `_solve` method to optimize the model coefficients. The objective of this optimization process is to maximize the likelihood or, equivalently, minimize the loss function, while taking into account fairness constraints related to the sensitive variables.

We employ `GridSearchCV`, a tool from `scikit-learn`, to determine the optimal parameters for the logistic models under investigation. In this procedure, various `covariance_threshold` values, ranging from 0.01 to 1.00, are considered. Logistic models are fitted using these thresholds, and the estimated parameters are subsequently used to make predictions and evaluate the performance metrics under study.

During this process:

- **Training and Evaluation:**
  - `GridSearchCV` trains the model for each combination of the `covariance_threshold` parameter values specified in the grid.
  - For each combination, the model is evaluated using cross-validation with 5 splits (20% of the data for testing in each split).

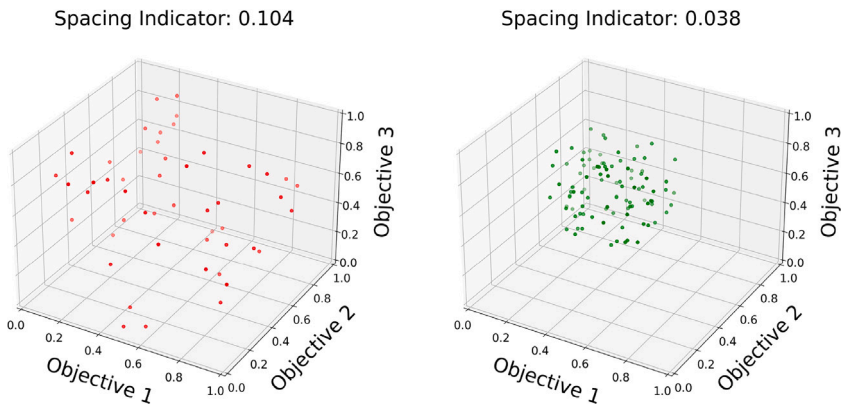


Fig. 5. Comparison of two Pareto fronts according to SP metric.

- Average Precision Score (AP) is calculated for each model trained with each parameter combination, and this metric is used to select the best model.
- Besides the AP, we also obtained for each `covariance_threshold`, the absolute difference of Demographic Parity (Eq. (1)) for each sensitive attribute. These values were used to calculate the metrics presented in Section 4.3.

• **Calculating Log Loss:**

- The log loss is calculated for each `covariance_threshold` parameter configuration. This calculation involves:
  - Fitting the model with each `covariance_threshold` value.
  - Obtaining the predicted probabilities from the fitted model.
  - Calculating the log loss using these predicted probabilities and the true labels.

Log loss assesses the quality of the predicted probabilities generated by the model. It measures how well the model's predictions match the true labels, penalizing incorrect predictions made with high confidence more severely.

4.4.1. Experimental procedure

We start our study by assessing the quality of the proposed model (Eq. (4)) through a comparison of the logistic regression loss function across various simulation scenarios (Fig. S1), with varying sample sizes, to assess whether the log loss function stabilizes as the covariance threshold increases, thereby allowing for a more flexible model about the sensitive attribute.

In the subsequent analysis, we compared models incorporating the demographic parity constraint on a single sensitive attribute, as proposed by Zafar (Eq. (3)), with our proposed model (Eq. (4)), which constraints for two or more sensitive attributes. The analysis was conducted by plotting the AP score against the difference in demographic parity between the groups of the sensitive attributes, with each point representing a specific threshold as demonstrated in Fig. 4. From these plots, the Pareto front was constructed, and the metrics were evaluated according to the methodology described in Section 4, using the worst-case scenario [0,1,..,1] as the reference point.

To gain deeper insights into the behavior of the proposed metrics for model comparison, we analyzed the differences between models that account for a single sensitive attribute and those that incorporate two or three sensitive attributes, using the MIMIC dataset. This analysis was performed by fitting the model across a range of threshold values, from 0.01 to 1.00, applied to the constraint function. For models incorporating only one sensitive attribute, the attribute was selected based on the one exhibiting the largest disparity in demographic parity, as outlined in Section 3. This attribute was added first, followed by others in succession. The goal was to assess the impact on a model that already constraint based on the sensitive attribute with the largest group difference, and then observe the effect of adding constraints for attributes with smaller demographic disparities.

After conducting the experiments and calculating the differences in the obtained metrics (HV, NPD, and SP) for the various models with different numbers of thresholds used in the constraint, the Wilcoxon test [35] was applied to determine whether these differences were statistically significant. The Wilcoxon test was applied to check whether the differences in metrics between the models were significantly different from zero, indicating whether or not the variations in performance are statistically meaningful.

5. Results

In this section, we present the results quantitatively and objectively, based on the performance indicators defined in the methodology (Section 4.3). The evaluation considered metrics such as the Hypervolume difference (HV), the proportion of non-dominated solutions (NDP), and the Spacing Indicator (SP). To assess the statistical significance of differences between models, the Wilcoxon signed-rank test was applied.

For the Arrhythmia dataset, the comparison between models with one and two fairness constraints revealed a statistically significant difference in HV, with a mean increase of up to 0.1056 ($p = 0.001$), indicating a substantial expansion of the covered region in the solution space when a second constraint was added (Table 2). Similarly, NDP showed significant improvements ($p = 0.001$), with a mean increase of up to 0.0600, reflecting a higher proportion of non-dominated solutions. In contrast, SP did not exhibit a significant change ($p = 0.9658$), suggesting that the additional constraint did not enhance the uniformity of the Pareto front distribution.

In the case of MIMIC-III — Diabetes, all three indicators showed statistically significant differences between models with one and two constraints. HV increased by up to 0.0266 ($p = 0.001$), NDP decreased by up to −0.0100 ($p = 0.0001$), and SP varied by up to 0.0197 ($p = 0.001^{**}$). Despite the small absolute differences, their consistency across comparisons and low p-values confirm statistical relevance.

For MIMIC-III — Intracranial Hemorrhage, the HV difference reached 0.0304 ($p = 0.01$), while NDP decreased by up to −0.0131 ($p = 0.02^{*}$) and SP increased by up to 0.0526 ($p = 0.000$). These findings indicate that even in scenarios with fewer samples ($n = 526$), the model with two sensitive attributes outperformed the simpler version, although the improvement in SP was less pronounced compared to HV and NDP.

When considering three sensitive attributes (gender, age, and ethnicity) in the MIMIC-III — Diabetes dataset, the HV difference increased

Table 2
Comparison of the differences observed between models 3 and model 4 with 2 sensitive attributes.

Dataset	Const. model 3	Const. model 4	N. of thr.	HV Dif.	NDP Dif.	SP Dif.
Arrhythmia	Gender	Gender and Age	20	0.0499	0.0600	0.0078
Arrhythmia	Gender	Gender and Age	50	0.0797	0.0400	0.0002
Arrhythmia	Gender	Gender and Age	100	0.0827	0.0340	0.0009
Arrhythmia	Gender	Gender and Age	150	0.0838	0.0253	−0.0067
Arrhythmia	Gender	Gender and Age	200	0.0841	0.0190	0.0045
Arrhythmia	Gender	Gender and Age	250	0.0844	0.0220	−0.0013
Arrhythmia	Gender	Gender and Age	300	0.0887	0.0153	−0.0056
Arrhythmia	Gender	Gender and Age	350	0.0843	0.0137	−0.0028
Arrhythmia	Gender	Gender and Age	400	0.0839	0.0105	0.0020
Arrhythmia	Gender	Gender and Age	450	0.1019	0.0120	−0.0009
Arrhythmia	Gender	Gender and Age	1000	0.1056	0.0058	0.0042
		Wilcoxon test (p-value)		0.0010**	0.0010**	0.9658
MIMIC III — Diabetes	Gender	Gender and Age	20	0.0266	−0.0100	0.0197
MIMIC III — Diabetes	Gender	Gender and Age	50	0.0266	−0.0040	0.0197
MIMIC III — Diabetes	Gender	Gender and Age	100	0.0266	−0.0020	0.0197
MIMIC III — Diabetes	Gender	Gender and Age	150	0.0266	−0.0013	0.0197
MIMIC III — Diabetes	Gender	Gender and Age	200	0.0266	−0.0010	0.0197
MIMIC III — Diabetes	Gender	Gender and Age	250	0.0218	−0.3944	0.0009
MIMIC III — Diabetes	Gender	Gender and Age	300	0.0266	−0.0007	0.0197
MIMIC III — Diabetes	Gender	Gender and Age	350	0.0266	−0.0006	0.0197
MIMIC III — Diabetes	Gender	Gender and Age	400	0.0266	−0.0005	0.0197
MIMIC III – Diabetes	Gender	Gender and Age	450	0.0266	−0.0004	0.0197
MIMIC III — Diabetes	Gender	Gender and Age	1000	0.0266	−0.0002	0.0197
		Wilcoxon test (p-value)		0.001**	0.0001**	0.001**
MIMIC III — Int. hem.	Age	Age and Ethnicity	20	0.0154	−0.0118	0.0526
MIMIC III — Int. hem.	Age	Age and Ethnicity	50	0.0304	−0.0131	0.0530
MIMIC III — Int. hem.	Age	Age and Ethnicity	100	0.0043	−0.0021	0.0482
MIMIC III — Int. hem.	Age	Age and Ethnicity	150	0.0114	−0.0013	0.0448
MIMIC III — Int. hem.	Age	Age and Ethnicity	200	0.0023	−0.0031	0.0466
MIMIC III — Int. hem.	Age	Age and Ethnicity	250	0.0109	−0.0006	0.0387
MIMIC III — Int. hem.	Age	Age and Ethnicity	300	0.0023	−0.0020	0.0466
MIMIC III — Int. hem.	Age	Age and Ethnicity	350	0.0023	−0.0020	0.0466
MIMIC III — Int. hem.	Age	Age and Ethnicity	400	0.0023	−0.0015	0.0466
MIMIC III — Int. hem.	Age	Age and Ethnicity	450	0.0032	−0.0013	0.0469
MIMIC III — Int. hem.	Age	Age and Ethnicity	1000	0.0032	−0.0008	0.0470
		Wilcoxon test (p-value)		0.01**	0.02*	0.000**

The Wilcoxon signed-rank test is used to determine whether the median of the differences between observations is significantly different from zero. In this case, it is applied to test whether the median difference between the values in each column is equal to zero or not. A p-value less than 0.05 (*) or 0.01 (**) indicates that there is significant evidence to reject the null hypothesis that the median difference is zero.

to 0.0489 ($p = 0.001$), and NDP reached 0.0200. SP showed a negative variation of −0.0207 ($p = 0.001$), suggesting a more uniform distribution of solutions (Table 3).

For MIMIC-III — Intracranial Hemorrhage, the model with three constraints also achieved statistically significant improvements across all indicators: $HV = 0.0239$, $NDP = 0.0368$, and $SP = 0.0241$ ($p = 0.002^{*}$), reinforcing the robustness of the proposed approach even in smaller datasets (Table 3).

Table 4 summarizes the best observed performance for each model, considering the optimal threshold. For instance, in MIMIC-III — Diabetes, the model with three constraints attained the highest HV (0.3169) and NDP (0.0280). In the Arrhythmia dataset, the best performance was obtained using 1000 thresholds, with $HV = 0.5995$, $NDP = 0.0094$, and $SP = 0.0159$, confirming that increasing the number of fairness constraints can substantially enhance the trade-off between utility and fairness.

6. Discussion

In all scenarios, we consistently observed a decrease in log loss as the sample size increased, indicating the robust performance of the proposed model (Fig. 6). In Scenario 1, where groups are balanced concerning the threshold for both sensitive variables (Fig. S1), the logistic loss function stabilizes rapidly. Beyond a threshold of 0.2, the loss remains steady, regardless of sample size. Scenarios 2 and 4 demonstrate similar trends, with classes being imbalanced for at least one sensitive variable. Nevertheless, this imbalance is uniformly distributed across both classes of the response variable. In these scenarios, we also observe a swift stabilization of the logistic loss function, though with more pronounced variation depending on sample size. Conversely, in Scenarios 3 and 5, where there is an imbalance in at least one sensitive variable, but this imbalance is reversed across categories of the response variable, the model requires a more extended period to achieve stabilization. This phenomenon introduces additional complexity to the modeling process. These results suggest that tailoring models to different sample sizes and balancing scenarios effectively optimizes logistic regression loss. This approach provides a deeper understanding of the interplay between fairness metrics and model performance, illustrating how variations in fairness parameters influence maximum likelihood and classification quality. It underscores the importance of considering both sample size and data balance in the development of fair and efficient classification models.

Analyzing the Arrhythmia dataset, we observed significant differences in hypervolumes and NDP between the two models (Table 2). Despite the relatively small sample size of 452 instances, the addition of a second constraint had a notable effect on the hypervolume calculation. This impact was possibly caused by the substantial disparity in demographic parity between the groups, as well as differences in recall and selection rates by gender (Fig. 1). Interestingly, despite the limited sample size, the hypervolume metric demonstrated remarkable stability, showing a consistent increase as more thresholds were incorporated into the Pareto front analysis. In contrast, no significant difference was found in the SP metric, which measures the uniformity across the observed Pareto fronts.

For the MIMIC III Diabetes dataset, we also found a significant difference in HV when comparing the two models, although this difference was smaller than that observed in the previous dataset (Table 2). This dataset, with a larger sample size (1635 instances), exhibited consistent values across different thresholds, resulting in a statistically significant difference from zero, and remained stable starting from a threshold number of 20. Although the demographic parity difference between the groups was not as pronounced, we observed significant

Table 3
Comparison of the differences observed between models (3) and (4) with three sensitive attributes.

Dataset	Const. model 3	Const. model 4	N. of thr.	HV Dif.	NDP Dif.	SP Dif.
MIMIC III — Diabetes	Gender	Age, Ethnicity and Gender	20	0.0483	0.0200	−0.0131
MIMIC III — Diabetes	Gender	Age, Ethnicity and Gender	50	0.0489	0.0080	−0.0162
MIMIC III — Diabetes	Gender	Age, Ethnicity and Gender	100	0.0486	0.0060	−0.0169
MIMIC III — Diabetes	Gender	Age, Ethnicity and Gender	150	0.0487	0.0053	−0.0174
MIMIC III — Diabetes	Gender	Age, Ethnicity and Gender	200	0.0488	0.0040	−0.0207
MIMIC III — Diabetes	Gender	Age, Ethnicity and Gender	250	0.0491	0.0024	−0.0204
MIMIC III — Diabetes	Gender	Age, Ethnicity and Gender	300	0.0488	0.0027	−0.0207
MIMIC III — Diabetes	Gender	Age, Ethnicity and Gender	350	0.0488	0.0023	−0.0207
MIMIC III — Diabetes	Gender	Age, Ethnicity and Gender	400	0.0488	0.0020	−0.0207
MIMIC III — Diabetes	Gender	Age, Ethnicity and Gender	450	0.0488	0.0018	−0.0207
MIMIC III — Diabetes	Gender	Age, Ethnicity and Gender	1000	0.0488	0.0008	−0.0207
		Wilcoxon test (p-value)		0.0010**	0.0010**	0.0010**
MIMIC III — Int. hemor.	Age	Age, Ethnicity and Gender	20	0.0121	0.0368	0.0002
MIMIC III — Int. hemor.	Age	Age, Ethnicity and Gender	50	0.0081	0.0306	0.0170
MIMIC III — Int. hemor.	Age	Age, Ethnicity and Gender	100	0.0103	0.0362	0.0063
MIMIC III — Int. hemor.	Age	Age, Ethnicity and Gender	150	0.0215	0.0245	0.0241
MIMIC III — Int. hemor.	Age	Age, Ethnicity and Gender	200	0.0226	0.0166	0.0201
MIMIC III — Int. hemor.	Age	Age, Ethnicity and Gender	250	0.0226	0.0137	0.0197
MIMIC III — Int. hemor.	Age	Age, Ethnicity and Gender	300	0.0234	0.0132	0.0190
MIMIC III — Int. hemor.	Age	Age, Ethnicity and Gender	350	0.0217	0.0093	0.0197
MIMIC III — Int. hemor.	Age	Age, Ethnicity and Gender	400	0.0239	0.0074	0.0193
MIMIC III — Int. hemor.	Age	Age, Ethnicity and Gender	450	0.0227	0.0063	0.0197
MIMIC III — Int. hemor.	Age	Age, Ethnicity and Gender	1000	0.0220	0.0034	0.0179
		Wilcoxon test (p-value)		0.002*	0.002*	0.002*

The Wilcoxon signed-rank test is used to determine whether the median of the differences between observations is significantly different from zero. In this case, it is applied to test whether the median difference between the values in each column is equal to zero or not. A p-value less than 0.05 (*) or 0.01 (**) indicates that there is significant evidence to reject the null hypothesis that the median difference is zero.

Table 4
Comparison of models with constraints on different sensitive attributes, considering the proposed metrics, with the best number of thresholds observed within the range of 0.01 to 1.

Dataset	Sensitive attributes	Threshold	HV	1 - HV	NDP	SP
MIMIC III — Diabetes	Age	50	0.224	0.776	0.0080	0.0000
MIMIC III — Diabetes	Ethnicity	50	0.257	0.743	0.0120	0.0169
MIMIC III — Diabetes	Gender	50	0.268	0.732	0.0200	0.0262
MIMIC III — Diabetes	Gender and Ethnicity and Age	50	0.3169	0.6831	0.0280	0.0100
MIMIC III — Int. hemor.	Ethnicity	400	0.4921	0.5079	0.0091	0.0626
MIMIC III — Int. hemor.	Gender	400	0.5343	0.4657	0.0071	0.0122
MIMIC III — Int. hemor.	Age	400	0.6069	0.3931	0.0155	0.0077
MIMIC III — Int. hemor.	Age and Ethnicity and Gender	400	0.6308	0.3692	0.0229	0.0270
Arrhythmia	Age	1000	0.3621	0.6379	0.0006	0.0842
Arrhythmia	Gender	1000	0.4939	0.5061	0.0036	0.0117
Arrhythmia	Gender and Age	1000	0.5995	0.4005	0.0094	0.0159

differences in precision and recall when comparing age groups, which likely contributed to the significant results (Fig. 2). Additionally, the NDP and SP also showed significant differences. The NDP exhibited extremely low, though significant, values that were negative, which was unexpected when the model was improved. The SP metric, on the other hand, did not demonstrate improved uniformity for the more comprehensive model. When applying three constraints (one for each sensitive attribute) the model exhibited enhanced performance compared to the version with a single constraint (Table 3). Significant differences were observed across all evaluated metrics. As anticipated, the hypervolume (HV) difference became more pronounced with the addition of multiple constraints, as opposed to models with only one constraint. The non-dominated points (NDP) indicator behaved as expected, with positive values indicating an increased number of points on the Pareto front for the more comprehensive model. Additionally, the spacing indicator (SP) showed improvement, with lower values signifying a more uniform distribution of points on the Pareto front in the more constrained model.

For the intracranial hemorrhage dataset in MIMIC III, a significant difference in HV was also observed when comparing models with a single constraint versus two constraints (Table 2). However, this difference stabilized more slowly, with lower values compared to other datasets, even after using 300 thresholds. This is likely due to the smaller sample size of this dataset (526 instances) and the minimal disparity in demographic parity between the two ethnic groups. Although the inclusion of the ethnicity constraint as a sensitive attribute was significantly different from zero, its impact was limited due to the low observed variation, possibly because the age group differences had already accounted for much of this variation (Fig. 3). Regarding the NDP and SP indicators, similar behavior was observed as in the diabetes dataset. Additionally, a further constraint was added to the model using gender as a sensitive attribute. Once again, we observed a greater difference in the HV metric compared to the model with only a single constraint. However, in this dataset, the SP indicator did not perform as well as it did in the diabetes dataset, suggesting that this metric yielded better results in larger datasets and when comparing models with more pronounced differences (Table 4).

To complete, Table 4 presents the results of the fitted models with the optimal thresholds, revealing the greatest differences in Hypervolume (HV). As anticipated, HV generally favors models that incorporate constraints on both sensitive attributes. For the Diabetes dataset, a threshold of 50 points shows substantial differences between models with varying numbers of constraints. Conversely, for the Intracranial Hemorrhage dataset, a threshold of 400 reveals more pronounced differences in the observed volumes, while for the Arrhythmia dataset, the highest tested threshold of 1000 achieved the best performance. The NDP indicator aligned with expectations, displaying higher values for the more comprehensive models. However, the Spacing Indicator (SP) did not follow the anticipated trend of increased uniformity with the more thorough models, as greater uniformity was not consistently observed with the more complex models. Additionally, the SP showed reduced performance for the Arrhythmia and Intracranial Hemorrhage datasets.

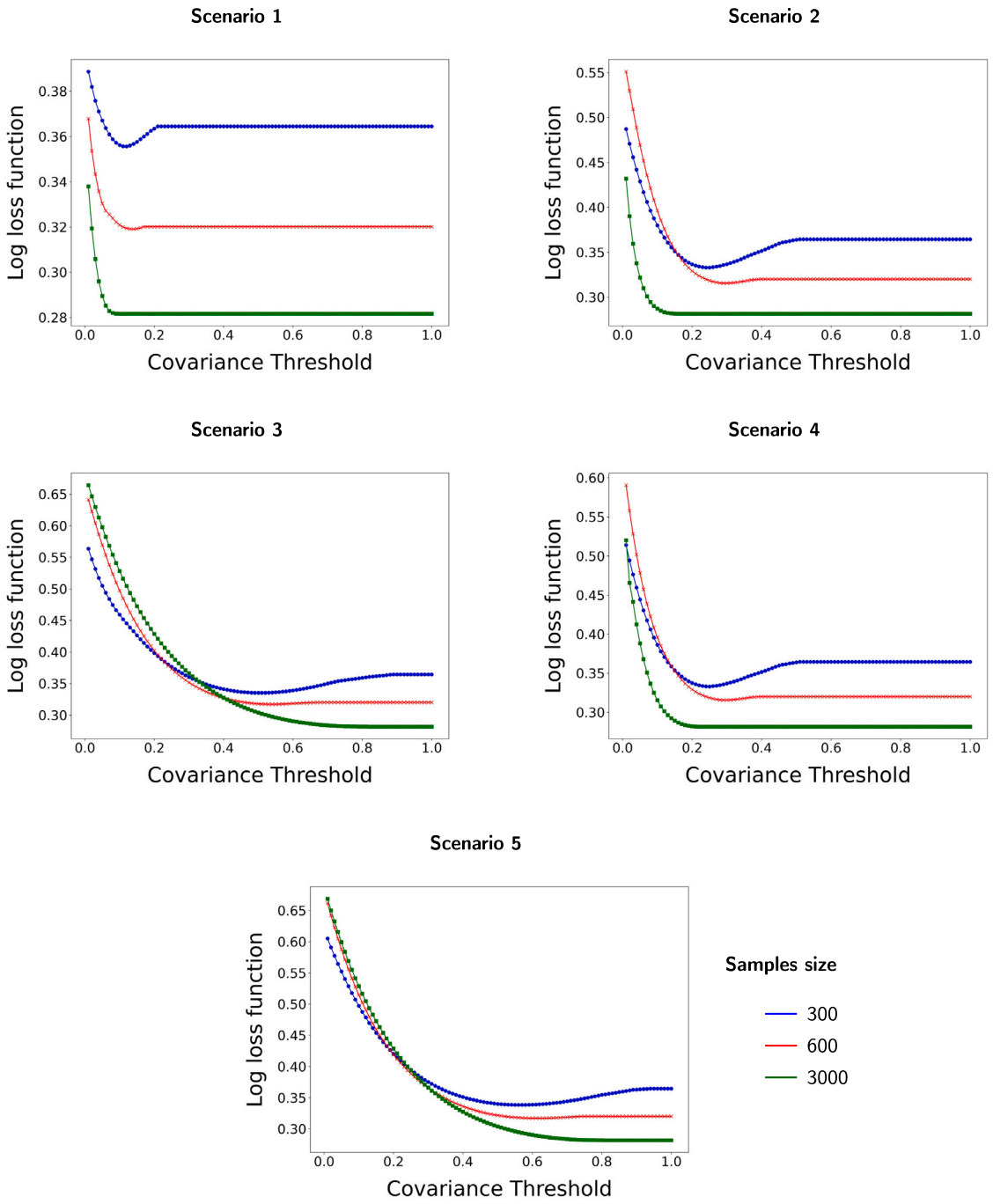


Fig. 6. Comparison of the performance of Model (4) as a function of each threshold according to different simulated sample sizes.

7. Conclusion

This study proposes a new approach to promoting fairness in machine learning models by incorporating multiple fairness constraints directly into the optimization process. Unlike previous methods that address sensitive attributes in isolation, our formulation allows for the simultaneous consideration of variables such as age, gender, and ethnicity, supporting a more comprehensive and intersectional analysis.

The proposed model ensures fairness by explicitly incorporating multiple demographic parity constraints into the optimization process. Unlike traditional methods that address fairness attribute by attribute, our approach allows for the simultaneous control of several sensitive variables – specifically gender, age, and ethnicity – enabling an intersectional evaluation of fairness. Empirical results on both small (Arrhythmia) and large-scale (MIMIC-III) datasets demonstrate that the model substantially reduces disparities in positive prediction rates and true positive rates across protected groups. For instance, demographic parity differences dropped from 0.4543 to below 0.1 for gender and age groups, and from 0.0110 to below 0.005 for ethnic groups. Moreover, fairness indicators such as Hypervolume (HV), Non-Dominated Proportion (NDP), and Spacing (SP) improved without compromising predictive performance (AP, precision, and recall). These findings confirm that the model achieves a more balanced distribution of predictions across groups, reducing bias and promoting equitable outcomes.

Using datasets of different scales (Arrhythmia and MIMIC-III), we show that jointly applying demographic parity constraints significantly reduces disparities across groups without compromising performance metrics such as AP, precision, and recall. Parity differences dropped,

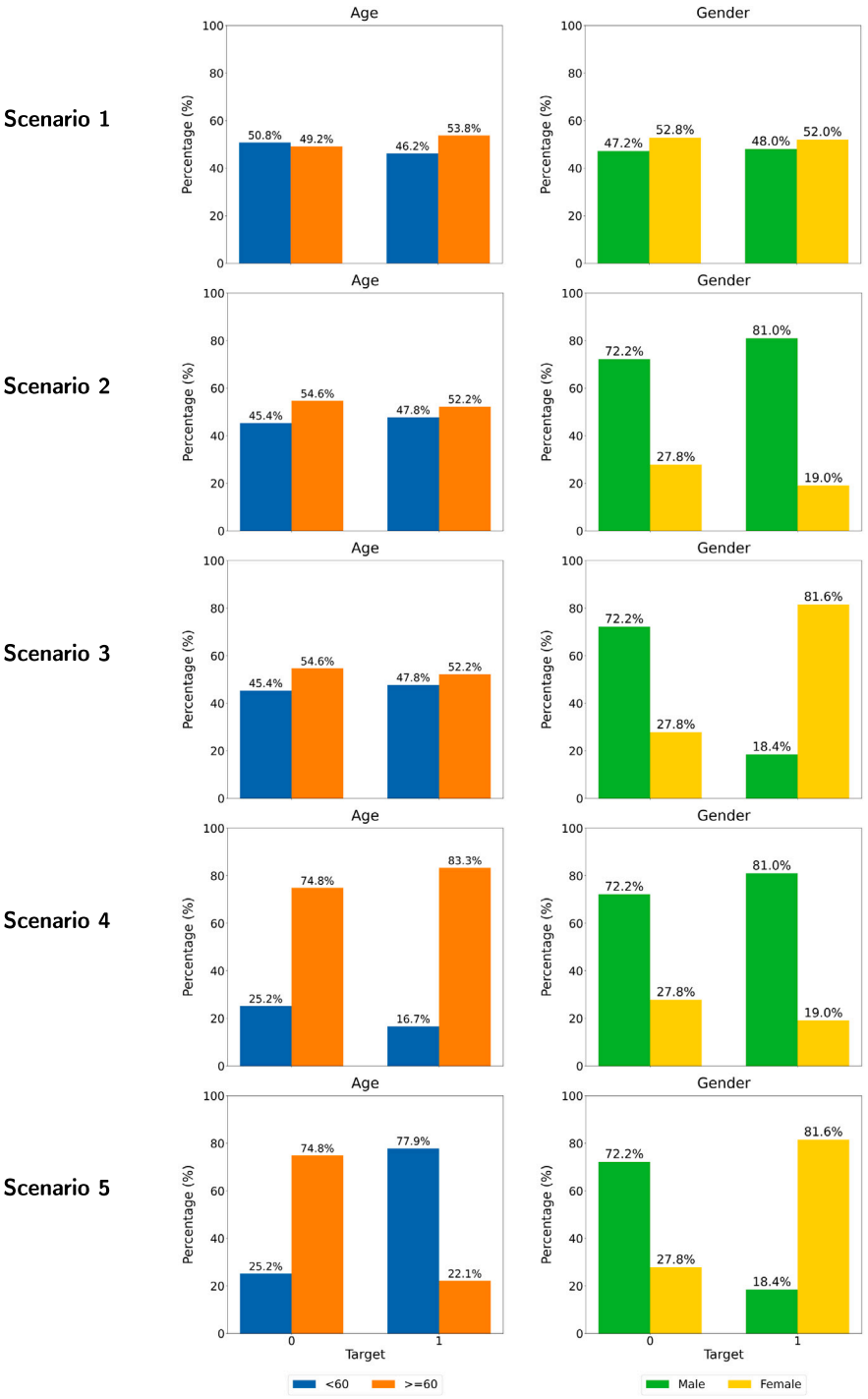


Fig. S1. Distribution of the response by age for each tested simulation scenario.

for instance, from 0.4543 to below 0.1 for gender and age groups, and from 0.0110 to under 0.005 for ethnic groups.

To evaluate the trade-offs between fairness and performance, we adopted tools from multi-objective optimization, including Hypervolume (HV) and the Non-Dominated Proportion (NDP), a metric proposed in this work. The results indicate that models with multiple constraints achieve higher HV and NDP values, reflecting a better balance between fairness and accuracy, and outperforming traditional single-constraint approaches. Wilcoxon tests confirmed the statistical significance of these gains.

In the clinical application for diabetes detection using MIMIC-III, the model reduced disparities in positive prediction rates and true positive rates among historically underrepresented subgroups, such as women, ethnic minorities, and older patients. Indicators like HV, NDP, and SP showed relevant improvements without compromising accuracy, leading to more equitable and reliable diagnoses.

Despite these advances, the use of logistic regression limits the ability to capture nonlinear interactions. Applying fairness constraints to more complex models, such as neural networks, remains challenging and requires tailored methodologies. Moreover, the focus on tabular

data may limit the generalizability of results to domains such as images or time series. Future work should explore these directions.

CRediT authorship contribution statement

Maíra Blumer Fatoretto: Writing – review & editing, Writing – original draft, Visualization, Validation, Methodology, Investigation, Funding acquisition, Formal analysis, Data curation, Conceptualization. **Gökhan Özbulak:** Investigation. **André Anjos:** Project administration, Methodology, Investigation, Funding acquisition. **Lilian Berton:** Resources, Project administration, Methodology, Funding acquisition, Formal analysis.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Maira Blumer Fatoretto reports financial support was provided by State of Sao Paulo Research Foundation. If there are other authors, they declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

We acknowledge the financial support received from the São Paulo Research Foundation (FAPESP), under Grant Numbers [21/14725-3 and 23/12468-9].

We acknowledge Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq).

We acknowledge Projeto Sistema de Computação Petaflópica do SINAPAD/2014 (Santos Dumont), process 01.14.192.00.

We also appreciate the support from Swiss National Science Foundation (SNSF), Switzerland under Grant Number [200021E_214653].

Appendix. Supplementary materials

See Fig. S1.

Data availability

Data will be made available on request.

References

[1] S. Barocas, M. Hardt, A. Narayanan, Fairness and Machine Learning: Limitations and Opportunities, MIT Press, 2023.

[2] R.T. Rabonato, L. Berton, A systematic review of fairness in machine learning, AI Ethics (2024) 1–12.

[3] D. Pessach, E. Shmueli, A review on fairness in machine learning, ACM Comput. Surv. 55 (3) (2022) 1–44.

[4] Z. Tang, J. Zhang, K. Zhang, What-is and how-to for fairness in machine learning: A survey, reflection, and perspective, ACM Comput. Surv. 55 (13s) (2023) 1–37.

[5] D. Queiroz, A. Carlos, A. Anjos, L. Berton, Fair foundation models for medical image analysis: Challenges and perspectives, 2025, arXiv preprint arXiv:2502.16841.

[6] P. Garg, J. Villasenor, V. Foggo, Fairness metrics: A comparative analysis, in: 2020 IEEE International Conference on Big Data (Big Data), IEEE, 2020, pp. 3662–3666.

[7] J.S. Franklin, K. Bhanot, M. Ghalwash, K.P. Bennett, J. McCusker, D.L. McGuinness, An ontology for fairness metrics, in: Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society, 2022, pp. 265–275.

[8] H. Gong, C. Guo, Influence maximization considering fairness: A multi-objective optimization approach with prior knowledge, Expert Syst. Appl. 214 (2023) 119138.

[9] S. Dandl, F. Pfisterer, B. Bischl, Multi-objective counterfactual fairness, in: Proceedings of the Genetic and Evolutionary Computation Conference Companion, 2022, pp. 328–331.

[10] H. Wu, C. Ma, B. Mitra, F. Diaz, X. Liu, A multi-objective optimization framework for multi-stakeholder fairness-aware recommendation, ACM Trans. Inf. Syst. 41 (2) (2022) 1–29.

[11] K. Padh, D. Antognini, E. Lejal-Glaude, B. Faltings, C. Musat, Addressing fairness in classification with a model-agnostic multi-objective algorithm, in: Uncertainty in Artificial Intelligence, PMLR, 2021, pp. 600–609.

[12] S. Liu, L.N. Vicente, Accuracy and fairness trade-offs in machine learning: A stochastic multi-objective approach, Comput. Manag. Sci. 19 (3) (2022) 513–537.

[13] Q. Hao, Q. Xu, Z. Yang, Q. Huang, Pareto optimality for fairness-constrained collaborative filtering, in: Proceedings of the 29th ACM International Conference on Multimedia, 2021, pp. 5619–5627.

[14] G. Özbulak, O. Jimenez-del Toro, M. Fatoretto, L. Berton, A. Anjos, A multi-objective evaluation framework for analyzing utility-fairness trade-offs in machine learning systems, 2025, arXiv preprint arXiv:2503.11120.

[15] E. Zitzler, K. Deb, L. Thiele, Comparison of multiobjective evolutionary algorithms: Empirical results, Evol. Comput. 8 (2) (2000) 173–195.

[16] N. Riquelme, C. Von Lücken, B. Baran, Performance metrics in multi-objective optimization, in: 2015 Latin American Computing Conference, CLEI, IEEE, 2015, pp. 1–11.

[17] C. Audet, J. Bigeon, D. Cartier, S. Le Digabel, L. Salomon, Performance indicators in multiobjective optimization, European J. Oper. Res. 292 (2) (2021) 397–422.

[18] E. Zitzler, Evolutionary Algorithms for Multiobjective Optimization: Methods and Applications, vol. 63, Shaker Ithaca, 1999.

[19] E. Del Barrio, P. Gordaliza, J.-M. Loubes, Review of mathematical frameworks for fairness in machine learning, 2020, arXiv preprint arXiv:2005.13755.

[20] M.B. Zafar, I. Valera, M.G. Rogriguez, K.P. Gummadi, Fairness constraints: Mechanisms for fair classification, in: Artificial Intelligence and Statistics, PMLR, 2017, pp. 962–970.

[21] A. Cotter, M. Gupta, H. Jiang, N. Srebro, K. Sridharan, S. Wang, B. Woodworth, S. You, Training well-generalizing classifiers for fairness metrics and other data-dependent constraints, in: International Conference on Machine Learning, PMLR, 2019, pp. 1397–1405.

[22] J. Komiyama, A. Takeda, J. Honda, H. Shimao, Nonconvex optimization for regression with fairness constraints, in: International Conference on Machine Learning, PMLR, 2018, pp. 2737–2746.

[23] H.A. Guvenir, B. Acar, G. Muderrisoglu, M. Tanik, D. Dua, Arrhythmia data set, UCI Mach. Learn. Repos. (1998) URL https://archive.ics.uci.edu/ml/datasets/Arrhythmia.

[24] A.E.W. Johnson, T.J. Pollard, L. Shen, et al., The MIMIC-III clinical database: A free intensive care unit database for research, Sci. Data 3 (2016) 160035, http://dx.doi.org/10.1038/sdata.2016.35.

[25] A. Castelnovo, R. Crupi, G. Greco, D. Regoli, I.G. Penco, A.C. Cosentini, A clarification of the nuances in the fairness metrics landscape, Sci. Rep. 12 (1) (2022) 4209.

[26] T. Calders, S. Verwer, Three naive bayes approaches for discrimination-free classification, Data Min. Knowl. Discov. 21 (2010) 277–292.

[27] M. Vincent, scikit-fairness, 2019, https://github.com/koaning/scikit-fairness.

[28] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, E. Duchesnay, Scikit-learn: Machine learning in Python, J. Mach. Learn. Res. 12 (2011) 2825–2830.

[29] C. Coello Coello, Evolutionary multi-objective optimization: a historical view of the field, IEEE Comput. Intell. Mag. 1 (1) (2006) 28–36, http://dx.doi.org/10.1109/MCI.2006.1597059, URL https://ieeexplore.ieee.org/abstract/document/1597059. Conference Name: IEEE Computational Intelligence Magazine.

[30] K. Deb, Multi-objective optimisation using evolutionary algorithms: An introduction, in: L. Wang, A.H.C. Ng, K. Deb (Eds.), Multi-Objective Evolutionary Optimisation for Product Design and Manufacturing, Springer, London, 2011, pp. 3–34, http://dx.doi.org/10.1007/978-0-85729-652-8_1.

[31] N. Riquelme, C. Von Lucken, B. Baran, Performance metrics in multi-objective optimization, in: 2015 Latin American Computing Conference (CLEI), IEEE, Arequipa, Peru, 2015, pp. 1–11, http://dx.doi.org/10.1109/CLEI.2015.7360024, URL http://ieeexplore.ieee.org/document/7360024/.

[32] E. Zitzler, L. Thiele, M. Laumanns, C. Fonseca, V. Da Fonseca, Performance assessment of multiobjective optimizers: an analysis and review, IEEE Trans. Evol. Comput. 7 (2) (2003) 117–132, http://dx.doi.org/10.1109/TEVC.2003.810758, URL http://ieeexplore.ieee.org/document/1197687/.

[33] S. Jiang, Y.-S. Ong, J. Zhang, L. Feng, Consistencies and contradictions of performance metrics in multiobjective optimization, IEEE Trans. Cybern. 44 (12) (2014) 2391–2404, http://dx.doi.org/10.1109/TCYB.2014.2307319, URL http://ieeexplore.ieee.org/document/6766232/.

[34] J.R. Schott, Fault tolerant design using single and multicriteria genetic algorithm optimization (Ph.D. thesis), Massachusetts Institute of Technology, 1995.

[35] F. Wilcoxon, Individual comparisons by ranking methods, in: Breakthroughs in Statistics: Methodology and Distribution, Springer, 1992, pp. 196–202.