Noname manuscript No.
(will be inserted by the editor)

Enforcing fairness using ensemble of diverse Pareto-optimal models

Vitória Guardieiro · Marcos M. Raimundo · Jorge Poco

Received: date / Accepted: date

Abstract One of the main challenges of machine learning is to ensure that its applications do not generate or propagate unfair discrimination based on sensitive characteristics such as gender, race, and ethnicity. Research in this area typically limits models to a level of discrimination quantified by an equity metric (usually the "benefit" discrepancy between privileged and non-privileged groups). However, when models reduce bias, they may also reduce their performance (e.g., accuracy, F1 score). Therefore, we have to optimize contradictory metrics (performance and fairness) at the same time. This problem is well characterized as a multi-objective optimization (MOO) problem. In this study, we use MOO methods to minimize the difference between groups, maximize the benefits for each group, and preserve performance. We search for the best trade-off models in binary classification problems and aggregate them using ensemble filtering and voting procedures. The aggregation of models with different levels of benefits for each group improves robustness regarding performance and fairness. We compared our approach with other known methodologies, using logistic regression as a benchmark for comparison. The proposed methods obtained interesting results: i) multi-objective training found models that are similar to or better than the adversarial methods and are more diverse in terms of fairness and accuracy metrics, ii) multi-objective selection was able to improve the balance between fairness and accuracy compared to selection with a single metric, and iii) the final predictor found models with higher fairness without sacrificing much accuracy.

Vitória Guardieiro
Fundação Getúlio Vargas, Brazil, E-mail: vitoriaguardieiro@gmail.com,

Marcos M. Raimundo
Fundação Getúlio Vargas, Brazil, and University of Campinas, Brazil E-mail: mraimundo@ic.unicamp.br,

Jorge Poco
Fundação Getúlio Vargas, Brazil, E-mail: jorge.poco@fgv.br

Keywords Algorithmic Fairness, Bias Mitigation, Multi-objective Optimization, Machine Learning, Ensemble Learning, Classification.

1 Introduction

Several incidents revealed unfair and discriminatory behavior by artificial intelligence models: (1) Amazon's AI discriminated against women penalizing resumes that contained the word “women's,” as in “women's chess club captain” [1]. (2) AI-powered COMPAS system discriminated against African-Americans, predicting a higher risk of criminal re-offending for African-Americans than Caucasians with the same profile [2–4]. These and many other examples [5, 6] occur because these models simulate the human behavior embedded in data. The most accurate artificial intelligence model might propagate such injustices if the data contain unfair or discriminatory decisions.

Fair AI aims to find accurate models that can reduce the biases in the society portrayed in the data. Different techniques induce fairness in models; a particular class is in-processing methods that reduce discrimination in AI models by constraining the bias in the training/learning process and tweaking model parameters after the learning process. When the goal of achieving fairness is also considered, the resulting model performance metrics often worsen. Several equity metrics have been shown to directly conflict with accuracy [7–9], in some cases only under certain conditions [10, 11]. Based on this, if we were to employ these metrics directly, increasing fairness (*i.e.*, reducing discrimination) could reduce maximum model accuracy (and vice versa). These facts constitute the ideal conditions for employing a posteriori multi-objective optimization in model training.

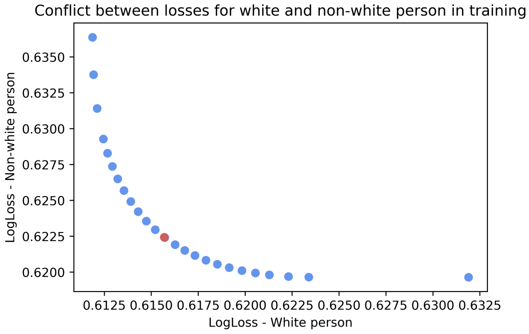


Fig. 1: The conflict between learning errors for White and non-White individuals on the ProPublica dataset. Each point is a different logistic regression classifier with its respective learning errors. The blue points were generated by a multi-objective approach, while the red point is the default logistic regression classifier.

The conflict between fairness and performance is often due to the trade-off between performance for each group of individuals. Each point in Figure 1 shows a Pareto-optimal between erring prediction for White and non-White individuals (errors from logistic regression on ProPublica dataset [12] from COMPAS system) — it shows that increasing one's performance decreases another. Also, the red point in Figure 1 shows a higher error for non-White people when using a default logistic regression. In this example, we can see a diversity of preference combinations, ranging from favoring White to favoring non-White individuals and the naivety of a default learner (in red), which slightly favors White individuals. This approach of sampling diverse options is one of the advantages of reducing a fairness metric because we might need a model that favors a class harder to learn (usually the discriminated class; *e.g.*, non-White individuals) on the training set to be effective in real life. We wish to produce a single result from several models created by multi-objective optimization. The simplest way to do this is to select a single model from the provided ensemble based on the trade-off between objectives. However, previous studies showed that ensemble aggregation is a better strategy in various contexts such as unbalanced classification [13], multi-task classification [14], and multi-class classification [15]. In addition, model subset selection allows a decision-maker or a machine to select the solution that best reduces inequality between social groups without causing a drastic performance loss.

Our methodology adjusts models for binary classification, minimizing the learning error and the discrimination. The base classifier is a logistic regression modeled as a multi-objective problem to optimize several metrics simultaneously. The contribution is a framework consisting of three steps. First, we indirectly model the discrimination metrics using two different objective functions: one based on accuracy and the other based on an acceptance metric (*i.e.*, a function that evaluates the average probability of a group achieving the desired outcome). Second, multi-objective optimization finds a set of models with the best trade-off among the objective functions. Third, the decision-maker selects and aggregates models using ensemble learning, thus creating a single, application-oriented, robust predictor.

The main contributions of this research are:

- Adaptations to the logistic regression formulation to allow optimization of both performance and classification rates of different groups of individuals;
- A framework that uses a multi-objective method on flexible adaptations of logistic regression to find a good representation of the trade-offs between performance and discrimination metrics, thus creating a set of classifiers;
- An ensemble-based selection and aggregation procedure that allows the decision-maker to choose multiple classifiers with distinct trade-offs between accuracy and performance metrics, creating more robust predictions;
- A set of experiments comparing the ability of methods described in the literature to create diverse models and the quality of the ensemble output when models are selected using different user preferences.

2 Related Work

Research on discrimination and injustice generated by artificial intelligence addresses different perspectives, from quantifying discrimination through metrics [16–18] to developing strategies to reduce it. This study proposes two approaches to reduce bias in machine learning models using multi-objective optimization.

Fair models usually modify well-known strategies such as logistic regression. We can divide them into three families [19,20]: (1) **Pre-processing** techniques compensate bias contained in the database before the creation of the artificial intelligence model, for instance, by creating a new representation of each sample [21], or by creating weights for each sample forcing the model to discriminate less [22]. (2) **In-processing** techniques modify the learning algorithms to address discrimination during the training phase, for instance, by constraining the model to meet a discrimination metric [23] or performing a sequence of classifications adding weights to reduce discrimination [24]. (3) **Post-processing** techniques change components of the artificial intelligence model after learning, for instance, by tweaking classification thresholds [25] or using adversarial learning [26]. Since this study modifies the optimization problem in training, thus we deepen the in-processing literature.

A subgroup of in-processing approaches re-model optimization problem that consists of the training phase. In general, they either add a fairness regularizer that penalizes discrimination in the objective function of the model [27–29] or add a discrimination-aware constraint — for instance, limiting the difference between learning errors between sensitive and nonsensitive groups [23] or limiting divergence among groups' odds of obtaining a resource [30]. However, constrained or penalized optimizations usually use a parameter to find a single model with a specific performance vs. fairness trade-off, forcing the decision-maker to make a priori decisions about the preference between performance and fairness [31], or arbitrarily vary the parameter to search for a trade-off. More theoretical research shows that optimizing some fairness metrics will generate conflict with accuracy (accuracy vs. demographic parity [7, 8], accuracy vs. coefficient of variation [9]). It is noteworthy that equal opportunity might also conflict with accuracy, but it can be alleviated under ideal conditions [10], and group fairness and ideal fairness can also be conflicting [9] due to a poor conceptualization of the problem [11].

This trade-off between fairness and model performance supports using multi-objective optimization methods capable of simultaneously dealing with several conflicting objectives [32]. This approach has been used in previous studies and can find the set of solutions for the optimization problem that have the best trade-offs between the objectives (the Pareto optimal solutions) or a single solution that minimizes the objective with the highest value (e.g., the minimax approach). Following the Pareto approach, previous work aimed to simultaneously optimize the model performance (in terms of accuracy or loss functions) and fairness (in terms of one or more pre-defined metrics) by optimizing the model's hyperparameters [33]; by optimizing the model's parameters via a stochastic multi-objective algorithm [34] and a model-agnostic gradient-based multi-objective algorithm [35]; and by evolving a population of models [36, 37]. Meanwhile, the method proposed in [38] considers the learning loss for each group that might suffer from discrimination as the objective function and finds the minimax solution.

One of the approaches presented in this work also uses groups' losses as its objective. However, unlike [38], we follow a Pareto approach that finds multiple solutions. We also propose a multi-objective modeling called *acceptance loss* that focuses on another category of fairness beyond error-based losses. By doing so, we are exploring multi-objective optimization to find models with different importance for each group (discriminated and non-discriminated). This way, we create models that favor each group in different intensities, representing an advantage over models that constrain fairness metrics [27–29]. Suppose one group has more challenging samples (in training). In that case, it will never privilege the other group, which might be a fairer model in real life (simulated by validation and test sets). Our proposal can generate more diverse models for users (or a metric) to select a model or create an ensemble by aggregating multiple chosen models.

The use of ensemble learning for training fair models has been explored in previous studies. Grgic-Hlaca et al. [39] propose using an ensemble method that randomly selects a classifier from a pool. This approach was later compared with majority voting ensembles with hard and soft voting [40]. The conclusion was that each ensemble strategy works best depending on the fairness metric used. Bhargava et al. [41] use an explainability technique (LIME) to produce a pool of classifiers for a soft majority voting ensemble. Another studied approach is the combination of bagging and boosting [42]. Focusing only on boosting, some works [43, 44] propose adapting AdaBoost so that the weights of the mispredicted instances are updated considering the fairness of the ensemble members. Furthermore, other works suggest adaptations to the training of Random Forests [45–48]. Lastly, focusing on the fairness-performance trade-off, Chen et al. [49] train a performance-focused and a fairness-focused model and then combine the probabilities assigned by each one.

A key characteristic of the ensemble techniques is the requirement for a pool of diverse models. The Pareto approach for multi-objective optimization is capable of generating such a pool. Moreover, the models are trained to have the optimal trade-offs between the objective functions in the training set. Because of this, diverse studies have proposed the use of Pareto optimal models as the pool for ensemble techniques [13–15]. Recently, Zhang et al. [37] proposed to mitigate unfairness using an ensemble of Pareto optimal models. However, this study evolves populations of models to optimize accuracy and ten fairness metrics simultaneously. In comparison, here we propose optimizing the loss function for each group via a deterministic optimization algorithm, with a much smaller number of objectives (in most cases) than the previously mentioned study.

3 Training models that optimize multiple objectives

In this section, we first propose the training of models that simultaneously optimize a set of conflicting metrics through multi-objective optimization. Then, we define what it means for a model to be optimal (or Pareto-optimal) in this scenario. After that, we present how the training can be carried out. Furthermore, we describe the restrictions on models and functions to make this possible. In the next section, we propose our objective functions.

Suppose $\mathcal{D} = \{(\mathbf{x}_i, y_i, a_i)\}_{i=1}^{N}$ is a dataset where $\mathbf{x}_i \in \mathcal{X}$ are the input features for an individual i, $y_i \in \mathcal{Y}$ is the output target, $a_i \in \mathcal{A}$ is a *protected feature* (*e.g.*, gender or ethnicity), and N is the number of individuals (samples). Let $f_\theta \in \mathcal{H}$ be a model trained to infer y from $\mathbf{x}$ with parameters $\theta \in \Theta$, $f_\theta(\mathbf{x}) : \mathcal{X} \to \mathcal{Y}$. Also, consider $g_1(\theta), g_2(\theta), \ldots, g_m(\theta)$ as conflicting functions of the model f_θ evaluated on $\mathcal{D}$ (for instance, loss functions), which are called objective functions. Because we have multiple conflicting objectives (for example the classification loss of different learning tasks [14]), the learning of f_θ can be conceived as a multi-objective optimization problem (MOOP) [50]:

$$\begin{aligned} \underset{\theta}{\text{minimize}} \quad & G(\theta) = [g_1(\theta), \ldots, g_m(\theta)] \\ \text{subject to} \quad & \theta \in \Theta \\ & G : \Theta \to \Psi, \Psi \subset \mathbb{R}^m \end{aligned} \tag{1}$$

In multi-objective optimization, optimality is defined by the concept of dominance. For two possible parameters θ_i and θ_j, it is said that θ_i **dominates** θ_j, noted as $G(\theta_i) \prec G(\theta_j)$, if $g_k(\theta_i) \leq g_k(\theta_j), \forall k \in 1, \ldots, m$ and $\exists\, l : g_l(\theta_i) < g_l(\theta_j)$. This means that we prefer θ_i over θ_j for any chosen trade-off of the functions g. So, we can define an optimal or Pareto-optimal model as:

Definition 31 (Pareto-optimal model) *A model $f_{\theta^*} \in \mathcal{H}$ is Pareto-optimal if it is not dominated by any other model $f_\theta \in \mathcal{H}$,* i.e.*, $\nexists f_\theta \in \mathcal{H} | G(\theta) \prec G(\theta^*)$.*

The set of all Pareto-optimal models is called **Pareto frontier**, which we wish to find. With the Pareto frontier, the decision-maker can choose the trade-off of the objective functions $g_1(\theta), \ldots, g_m(\theta)$ knowing exactly how much favoring a function $g_i(\theta)$ impacts the other objective functions.

It is possible to transform the MOOP into a single objective optimization problem by multiplying $g_1(\theta), \ldots, g_m(\theta)$ by a weight vector $\mathbf{w} \in \mathbb{R}^m$. This is called the weighted sum method [51], and the weighted optimization problem is defined as:

$$\begin{aligned} \underset{\theta}{\text{minimize}} \quad & \mathbf{w}^T G(\theta) \\ \text{subject to} \quad & \theta \in \Theta \\ & G : \Theta \to \Psi, \Psi \subset \mathbb{R}^m \end{aligned} \tag{2}$$

where $w_i \geq 0, \forall i \in \{1, 2, \ldots, m\}$ and $\mathbf{w}^T \mathbf{1} = 1$.

If the functions $g_1(\theta), \ldots, g_m(\theta)$ are convex, then every solution to Equation 2 is a Pareto-optimal model. If the hypothesis space $\mathcal{H}$ is convex, then for every Pareto-optimal model f_θ there is a weight vector $\mathbf{w}$ such that θ is the solution for the weighted sum method of $\mathbf{w}$ [50]. So, if both the functions and the hypothesis space are convex, it is possible to find a good representation of the Pareto-frontier of Equation 1 by solving the weighted sum method for a well-sampled set of weights. This strategy is precisely what the NISE (Noninferior Set Estimation) method [52] does.

However, NISE fails when dealing with problems with three or more objectives [53]. For this reason, we will employ the MONISE (Many-Objective Noninferior Set Estimation) method [53], which extends NISE to more than two objective functions, and it has a good performance for convex problems with high-dimensionality [53], thus justifying its use. From a set of Pareto-optimal models, MONISE can find a new weight vector $\mathbf{w}$ that corresponds, when solving Equation 2, to a new Pareto-optimal model. Through an iterative procedure, we can employ MONISE to find the R most representative models from the Pareto frontier.

Therefore, given the dataset $\mathcal{D}$, the hypothesis space $\mathcal{H}$, the objective functions $g_1(\theta), \ldots, g_m(\theta)$, and a way to optimize the weighted problem (defined in Equation 2), we are able to find a well-representative subset of the Pareto frontier of (defined in Equation 1). We approach the problem of training fairness-aware models through the multi-objective optimization of protected-group-based functions of the model parameters. After generating this set of models, we can employ ensemble filtering and aggregation to combine them and perform the prediction.

4 Fairness as objective functions

In this study, we focus mainly on Group Fairness, where the dataset $\mathcal{D}$ is stratified into $|\mathcal{A}|$ groups or sub-populations based on the possible values for the protected feature. Let $\mathcal{D}^a$ be the group of individuals with protected (or sensitive) feature equal to $a \in \mathcal{A}$, *i.e.*, $\mathcal{D}^a = \{(\mathbf{x}_i, y_i) | a_i = a\}$. Under Group Fairness, the model f_θ is considered *fair* if a chosen function $\delta(f_\theta)$ achieves similar values for each $\mathcal{D}^a$ for every $a \in \mathcal{A}$. Different functions δ define different fairness metrics.

Regardless of which fairness metric is considered, there is a conflict between the fairness and prediction performance of a model. For that reason, we modify the model training to consider both. Usually, that is done either by optimizing accuracy while constraining fairness or optimizing fairness while constraining accuracy. However, that forces the decision-maker to choose over the fairness/accuracy trade-off without knowing how much favoring one disfavors the other. Although training models with several different constraint values can provide information to the decision-maker, there is no guarantee that the trained models are Pareto-optimal for fairness/accuracy, especially when using more than two metrics to evaluate the model.

Therefore, we propose two multi-objective optimization formulations that find a set of Pareto-optimal models representing the best trade-off solutions that favor each group (privileged and unprivileged) differently. One formulation models the classification loss for each group (Section 4.1), and the other models the access to a beneficial output (the desired outcome) loss (Section 4.2). The advantage of treating each group's classification/beneficial output loss as an objective relies on finding different models; each one focuses on privileged and unprivileged groups instead of minimizing fairness. This last minimization might always favor the same group, while the Pareto-representation will find a good representation of trade-offs for both groups. Those functions are defined for binary classification problems ($y_i \in \{0, 1\}$ for $i \in (1, \dots, N)$), and the models are logistic regressions.

4.1 Objectives for Equal Risk Fairness

Some of the definitions of a fair model are based on predictive risk, that is, f_θ is seen as fair if the expected loss $E_{X,Y|A=a}[l(Y, f_\theta(X))]$ satisfies some restriction for the $|\mathcal{A}|$ sub-populations. Such a restriction can be imposed over the expected loss of each sub-population to be exactly equal, or their difference is lower than the desired value, in what is called Equal Risk Fairness. Another approach is to find the model f_θ that has the minimum possible risk for the group with maximum risk [38]. In our first approach, we propose multi-objective modeling of Equal Risk Fairness, where the expected loss of each group is an objective function.

Let $l^a(\theta)$ be the loss of the model f_θ for the group $\mathcal{D}^a$. For simplicity, consider that $\mathcal{A} = \{1, 2, \dots, |\mathcal{A}|\}$. We propose to use the $l^a(\theta)$ as objective functions for the multi-objective optimization problem (defined in Equation 1). That way, the weighted problem becomes:

$$\begin{aligned} \underset{\theta}{\text{minimize}} \quad & \sum_{j=1}^{|\mathcal{A}|} w_j \cdot l^j(\theta) \\ \text{subject to} \quad & \theta \in \Theta \end{aligned} \tag{3}$$

Theorem 1 *If exists a model θ that equals the losses among sensible groups $l^1(\theta) = \ldots = l^{|A|}(\theta)$, then there is a weight vector $\mathbf{w}$ that finds θ^* using Equation 3 is such a way that $\mathbf{l}(\theta) = \{l^1(\theta), \dots, l^{|A|}(\theta)\}$ is weakly dominated* [1] *by $\mathbf{l}(\theta^*)$.*

Proof Given that θ is feasible, then, exists a Pareto-optimal solution θ^* that $l^i(\theta^*) \leq l^i(\theta), \forall i \in \{1, \dots, |A|\}$. And, given that $l^1(\bullet), \dots, l^{|A|}(\bullet)$ are convex functions by the Theorem 3.1.4 in [50], then, exists $\mathbf{w}$ that finds θ^*.

This theorem shows that it is possible to find models that dominate a model with equal/similar losses. The advantage of finding such a model is that we allow the optimization to find a model with better accuracy since we allow both groups to be better, thus dominating the equal/similar solution.

[1] For two possible parameters θ_i and θ_j, it is said that θ_i **weakly dominates** θ_j, noted as $G(\theta_i) \preceq G(\theta_j)$, if $g_k(\theta_i) \leq g_k(\theta_j), \forall k \in 1, \dots, m$.

4.1.1 Logistic Regression

Logistic regression models define the probability of an individual $\mathbf{x}_i$ having target $y_i = 1$ as the sigmoid function:

$$p(y_i = 1|\mathbf{x}_i, \theta) = \frac{1}{1 + e^{-\theta^T \mathbf{x}_i}}$$

where the parameter θ is the maximum likelihood estimator over the training set. The loss function of the logistic regression is $l(\theta) = -\sum_{i=1}^{N} \log p(y_i|\mathbf{x}_i, \theta)$.

Training a logistic regression that optimizes Equation 3 means finding θ that minimizes:

$$\begin{aligned} \sum_{j=1}^{|\mathcal{A}|} w_j \cdot l^j(\theta) &= -\sum_{j=1}^{|\mathcal{A}|} w_j \left[\sum_{(\mathbf{x}_i, y_i) \in \mathcal{D}^j} \log p(y_i|\mathbf{x}_i, \theta) \right] \\ &= -\sum_{i=1}^{N} w_{a_i} \log p(y_i|\mathbf{x}_i, \theta) \end{aligned} \tag{4}$$

Thus, given the weights w_j for $j = 1, \ldots, |\mathcal{A}|$, we can train a logistic regression model that optimizes the weighted problem on Equation 3 by training a standard logistic regression in which the samples are weighted according to their protected attribute. The weights that best find a representation of the Pareto-frontier are calculated through the iterative process with the MONISE method, as explained in Section 3.

4.2 Objectives for Demographic Parity Fairness

One limitation of our previous approach is that it does not consider the rate of the classifications achieved by each group, which is the basis of some Fairness definitions, such as Demographic Parity. This can limit the approach's ability to find suitable solutions under such a Fairness definition. For instance, if we consider a case where the sensitive attribute is highly correlated with the class label, then our approach based on Equal Risk Fairness may not explore models which imply similar classifications for each group. Even in more moderate scenarios, such an approach will not necessarily generate a set of diverse models considering the proportion of class labels for each class since this is not one of its objectives. Therefore, we propose a second approach, based on the Demographic Parity definition of Fairness, where the objectives explicitly consider the proportion of class labels for each group.

More formally, consider that $y_i = 1$ indicates a beneficial output and that the protected attribute is binary, with $a_i = 1$ indicating that the individual i is part of the protected group and $a_i = 0$ that it is not. Under the **Demographic Parity** definition, a model f_θ is considered fair if the probability of a random individual with features $\mathbf{X}$ to receive a beneficial classification ($f_\theta(\mathbf{X}) = 1$) is the same whether the individual is part of the protected group ($A = 1$) or not $(A = 0)$, so $P(f_\theta(\mathbf{X}) = 1|A = 0) = P(f_\theta(\mathbf{X}) = 1|A = 1)$. As it is often not possible to train a perfectly fair model, the demographic parity metric is defined as the proportion of the probabilities of receiving a positive rating given whether or not they belong to the protected group:

$$DP = \min\left(\frac{P(f_\theta(\mathbf{X}) = 1|A = 1)}{P(f_\theta(\mathbf{X}) = 1|A = 0)}, \frac{P(f_\theta(\mathbf{X}) = 1|A = 0)}{P(f_\theta(\mathbf{X}) = 1|A = 1)}\right) \tag{5}$$

The Demographic Parity is not convex, therefore it can not be used as an objective function for our Multi-Objective formulation. To train fair models under Demographic Parity, we define an *acceptance loss* function α_a, which is equal to the loss of the model f_θ for the group $\mathcal{D}^a$ if all individuals had target equal to 1 (*i.e.*, we calculate the loss of achieving the beneficial output, $\alpha_a = E_{p(X|A=a)}[\ell(f_\theta(X), 1)]$)). For the logistic regression, the acceptance loss function is given by $\alpha_a = -\sum_{(x_i)\in D^a} \log p(y = 1|x_i, \theta)$. Since not all individuals have target $y_i = 1$, α_a is conflicting with prediction performance, so we also use the overall loss as an objective function, resulting in the following weighted problem:

$$\begin{aligned} \underset{\theta}{\text{minimize}} \quad & \sum_{j=1}^{|\mathcal{A}|} w_j \cdot \alpha^j(\theta) + w_{|\mathcal{A}|+1} \cdot l(\theta) \\ \text{subject to} \quad & \theta \in \Theta \end{aligned} \tag{6}$$

Theorem 2 *For any model θ that equalize the acceptance among sensible groups $a^1(\theta) = \ldots = a^{|A|}(\theta)$, exists a weight vector $\mathbf{w}$ that finds θ^* using Equation 6 is such a way that $\mathbf{v}(\theta) = \{a^1(\theta), \ldots, a^{|A|}(\theta), l(\theta)\}$ is weakly dominated [1] by $\mathbf{v}(\theta^*)$.*

Proof Given that θ is feasible, exists a Pareto-optimal solution θ^* that $v_i(\theta^*) \leq v_i(\theta), \forall i \in \{1, \ldots, |A| + 1\}$. And given that $v_1(\bullet), \ldots, v_{|A|}(\bullet)$ are convex functions, by the Theorem 3.1.4 in [50] exists $\mathbf{w}$ that finds θ^*.

5 Proposed Framework and Model Aggregation

The proposed framework explores multiple Pareto-optimal models to allow robust classifiers for both accuracy and fairness metrics. Figure 2 depicts this framework in two phases: *Model Generation* (where a set of models are created); and *Model Aggregation* (where the models are combined to create a single predictor). In Section 4), we explained that *Weighted Problem* in *Model Generation* consists of modeling the classification problem to evidence the conflict of fairness goals among the groups that might suffer discrimination. Also, in Section 3, we describe how the *Pareto-optimal Models* are generated using MONISE. In this section, we will discuss *Model Aggregation* strategies to deal with the multiple models generated by Multi-Objective optimization to find a *Final Model* that will be able to predict values and be evaluated.

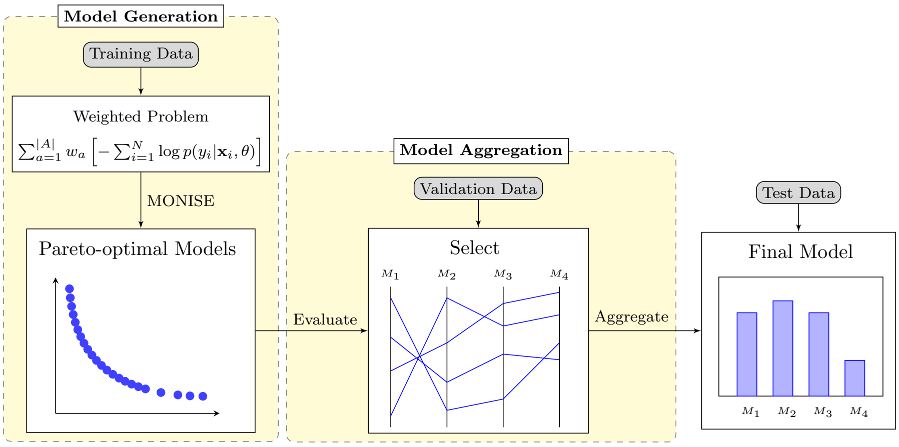


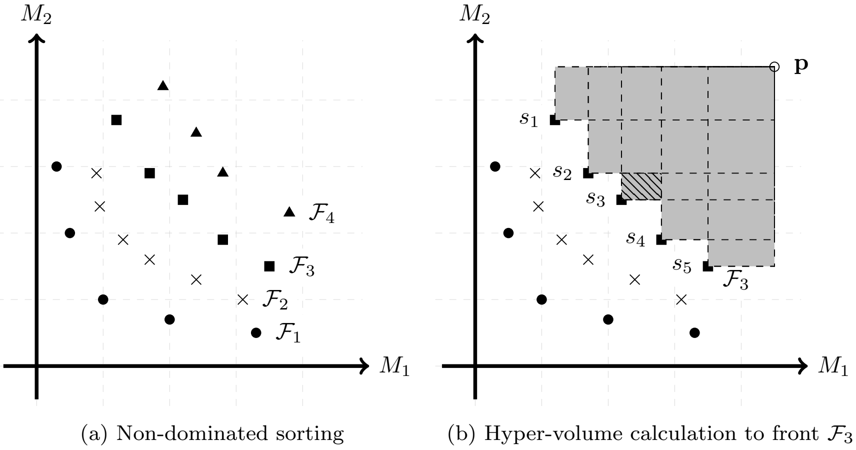
Fig. 2: Visual description of the proposed methodology

5.1 Model Aggregation

Ensemble methodologies combine multiple individual models generated for the same task into a more robust model than every one of them [54]. As shown in [14] and explored in our experiments, the Pareto-optimal models compose a well-diverse set. Each model specializes in a trade-off between different tasks or groups, which favors the implementation of ensemble methodologies. From a subset of Pareto-optimal models, selected according to some filter, we propose to combine them with a *hard majority voting* method: for each sample x_i, each model j will predict its classification $f(x, \theta_j)$ (the vote), and the outcome consists of the class with more votes.

Remember that we find R Pareto-optimal models. We propose filtering the following two steps (evaluated over a validation dataset):

1. **Select models above minimum performance:** in this step, we select the best $R_1 \leq R$ models w.r.t. their validation accuracy — this step is necessary to ensure the consistency of the results obtained. If we did not impose such a restriction, it could result in models that are fair for the training and validation data but with too low accuracy to be a useful predictor.
2. **Select according to a Multi-Objective Sorting using metrics as dimensions:** in this step, we select the best $R_2 \leq R_1$ models w.r.t. multiple metrics in the validation set. To choose models using various metrics simultaneously, we resort to *Non-Dominated Sorting* and *Hyper-volume Metric*, which we will explain later. We employ this sorting approach because we assume that our decision-maker can not select a single performance/fairness metric, and this approach will ensure that the selected models are the most representative of the trade-off.



(a) Non-dominated sorting (b) Hyper-volume calculation to front $\mathcal{F}_3$

Fig. 3: Strategies to select models using multiple metrics M_1 and M_2.

Non-Dominated Sorting is a procedure that sorts solutions to a multi-objective optimization problem. Figure 3a shows the division of solutions into sub-sets called **fronts**: $\mathcal{F}_1, \mathcal{F}_2, \mathcal{F}_3, \mathcal{F}_4$. The first front, $\mathcal{F}_1$, is the set of non-dominated solutions. The second front, $\mathcal{F}_2$, is the subset of solutions that become non-dominated when we remove the first front from the set. The remainder fronts, $\mathcal{F}_3$ and $\mathcal{F}_4$, are found the same way until we allocate all the solutions to a front. Following the example in Figure 3, suppose that we want to select 15 solutions. Thus we can choose the fronts $\mathcal{F}_1$ and $\mathcal{F}_2$, but adding the whole front $\mathcal{F}_3$ would result in 16 solutions. To selectively remove the additional solution, we will resort to a hyper-volume metric.

Hyper-volume Metric, also known as *size of dominated space*, is a procedure that calculates an index of how good the representation of a Pareto-frontier is. Figure 3b depicts in gray the hyper-volume for the solutions in $\mathcal{F}_3$. Knowing a reference point $\mathbf{p}$, the hyper-volume is the union of the hyper-cubes (in the example, rectangles) formed by $\mathbf{p}$ and each solution in $\mathcal{F}_3$. Another important trait is that we can calculate each solution's contribution to the hyper-volume. Figure 3b shows a hatchet area that corresponds to the contribution of the solution s_3. This contribution is the "lost" hyper-cube (in the example, rectangle) if we remove the solution s_3. Note that s_3 has the lowest contribution to the hyper-volume. Thus we can remove this solution and find the 15 solutions we need.

The procedure of selecting the fronts using *Non-Dominated Sorting* until we find more solutions than we need, then removing the solutions with the lowest *Hyper-volume Metric* is by SMS-EMOA [55]. We refer the reader to that document for further details.

Given the set of models that satisfied the minimum performance restriction, we evaluate them in three different fairness metrics: Demographic Parity [56]; Equality of Opportunity [25]; Coefficient of Variation [57]. The Non-Dominated Sorting is performed over those metrics, with or without accuracy as another metric. Therefore, the trade-off between fairness and performance is controlled by the following parameters: R_1, which is the number of models filtered by accuracy, so lower values of R_1 place higher importance on performance; R_2, which is the number of models selected through the non-dominated sorting of the fairness metrics, thus lower values of R_2 gives more emphasis to fairness. Next, we evaluate the proposed framework in a series of experiments regarding the diversity of the Pareto-optimal models, different forms of selecting the models for the aggregation, and compare the final result with other well-used methodologies.

6 Experiments

This section evaluates the performance of the methodologies proposed performance in four experiments. The first experiment evaluates the behavior of the Pareto-optimal models in their respective objective functions, comparing them with other approaches. The second aims to validate the advantage of the diversity of models generated by our approach. The third explores ensemble learning to aggregate the models generated in the first experiment in a more robust model. Finally, the last experiment compares the aggregated proposed models to other competing approaches.

6.1 Experiments setup

This section presents the datasets used in our experiments and the methodologies used to obtain the results of the experiments.

6.1.1 Datasets

We use three real datasets differing in applications for the experiments: addressing credit, income, and crime. It is well-known that these contexts already express discriminatory bias that might be perpetuated in machine learning models. These datasets are frequently used in Fairness, they are not class-balanced, and Bellamy et al. did the pre-processing in [58].

ProPublica [12]: This dataset was collected using the COMPAS (Correctional Offender Management Profiling for Alternative Sanctions) risk tool. It includes data from 6,167 arrested individuals and 10 features, including the incident level, gender (not used as a sensitive attribute), and race (used as a sensitive attribute). The goal is to predict whether the individual will be arrested again in two years.

Adult [59]: Also known as Census Income, it contains information about 48,842 individuals from the 1994 United States Census. The 14 attributes include gender (not used as a sensitive attribute) and race (used as a sensitive

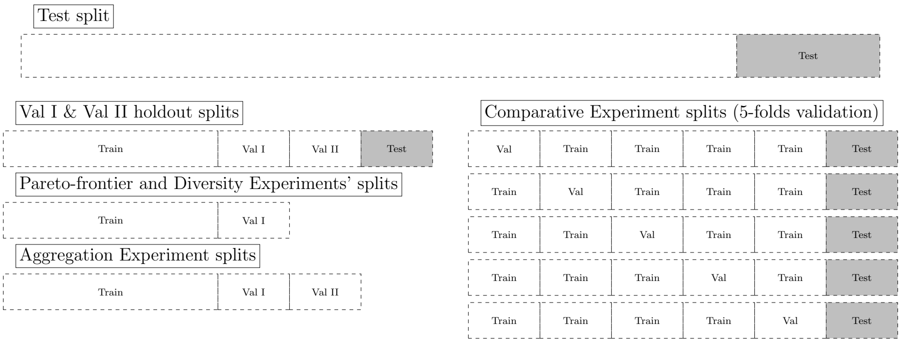


Fig. 4: Representation of the data split for the experiments.

attribute). The goal is to predict whether a particular individual receives more or less than $50,000 a year.

German [59]: The German Credit Data dataset contains 1,000 credit requests and 11 features, including credit amount, payment duration, request purpose, and personal information (including gender, which we use as a sensitive attribute). The goal is to predict whether the request was accepted or denied.

6.1.2 Data Splitting

To perform the experiments in Sections 6.3, 6.4, and 6.5, we adopted the data split depicted in Figure 4. First, we separate 30% of the data to be the Test Set that will be used only in the Comparative Experiment (Section 6.5). Then, for the Pareto-frontier, Diversity & Aggregation experiments, we split the remaining 70% of the data into Train (49% of the data), 10.5% into Validation I, and 10.5% into Validation II. For the Comparative experiment, we create a 5-fold cross-validation. The Pareto-frontier & Diversity experiments use the Train set to fit the parameters for the baseline and Pareto-optimal models and use the Validation I set to evaluate their properties. The Aggregation Experiment uses the Train set to fit a set of Pareto-optimal models, the Validation I set to select the models, and the Validation II to observe the performance of the aggregated predictors. The report of performances in Validation II is used to make decisions to conduct the Comparative Experiment; this is the reason not to use the Test set here. Thus, we only evaluate all models on the Test set in the Comparative experiment.

In this last experiment, we train the models in four Training folds, select the most appropriate model in the remaining Validation fold and evaluate the model in the Test Set. This procedure is done in the five visions generated in the 5-fold process, but the Test set is always the same (split before the 5-fold strategy). The final performance is the average for each of the five runs. Please note that although we used two validation sets for our experiments, our framework only requires one validation set.

6.1.3 Compared approaches

We compare our proposals with five similar strategies. Other methods also address binary classifications using Logistic Regression as a base classifier.

Logistic Regression (LogReg): It is our default baseline, it uses the sigmoid function $p(x, \theta) = \frac{e^{\theta^\top \phi(\mathbf{x})}}{1+e^{\theta^\top \phi(\mathbf{x})}} \in [0, 1]$ as a model, with p being the probability of the individual x being classified as 1. The training process consists of finding the best value for θ to minimize the classification error for the data in training. This approach lacks a fairness strategy, focusing only on maximizing performance.

Reweighting [60] (Reweight): It modifies the logistic regression to consider different weights for individuals. It separates individuals into groups according to the sensitive attribute and the outcome label — *e.g.*, if it is a binary classification and the sensitive feature is also binary, then we have four strata. For each stratum, the weight corresponds to a ratio between the expected probability of the strata (in a world without prejudice) and the actual probability presented in the data set. Thus, this strategy weights the optimization to find a θ parameter of the sigmoid function that compensates each stratum following the prejudice error.

Demographic Parity Classifier [56] (DemPar): It rewrites the logistic regression optimization problem to ensure the discrimination of the model does not exceed a specific value. As the name implies, it uses Demographic Parity as a fairness metric. However, due to the metric not being convex and other characteristics, the classifier does not use it as a constraint but adapts it in a condition that indirectly limits discrimination.

Equality of Opportunity Classifier [23] (EqOp): It is similar to the Demographic Parity Classifier, but its restriction is based on the Equality of Opportunity metric.

Minimax Pareto Fairness [38] (Minimax): It models the training as a multi-objective optimization problem in which the objective functions are the model error for the groups according to the sensitive attribute. Finding a single model seeks to minimize the greatest among the mistakes of the groups.

Adafair [43] (AdaFair): It is a boosting method based on AdaBoost where at each boosting round, the weights of the instances are updated considering the Fairness of the current ensemble members. The Fairness measure used is the Equalized Odds [25].

MAMOFair [35] (MAMOFair): It is a model-agnostic multi-objective approach based on a gradient that uses the binary cross-entropy as the performance objective and a hyperbolic tangent relaxation of a Fairness notion as a second objective. The Fairness notion must require a form of statistical parity across groups. In their experiments, the fairness metrics used are “Difference of Demographic Parity” and “Difference of Equality of Opportunity,” based on the same fairness notions used in this study.

The proposed approaches consist of selecting one of two proposed multi-objective models (described in Sections 4.1 and 4.2) and using MONISE to generate a set of Pareto-optimal models that will be selected and aggregated into an ensemble. MOOError and MOOAcep are the respective labels for methods using Error per group (described in Section 4.1) and Acceptance by group (described in Section 4.2) formulations.

6.1.4 Evaluation metrics

This study uses the well-known accuracy and a set of well-known discrimination metrics: Equal Opportunity [25], Demographic Parity [56], and Coefficient of Variation [9]. We give a quick explanation and justification of those metrics below:

- **Accuracy** (labeled as Acc): it measures the ratio of correctly classified samples.
- **Equal Opportunity** (labeled as EO): it measures (from 0 to 1) how close are the predictions from achieving the same probability of correctly classifying a beneficial outcome $P(f_\theta(X) = 1|Y = 1)$ for both groups $A = 0$ and $A = 1$ (if $P(f_\theta(X) = 1|A = 0, Y = 1) = P(f_\theta(X) = 1|A = 1, Y = 1)$, then $EO = 1$).
- **Demographic Parity** (labeled as DP): it measures (from 0 to 1) how close are the predictions from achieving the same probability of correctly classifying a beneficial outcome $P(f_\theta(X) = 1)$ for both groups $A = 0$ and $A = 1$ (if $P(f_\theta(X) = 1|A = 0) = P(f_\theta(X) = 1|A = 1)$, then $DP = 1$).
- **Coefficient of Variation** (labeled as CV): measures how much the beneficial outcome for each individual deviates from the mean beneficial outcome achieved by all individuals.

We select those metrics to evaluate different aspects of a fair model. Accuracy measures the overall capacity to make correct classifications. Equal opportunity and demographic parity measure how far the classifier can make a similar classification for any group on two aspects: correctly classifying the beneficial outcome and achieving the beneficial outcome. Finally, the coefficient of variation measures how much the classifier deviates from delivering the beneficial outcome for all individuals.

6.2 Pareto-frontier experiment

In our first experiment, we evaluate how the models generated by our approach behave under the proposed sets of objective functions (loss by group and acceptance by group), compared with competing approaches that aim for the same Fairness definition. The goal is to verify if the generated models are non-dominated and more diverse than the compared approaches. We also want to analyze how the Pareto-frontier behaves in different subsets of data.

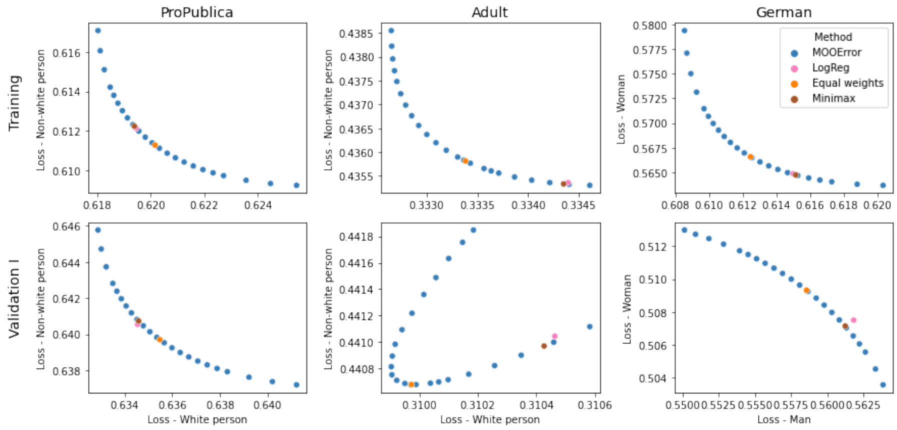


Fig. 5: Models generated by the error-based multi-objective approach (MOOErr) compared with models generated by other approaches based on loss: standard Logistic Regression (LogReg), Minimax Pareto Fairness (Minimax), and a Logistic Regression with weights on groups to balance the amount of data per group (Equal weights). The axes are the log loss for each group based on the protected feature, and the color indicates the approach. The column of plots indicates the dataset (ProPublica, Adult, and German), and the rows indicate if the metrics were evaluated on training (top row) or validation I data (bottom row).

For the approach that optimizes loss by group, we generated 20 models, evaluated them in the Training and Validation I datasets, and compared them with the standard Logistic Regression (LogReg) — with weights on groups to balance the amount of data per group (Equal weights) — and Minimax. Columns show the results for each dataset in Figure 5. Each point in the figure represents a different model, with the color indicating its approach. The compared models were in or near the Pareto frontier for all datasets. Notably, the Minimax model did not achieve the minimax point in the Training set. For instance, in the ProPublica dataset (first column of Figure 5), there are Pareto models with a lower loss for White individuals in the Training set, which is the group with the maximum loss. This outcome is likely due to Minimax's use of a validation dataset during training, working with a conflict between training and validation minimax. Moreover, in the ProPublica dataset example, the non-White category is the group with maximum loss in the Validation I dataset. Therefore the minimax model can not decrease the loss for White individuals because it would increase the loss for non-White individuals.

Regarding the results obtained for the Adult and German datasets in Training and Validation I, note that some models not dominated in training become dominated in validation. We got the same results for both the proposed and compared approaches. This effect indicates that the Pareto frontier has different generalizations in learning in each group for each dataset. For instance, ProPublica seems to have equal loss performance in each group. Adult generalizes poorly when learning is focused on a specific group (generating dominated solutions). Finally, the German dataset has a slight generalization problem for more balanced learning (between groups) but not enough to create useless (dominated) models. We believe this is an exciting point, requiring a theoretical and deep analysis in independent research. Nonetheless, the diversity of the approach proposed guarantees to find a set of solutions that remains non-dominated (and dominates some of the compared techniques). It motivates the implementation of a Non-Dominated Sorting filter for the model aggregation (explored in Section 6.4) to ensure that the selected models are the best in training and validation.

For the acceptance-based approach, we generated 150 models. We compared them with other 150 models generated by the regularization and constraint parameters of the Demographic Parity Classifier (DemPar) — both approaches that aim to ensure Demographic Parity [56]. Figure 6 shows the results, in which the models generated by MOOAcep are represented with circles and the ones generated by DemPar by triangles. The horizontal axis of each plot indicates the acceptance loss achieved by one group and the vertical axis by the other. The color of each point indicates its loss considering all individuals. Also, the dashed line represents the region where the acceptance is the same for both groups, so the distance of a model to the line indicates its Demographic Parity; the more distant, the more unfair the model.

Analyzing and comparing the models, we have observed that the DemPar models are significantly less diverse both in acceptance by the group and in classification loss. However, both approaches have the same quantity of models; the MOOAcep models are significantly more dispersed than the first ones. With the MOOAcep models, the trade-off among minimum acceptance loss, and classification loss, as well as among the groups' acceptance; the closer to the axis, the brighter the model's point is, which indicates a more significant loss. The MOOAcep models achieve significantly lower acceptance losses than the DemPar models, which means that, given the same loss, the first ones predict more beneficial outputs than the second ones. Also, MOOAcep models with both high and low classification losses, which motivates the filtering by accuracy before aggregating the models, will be explored in the following experiments.

Worth mentioning that even when the multi-objective method gives preference to discriminated groups, the optimization does not find any model with more acceptance for discriminated (non-White and woman) than non discriminated (White and man) groups. It shows that, when such preference is given, the acceptance becomes equal, which may increase the loss. This relevant qualitative result shows how much the dataset is biased towards the non-discriminated group.

This experiment suggests that the multi-objective approaches can generate set Pareto-optimal models that are highly diverse for the proposed objective functions. However, this diversity might come with a subset of models with higher loss than the compared approaches, as shown in Figure 6. The experi-

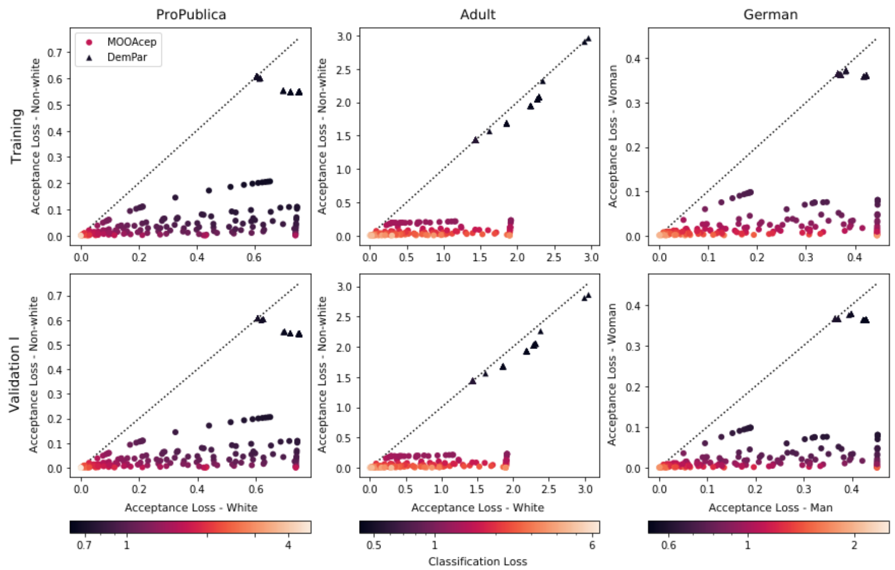
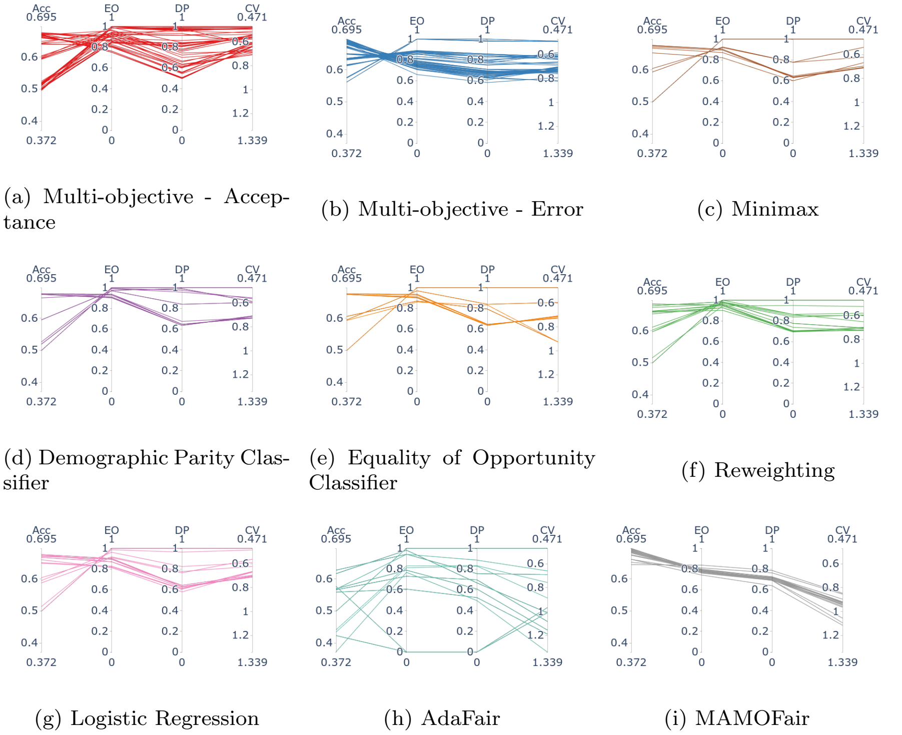


Fig. 6: Models generated by the acceptance-based multi-objective approach (circles) compared with models generated by the Demographic Parity Classifier (triangles). The axes are the acceptance loss for each group based on the protected feature, and the color indicates the model loss for all individuals. The column of plots indicates the dataset (ProPublica, Adult, and German), and the rows indicate if the metrics were evaluated on training (top row) or validation I data (bottom row). The line represents the region where the acceptance loss is the same for both groups.

ment also shows the importance of making reasonable filtering procedures to discard such models. In the following experiments, we will evaluate if the models are diverse for performance and Fairness metrics beyond the proposed loss and acceptance metrics. We will also study how to select the Pareto-optimal models to achieve a better model through aggregation.

6.3 Diversity experiment

In this experiment, we explored the ability of the proposed and baseline methods to find models with different metric profiles for the same data set. Such capability is essential for two main reasons. First, it allows the user to choose between the performance and fairness metrics better suited for their application. And second, it enables ensemble learning techniques to generate, from the diverse set of models, a single model that is more robust than the previous ones. For the proposed methodologies and MAMOFair, we analyzed the diversity of models generated by the multi-objective optimization. As for the baseline methodologies, we created sets of models by varying their hyperparameters: for Logistic Regression, Reweighting, and Minimax, we varied the penalization hyperparameter C; and for Demographic Parity and Equality of Opportunity Classifiers, we varied both the penalization hyperparameter C and the Fairness constraint hyperparameter. Lastly, for AdaFair, we analyzed the weak classifiers.



(a) Multi-objective - Acceptance

(b) Multi-objective - Error

(c) Minimax

(d) Demographic Parity Classifier

(e) Equality of Opportunity Classifier

(f) Reweighting

(g) Logistic Regression

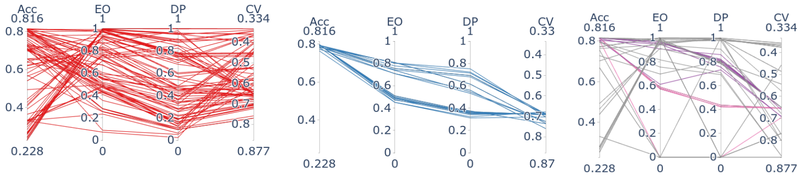
(h) AdaFair

(i) MAMOFair

Fig. 7: Models found by each strategy for the ProPublica dataset. Each plot represents a strategy, and each line represents one of the 150 models generated by that strategy. The points where the lines meet the axes indicate the value found for each of the metrics Accuracy (Acc), Equal Opportunity (EO), Demographic Parity (DP), and Coefficient of Variation (CV).

Figures 7, 8, and 9 show the models obtained using the ProPublica, German and Adult datasets. Each plot contains the 150 models generated by a single strategy (either proposed or compared). Each line represents a model, and the axes (Acc, EO, DP, CV) indicate the values obtained for the metrics Accuracy, Equal Opportunity, Demographic Parity, and Coefficient of Variation. The axis of the Coefficient of the Variation metric was inverted because, unlike the other metrics, we seek to minimize it. We evaluate each metric on validation data.

By analyzing the results obtained for the ProPublica dataset in Figure 7, we noted that multi-objective strategies ((a) and (b)) visually present a significantly larger number of lines w.r.t. other strategies, despite all having 150 models. The reason resides in the baselines generating models with practically equal values for all metrics, thus overlapping the lines. This difference indicates a greater diversity of models created by multi-objective strategies, allowing more flexible and robust results when applying ensemble learning strategies. Although the compared strategies could only detect models with higher accuracy and lower fairness or models with lower accuracy and higher fairness, the proposed strategies generated intermediate models, giving the decision-maker more options to choose from when using filtering/selection strategies.



(a) Multi-objective - Acceptance

(b) Multi-objective - Error

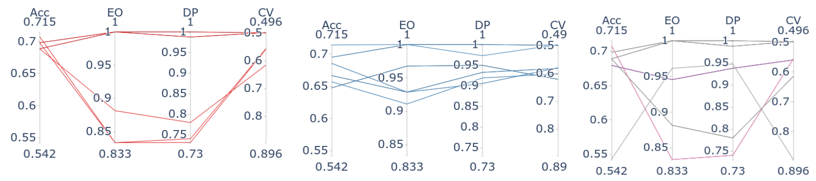
(c) Others

Fig. 8: Models found by each strategy for the Adult dataset. The first two plots present the results obtained by the proposed strategies, while the third presents the results for the compared strategies. Each line represents one of the 150 models generated by this strategy, and the points where they meet the axes indicate the value found for each of the metrics of Accuracy (Acc), Equal Opportunity (EO), Demographic Parity (DP) and Coefficient of Variation (CV).

For the Adult dataset (see Figure 8), the models generated by the multi-objective error-based strategy showed similar values or low values for accuracy and fairness compared to the other techniques (shown aggregated in (c)), having not considerably gained in diversity. However, the multi-objective method based on acceptance generated was more diverse, including models with better demographic parity and coefficient of variation than one model generated by other strategies.

Finally, for the German dataset (see Figure 9), the multi-objective strategies obtained a slightly greater variety of models than the other strategies, and the best models are better or at least comparable with the baselines.

These results indicate that our proposal is more diverse and more capable of finding models with higher Accuracy (Figure 7), Demographic Parity (Figures 7 and 8), and lower Coefficient of Variation (Figure 8) than other formulations. These qualities are essential for building good ensembles. Also, for the contenders, in Adult and ProPublica datasets, at least one of those



(a) Multi-objective - Acceptance

(b) Multi-objective - Error

(c) Others

Fig. 9: Models found by each strategy for the German dataset. The first two plots present the results obtained by the proposed strategies, while the third presents the results for the compared strategies. Each line represents one of the 150 models generated by this strategy, and the points where they meet the axes indicate the value found for each of the metrics of Accuracy (Acc), Equal Opportunity (EO), Demographic Parity (DP), and Coefficient of Variation (CV).

diversity-focused methods is diverse. Thus this problem is directly related to the lack of awareness of this feature by other methods.

6.4 Aggregation Experiment

In this experiment, we test different ways of selecting the models from the previous experiment to be aggregated using the process described in Section 5. First, we try filtering the models based on accuracy and then selecting the 10 best ones in a particular fairness metric evaluated in the Validation I set. After filtering, the models are aggregated through ensemble, the resulting model is evaluated in Validation II set. The results obtained for the ProPublica dataset are shown in Figure 10, with the first row containing the models generated through the error-based approach and the second row the acceptance-based approach. The horizontal axis indicates which procedure was used to filter the models: *All models* was generated without any filtering; *50Acc+DP* selected the 50 best models in Accuracy and then the 10 best in Demographic Parity; *20Acc+DP* selected the 20 best in Accuracy and 10 best in Demographic Parity; the other filters are named in the same pattern for different metrics: Equal Opportunity (EO) and Coefficient of Variation (CV).

The error-based aggregated models did not change much with the different filters. On the other hand, the acceptance-based approach presented very differently aggregated models, with Accuracy increasing and Fairness decreasing when selecting fewer and better models on Accuracy. They also varied with the Fairness metric used as the second filter; the models aggregated using CV were better in CV, and (*50Acc+CV*) were better also in DP than the other aggregated models. The models aggregated using DP had better or similar Fairness than the ones aggregated using EO. The most restricted filters

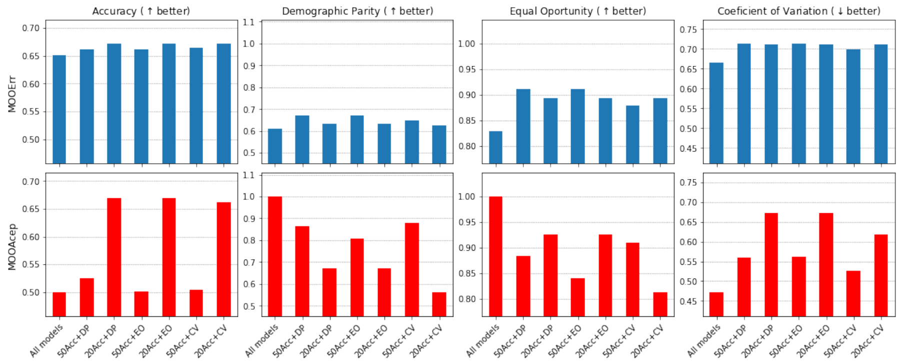


Fig. 10: Aggregation experiment selecting the models through single Fairness metrics for ProPublica dataset. The first row shows the models obtained with the error-based approach, and the second one obtained with the acceptance-based approach. Each column of plots shows the values obtained for a metric by the aggregated models using different selecting options.

showed the best values for Accuracy, but the acceptance-based models showed a significant increase in Fairness compared to the error-based ones.

Next, we experiment with sorting the models filtered by Accuracy through Non-Dominated Sorting using multiple metrics as objectives to select the 10 best models. We compare performing the sort with and without Accuracy as objective; the filters that are sorted with Accuracy are named *NDS(wAcc)* and the ones without *NDS*. Figure 11 presents the results for the ProPublica dataset.

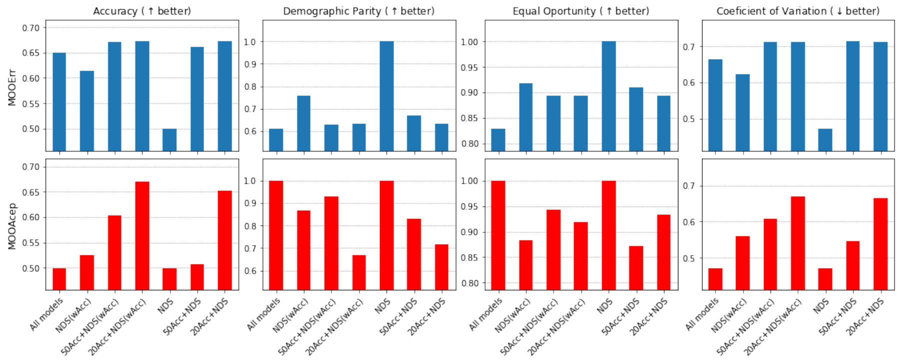


Fig. 11: Aggregation experiment selecting the models through Non-Dominated Sorting for the ProPublica dataset. The first row shows the models obtained with the error-based approach, and the second one obtained with the acceptance-based approach. Each column of plots shows the values obtained for a metric by the aggregated models using different selecting options.

The error-based models filtered by Non-Dominated sorting varied significantly more than when filtered by single metrics, especially when filtered only by sorting. When Accuracy is not considered for the selection (filter *NDS*), the resulting model improves on the Fairness metrics but decreases in Accuracy compared to the other filters' resulting models. Also, when Accuracy is considered during sorting, but without a previous filter over it (filter *NDS(wAcc)*), the resulting model improves on Fairness without decreasing so much in Accuracy.

For the acceptance-based models, sorting with Accuracy resulted in models with higher Accuracy and similar Fairness compared to the ones sorted without it. The filter *50Acc+NDS(wAcc)* resulted in a small decrease in Accuracy when compared to the most restrict models (*20Acc+NDS(wAcc)* and *20Acc+NDS*) and a significantly improve in Fairness, being the most balanced filter for the acceptance-based models. By comparing the filters based on single metrics and the ones with Non-Dominated Sorting with the acceptance-based models, we observed that the NDS filters with Accuracy can increase Fairness as much as the first type of filter, but with a smaller decrease in Accuracy; the NDS filters without Accuracy showed similar behavior to the first type of filters.

In this experiment, we could see that depending on the number of models selected using accuracy followed by a selection using one or multiple fairness metrics (with or without accuracy), we can find different metric patterns sacrificing more or less accuracy to achieve more or less fairness. The decision-maker can carefully use it to search for the desired profile of accuracy and fairness metrics. More importantly, it is possible to improve both Fairness and Accuracy through the selection of models to be aggregated when compared to aggregating all models; improving Fairness through the model selection enhances all chosen Fairness metrics simultaneously; filtering the models through Non-Dominated Sorting results in more balanced models than filtering based on individual metrics. In the next experiment, we evaluate how the aggregated models perform compared to other approaches.

6.5 Comparative Experiment

In this experiment, we compare the aggregated models with other approaches. Given the results of the previous experiment, we compared the error-based models filtered by Non-Dominated Sorting without Accuracy filter (*NDS (wAcc)*) and the acceptance-based models selected with Non-Dominated Sorting with Accuracy of the 50 best models in Accuracy (*50Acc+NDS(wAcc)*) with the competing approaches. We selected this ensemble profile because we want to increase fairness without sacrificing too much accuracy, and the previous experiment for the ProPublica dataset showed that. The compared models were trained with regularization, and their proposed standard parameters ($C = 1$ for LogReg, Reweight, DemPar, and EqOp, and $C = 1e^7$ for Minimax), and the constraint approaches were trained with the most strict constraint (covariance threshold of 0 for DemPar and EqOp). The results presented were obtained through a k-fold Cross-Validation with 5 folds.

Table 1 shows the results obtained for the ProPublica dataset. The proposed models were able to increase Fairness in comparison to the standard model (LogReg) without decreasing as much in Accuracy as the restriction-based ones (DemPar and EqOp). Also, they have the lowest Coefficient of Variation among all tested models.

Approach	Acc	EO	DP	CV
LogReg	0.658	0.788	0.664	0.735
Reweigh	0.643	0.938	0.887	0.764
DemPar	0.483	0.932	0.920	0.677
EqOp	0.589	0.950	0.870	0.952
Minimax	0.658	0.787	0.663	0.735
MAMOFair	0.643	0.852	0.745	0.843
AdaFair	0.612	0.985	0.976	0.662
MooAcep (ours)	0.505	0.992	0.973	0.479
MooErr (ours)	0.615	0.856	0.770	0.645

Table 1: Comparative experiment for the ProPublica dataset

The proposed models were the best at all Fairness metrics for the German dataset (Table 2). They decreased Accuracy compared to the standard Logistic Regression and Minimax but reported a higher Accuracy than Reweigh and EqOp. AdaFair only found trivial models for this dataset, so we did not include its results.

Approach	Acc	EO	DP	CV
LogReg	0.728	0.975	0.914	0.528
Reweigh	0.708	0.966	0.957	0.534
DemPar	0.713	0.994	0.987	0.508
EqOp	0.702	0.986	0.979	0.544
Minimax	0.728	0.978	0.917	0.531
MAMOFair	0.713	0.989	0.993	0.516
AdaFair*	-	-	-	-
MooAcep (ours)	0.722	0.982	0.942	0.506
MooErr (ours)	0.722	0.986	0.942	0.517

Table 2: Comparative experiment for the German dataset

Finally, for the Adult dataset (Table 3), the acceptance-based model (MooAcep) was able to increase both Demographic Parity and Equal Opportunity while decreasing the Coefficient of Variation. On the other hand, the error-based model shows a decrease in Demographic Parity and an increase in Coefficient of Variation. It is noteworthy that the filtering methods of the proposed approaches were not chosen considering the validation performance for this dataset, which may explain the obtained results.

Approach	Acc	EO	DP	CV
LogReg	0.802	0.706	0.434	0.679
Reweigh	0.801	0.940	0.722	0.679
DemPar	0.802	0.970	0.686	0.678
EqOp	0.801	0.993	0.668	0.674
Minimax	0.802	0.710	0.431	0.679
MAMOFair	0.802	0.959	0.681	0.869
AdaFair	0.787	0.925	0.772	0.658
MooAcep (ours)	0.793	0.908	0.648	0.663
MooErr (ours)	0.780	0.726	0.204	0.743

Table 3: Comparative experiment for the Adult dataset

The obtained results imply that we can improve Fairness with a smaller decrease in Accuracy compared to other approaches. Together with the results from the previous experiment, we could validate that our models were flexible enough to adapt to the decision-maker's need (in this case, us) to find the model with a specific trade-off between accuracy and fairness metrics. In other situations, it would be possible to adapt the parameters (such as the number of models selected using accuracy) to change the trade-off profile without needing try-and-error parameters in training — our approach creates diverse models to achieve the desired goal of decision-makers.

Finally, the proposed strategy can systematically decrease the Coefficient of Variation observed for all analyzed datasets. This result is interesting because the Coefficient of Variation is an equity metric based on individual rather than group outcomes. Some studies argue that equity-based on groups and individuals is conflicting, meaning that improving the balance between groups introduces unfairness between individuals [17,61]. However, other studies argue that injustice can be decomposed into a group and individual level (that can be improved [9]), and this apparent conflict is based on a philosophical misconception [11].

7 Discussion

This research achieves a milestone in creating multiple fairness-compatible models through a good representation of trade-offs of fairness and accuracy-related losses. In what follows, we discuss the characteristics of the proposed method.

High quality and diverse models. The Diversity Experiment shows the proposed models varying on a wide range in all metrics but always having models with good performance in some metrics. It happens because we offer flexible formulations that had their trade-offs correctly sampled, guaranteeing to find good models.

Decision-maker empowerment. The flexibility of the formulations combined with a good Pareto representation promoted by MONISE finds models to a vast majority of accuracy vs. fairness trade-offs. Consequently, the proposed method allows the decision-maker to explicitly choose the trade-off between accuracy and fairness metrics, thus combining a set of models to create an ensemble.

Focus on two groups on binary classification. This first attempt at exploring MONISE on fairness models shows an initial success. The following steps include (1) exploring multiple groups and (2) other machine learning problems such as multiclass classification and regression. Both approaches are feasible in the current framework but depend on research to see the impact of (1) many groups and (2) how to adapt the formulations to new problems.

Generality. The proposed framework, when using MONISE, is constrained to use convex objective functions. It limits the approach to simple models, such as Linear, Logistic, Multinomial regression, and Linear SVMs, and the losses (demographic parity and equal opportunity are non-convex functions). However, the concepts are still generalizable; we advocate that modeling the fairness problem as multi-objective and finding a suitable representation will discover helpful and diverse models that will generate good ensembles. Expanding this methodology to non-convex scenarios is a relevant challenge.

8 Conclusion

The conflict between accuracy and fairness metrics is a critical challenge in the fairness field. Any decision-maker must decide their preference and select a trade-off among the desired goals. This study tackles this problem using multi-objective optimization concepts to aggregate diverse models to achieve predictions with increased fairness without sacrificing too much accuracy. Our framework achieved this goal using three steps: (1) Modeling the classification as a multi-objective problem in two different ways; (2) Creating Pareto-front representation using MONISE; (3) Using accuracy and fairness metrics under multi-objective selection methods (Non-Dominate Sorting and hypervolume) to filter the models to be aggregated by the ensemble.

The two different models optimized the classification loss and the acceptance loss (how distant the classification is to achieve the desired outcome) per group instead of minimizing the difference between the groups (usually associated with a fairness objective). This proxy approach showed many advantages: (1) the flexibility of the models does not force a solution with equal performance per group; if a model has better performance for all groups, it finds the latter; (2) using MONISE generates diverse Pareto-optimal classifiers, which can be helpful because the performance in validation can change; and (3) this diversity allows the decision-maker to achieve their preference by adjusting the ensemble filtering and aggregation. In the experiments, we showed that it was possible to find models that did not sacrifice too much accuracy, but we could improve the fairness metrics significantly.

With this research, we showed the importance of approaching the fairness problem using multi-objective methodologies. The flexibility of the models combined with the diversity promoted by finding multiple Pareto-optimal models using MONISE creates diverse models that allow the decision-maker to create a final predictor aligned to their preferences. This strategy might be helpful to achieve competent models for other fairness goals (different metrics) with adapted new models, and it might be beneficial to other problems; however, the models need to be adjusted to find other proxy multi-objective models.

Declarations

This study was supported by Getulio Vargas Foundation and by the Coordination for the Improvement of Higher Education Personnel (CAPES). The authors thank Espaço da Escrita – Pró-Reitoria de Pesquisa – UNICAMP – for the language services provided.

References

1. Dastin, J.: Amazon scraps secret ai recruiting tool that showed bias against women (2018). URL https://www.reuters.com/article/us-amazon-com-jobs-automation-insight-idUSKCN1MK08G
2. Julia Angwin Jeff Larson, S.M., Kirchner, L.: Machine bias (2016). URL https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing
3. Dieterich, W., Mendoza, C., Brennan, T.: Compas risk scales: Demonstrating accuracy equity and predictive parity. Northpoint Inc **7**(7.4), 1 (2016)
4. Dressel, J., Farid, H.: The accuracy, fairness, and limits of predicting recidivism. Science advances **4**(1), eaao5580 (2018)
5. Howard, A., Borenstein, J.: The ugly truth about ourselves and our robot creations: the problem of bias and social inequity. Science and engineering ethics **24**(5), 1521–1536 (2018)
6. Osoba, O.A., Welser IV, W.: An intelligence in our image: The risks of bias and errors in artificial intelligence. Rand Corporation (2017)
7. Zhao, H., Gordon, G.: Inherent tradeoffs in learning fair representations. Advances in neural information processing systems **32** (2019)
8. Zliobaite, I.: On the relation between accuracy and fairness in binary classification (2015)
9. Speicher, T., Heidari, H., Grgic-Hlaca, N., Gummadi, K.P., Singla, A., Weller, A., Zafar, M.B.: A unified approach to quantifying algorithmic unfairness: Measuring individual & group unfairness via inequality indices. In: Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, KDD '18, p. 2239–2248. Association for Computing Machinery (2018)
10. Dutta, S., Wei, D., Yueksel, H., Chen, P.Y., Liu, S., Varshney, K.: Is there a trade-off between fairness and accuracy? a perspective using mismatched hypothesis testing. In: International Conference on Machine Learning, pp. 2803–2813. PMLR (2020)
11. Binns, R.: On the apparent conflict between individual and group fairness. In: Proceedings of the 2020 conference on fairness, accountability, and transparency, pp. 514–524 (2020)
12. Julia Angwin Jeff Larson, S.M., Kirchner, L.: Compas recidivism risk score data and analysis (2016). URL https://github.com/propublica/compas-analysis/
13. Raimundo, M.M., Von Zuben, F.J.: Multi-criteria analysis involving pareto-optimal misclassification tradeoffs on imbalanced datasets. In: 2020 International Joint Conference on Neural Networks (IJCNN), pp. 1–8. IEEE (2020)

14. Raimundo, M.M., Von Zuben, F.J.: Investigating multiobjective methods in multitask classification. In: 2018 International Joint Conference on Neural Networks (IJCNN), pp. 1–9 (2018). DOI 10.1109/IJCNN.2018.8489333
15. Raimundo, M.M., Drumond, T.F., Marques, A.C.R., Lyra, C., Rocha, A., Von Zuben, F.J.: Exploring multiobjective training in multiclass classification. Neurocomputing **435**, 307–320 (2021)
16. Kusner, M.J., Loftus, J., Russell, C., Silva, R.: Counterfactual fairness. Advances in neural information processing systems **30** (2017)
17. Dwork, C., Hardt, M., Pitassi, T., Reingold, O., Zemel, R.: Fairness through awareness. In: Proceedings of the 3rd innovations in theoretical computer science conference, pp. 214–226 (2012)
18. Grgic-Hlaca, N., Zafar, M.B., Gummadi, K.P., Weller, A.: The case for process fairness in learning: Feature selection for fair decision making. In: NIPS Symposium on Machine Learning and the Law, vol. 1, p. 2 (2016)
19. Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., Galstyan, A.: A survey on bias and fairness in machine learning. ACM Computing Surveys (CSUR) **54**(6), 1–35 (2021)
20. d'Alessandro, B., O'Neil, C., LaGatta, T.: Conscientious classification: A data scientist's guide to discrimination-aware classification. Big data **5**(2), 120–134 (2017)
21. Zemel, R., Wu, Y., Swersky, K., Pitassi, T., Dwork, C.: Learning fair representations. In: International conference on machine learning, pp. 325–333. PMLR (2013)
22. Kamiran, F., Calders, T.: Data preprocessing techniques for classification without discrimination. Knowledge and Information Systems **33**(1), 1–33 (2012)
23. Zafar, M.B., Valera, I., Rogriguez, M.G., Gummadi, K.P.: Fairness constraints: Mechanisms for fair classification. In: Artificial Intelligence and Statistics, pp. 962–970. PMLR (2017)
24. Agarwal, A., Beygelzimer, A., Dudik, M., Langford, J., Wallach, H.: A reductions approach to fair classification. In: J. Dy, A. Krause (eds.) Proceedings of the 35th International Conference on Machine Learning, *Proceedings of Machine Learning Research*, vol. 80, pp. 60–69. PMLR (2018). URL http://proceedings.mlr.press/v80/agarwal18a.html
25. Hardt, M., Price, E., Srebro, N.: Equality of opportunity in supervised learning. Advances in neural information processing systems **29** (2016)
26. Wadsworth, C., Vera, F., Piech, C.: Achieving fairness through adversarial learning: an application to recidivism prediction (2018)
27. Kamishima, T., Akaho, S., Sakuma, J.: Fairness-aware learning through regularization approach. In: 2011 IEEE 11th International Conference on Data Mining Workshops, pp. 643–650. IEEE (2011)
28. Kamishima, T., Akaho, S., Asoh, H., Sakuma, J.: Fairness-aware classifier with prejudice remover regularizer. In: Joint European Conference on Machine Learning and Knowledge Discovery in Databases, pp. 35–50. Springer (2012)
29. Berk, R., Heidari, H., Jabbari, S., Joseph, M., Kearns, M., Morgenstern, J., Neel, S., Roth, A.: A convex framework for fair regression. arXiv preprint arXiv:1706.02409 (2017)
30. Zafar, M.B., Valera, I., Rodriguez, M., Gummadi, K., Weller, A.: From parity to preference-based notions of fairness in classification. Advances in Neural Information Processing Systems **30** (2017)
31. Kearns, M., Roth, A.: The ethical algorithm: The science of socially aware algorithm design. Oxford University Press (2019)
32. Savic, D.: Single-objective vs. multiobjective optimisation for integrated decision support. Proceedings of the First Biennial Meeting of the International Environmental Modelling and Software Society **1**, 7–12 (2002)
33. Cruz, A.F., Saleiro, P., Belém, C., Soares, C., Bizarro, P.: A bandit-based algorithm for fairness-aware hyperparameter optimization. arXiv preprint arXiv:2010.03665 (2020)
34. Liu, S., Vicente, L.N.: Accuracy and fairness trade-offs in machine learning: A stochastic multi-objective approach. Computational Management Science pp. 1–25 (2022)
35. Padh, K., Antognini, D., Lejal-Glaude, E., Faltings, B., Musat, C.: Addressing fairness in classification with a model-agnostic multi-objective algorithm. In: Uncertainty in Artificial Intelligence, pp. 600–609. PMLR (2021)

36. Zhang, Q., Liu, J., Zhang, Z., Wen, J., Mao, B., Yao, X.: Fairer machine learning through multi-objective evolutionary learning. In: International Conference on Artificial Neural Networks, pp. 111–123. Springer (2021)
37. Zhang, Q., Liu, J., Zhang, Z., Wen, J., Mao, B., Yao, X.: Mitigating unfairness via evolutionary multi-objective ensemble learning. IEEE Transactions on Evolutionary Computation (2022)
38. Martinez, N., Bertran, M., Sapiro, G.: Minimax pareto fairness: A multi objective perspective. In: H.D. III, A. Singh (eds.) Proceedings of the 37th International Conference on Machine Learning, *Proceedings of Machine Learning Research*, vol. 119, pp. 6755–6764. PMLR (2020). URL http://proceedings.mlr.press/v119/martinez20a.html
39. Grgic-Hlaca, N., Zafar, M.B., Gummadi, K.P., Weller, A.: On fairness, diversity and randomness in algorithmic decision making. CoRR (2017)
40. Kenfack, P.J., Khan, A.M., Kazmi, S.A., Hussain, R., Oracevic, A., Khattak, A.M.: Impact of model ensemble on the fairness of classifiers in machine learning. In: 2021 International Conference on Applied Artificial Intelligence (ICAPAI), pp. 1–6. IEEE (2021)
41. Bhargava, V., Couceiro, M., Napoli, A.: Limeout: an ensemble approach to improve process fairness. In: Joint European Conference on Machine Learning and Knowledge Discovery in Databases, pp. 475–491. Springer (2020)
42. Iosifidis, V., Fetahu, B., Ntoutsi, E.: Fae: A fairnesskusner2017counterfactual-aware ensemble framework. In: 2019 IEEE International Conference on Big Data (Big Data), pp. 1375–1380 (2019). DOI 10.1109/BigData47090.2019.9006487
43. Iosifidis, V., Ntoutsi, E.: Adafair: Cumulative fairness adaptive boosting. In: Proceedings of the 28th ACM International Conference on Information and Knowledge Management, pp. 781–790 (2019)
44. Bhaskaruni, D., Hu, H., Lan, C.: Improving prediction fairness via model ensemble. In: 2019 IEEE 31st International Conference on Tools with Artificial Intelligence (ICTAI), pp. 1810–1814. IEEE (2019)
45. Zhang, W., Bifet, A., Zhang, X., Weiss, J.C., Nejdl, W.: Farf: A fair and adaptive random forests classifier. In: Pacific-Asia Conference on Knowledge Discovery and Data Mining, pp. 245–256. Springer (2021)
46. Zhang, W., Weiss, J.C.: Fair decision-making under uncertainty. In: 2021 IEEE International Conference on Data Mining (ICDM), pp. 886–895. IEEE (2021)
47. Zhang, W., Weiss, J.C.: Longitudinal fairness with censorship. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 36, pp. 12235–12243 (2022)
48. Abebe, S.A., Lucchese, C., Orlando, S.: Eifffel: enforcing fairness in forests by flipping leaves. In: Proceedings of the 37th ACM/SIGAPP Symposium on Applied Computing, pp. 429–436 (2022)
49. Chen, Z., Zhang, J., Sarro, F., Harman, M.: Maat: A novel ensemble approach to addressing fairness and performance bugs for machine learning software. In: The ACM Joint European Software Engineering Conference and Symposium on the Foundations of Software Engineering (ESEC/FSE) (2022)
50. Miettinen, K.: Nonlinear multiobjective optimization, vol. 12. Springer Science & Business Media (2012)
51. Cohon, J.L.: Multiobjective programming and planning, vol. 140. Courier Corporation (2004)
52. Cohon, J.L., Church, R.L., Sheer, D.P.: Generating multiobjective trade-offs: An algorithm for bicriterion problems. Water Resources Research **15**(5), 1001–1010 (1979)
53. Raimundo, M.M., Ferreira, P.A., Von Zuben, F.J.: An extension of the non-inferior set estimation algorithm for many objectives. European Journal of Operational Research **284**(1), 53–66 (2020). DOI https://doi.org/10.1016/j.ejor.2019.11.017. URL https://www.sciencedirect.com/science/article/pii/S0377221719309282
54. Rokach, L.: Ensemble-based classifiers. Artificial intelligence review **33**(1), 1–39 (2010)
55. Beume, N., Naujoks, B., Emmerich, M.: Sms-emoa: Multiobjective selection based on dominated hypervolume. European Journal of Operational Research **181**(3), 1653–1669 (2007). DOI https://doi.org/10.1016/j.ejor.2006.08.008. URL https://www.sciencedirect.com/science/article/pii/S0377221706005443
56. Zafar, M.B., Valera, I., Rodriguez, M.G., Gummadi, K.P., Weller, A.: From parity to preference-based notions of fairness in classification (2017)

57. Abdi, H.: Coefficient of variation. Encyclopedia of research design **1**, 169–171 (2010)
58. Bellamy, R.K., Dey, K., Hind, M., Hoffman, S.C., Houde, S., Kannan, K., Lohia, P., Martino, J., Mehta, S., Mojsilović, A., et al.: Ai fairness 360: An extensible toolkit for detecting and mitigating algorithmic bias. IBM Journal of Research and Development **63**(4/5), 4–1 (2019)
59. Dua, D., Graff, C.: UCI machine learning repository (2017). URL http://archive.ics.uci.edu/ml
60. Calders, T., Kamiran, F., Pechenizkiy, M.: Building classifiers with independency constraints. In: 2009 IEEE International Conference on Data Mining Workshops, pp. 13–18. IEEE (2009)
61. Corbett-Davies, S., Goel, S.: The measure and mismeasure of fairness: A critical review of fair machine learning. arXiv preprint arXiv:1808.00023 (2018)