



# Interpretable and Fair Mechanisms for Abstaining Classifiers

Daphne Lenders<sup>1,2(✉)</sup>, Andrea Pugnana<sup>3</sup>, Roberto Pellungrini<sup>4</sup>,  
Toon Calders<sup>1,2</sup>, Dino Pedreschi<sup>3</sup>, and Fosca Giannotti<sup>4</sup>

<sup>1</sup> Adrem Data Lab, University of Antwerp, Antwerp, Belgium  
[daphne.lenders@uantwerpen.be](mailto:daphne.lenders@uantwerpen.be)

<sup>2</sup> DigiTax, University of Antwerp, Antwerp, Belgium

<sup>3</sup> KDD Lab, University of Pisa, Pisa, Italy

<sup>4</sup> KDD Lab, Scuola Normale Superiore, Pisa, Italy

**Abstract.** Abstaining classifiers have the option to refrain from providing a prediction for instances that are difficult to classify. The abstention mechanism is designed to trade off the classifier’s performance on the accepted data while ensuring a minimum number of predictions. In this setting, often fairness concerns arise when the abstention mechanism solely reduces errors for the majority groups of the data, resulting in increased performance differences across demographic groups. While there exist a bunch of methods that aim to reduce discrimination when abstaining, there is no mechanism that can do so in an explainable way. In this paper, we fill this gap by introducing Interpretable and Fair Abstaining Classifier (IFAC), an algorithm that can reject predictions both based on their uncertainty and their unfairness. By rejecting possibly unfair predictions, our method reduces error and positive decision rate differences across demographic groups of the non-rejected data. Since the unfairness-based rejections are based on an interpretable-by-design method, i.e., rule-based fairness checks and situation testing, we create a transparent process that can empower human decision-makers to review the unfair predictions and make more just decisions for them. This explainable aspect is especially important in light of recent AI regulations, mandating that any high-risk decision task should be overseen by human experts to reduce discrimination risks. (Code and Appendix for this work is available on: <https://github.com/calathea21/IFAC>).

**Keywords:** Reject Option · Fair ML · Interpretable ML

## 1 Introduction

Over the last 15 years, much research has been conducted on creating fairness-aware classification algorithms. While a lot of work has been done on creating automatized solutions based on some mathematical definition of fairness, recently the call for more flexible approaches has been growing. Rather than trying to define or achieve fairness through one numeric measure for the entire

system, there is a growing recognition that we need to understand under which circumstances unfairness occurs, which groups are most affected by it, and which differences in the treatment of demographic groups might be justifiable [12, 45]. Because of the delicate and nuanced nature of these questions, there is also an increased consensus that automated algorithms cannot be used alone in the identification and resolution of bias, but instead should actively be overseen and adapted by human experts with sufficient knowledge about a domain and the historic biases in place. This call for human-in-the-loop approaches for algorithmic fairness is now even mandated by AI legislation, such as the EU AI Act [16]. Despite the clear call that human oversight and control are necessary, the legislation says little about how it should take place [16]. A way to put humans in the loop during the deployment of a system is provided by the framework of selective classification. The original idea behind this framework is to build a classifier that abstains from making a prediction when it is not certain about it. In other words, these models *reject* ambiguous instances and pass them to better decision models or human experts, to increase accuracy over all non-rejected instances. Even though this idea originally dates back to the 1970s [9], it has only barely been explored in the context of increasing the fairness of models, by abstaining from predictions that might be unfair. Ensuring the interpretability of such abstentions, and explaining why instances are seen as unfair can further empower humans to understand whether to override original decisions or not, and increase the overall fairness of the decision process [43].

In this work, we exploit this idea by proposing an Interpretable Fair Abstaining Classifier (IFAC) for building selective classifiers that do not only abstain from making decisions in cases of uncertainty but also in cases of unfairness. We do so by adding an inherently interpretable mechanism for unfairness-based rejections to a selective classifier, thus allowing the user to inspect the unfair decisions of the model and the instances they need to review.

The paper is organized as follows: in Sect. 2 we list the main papers in the literature relevant to our work, in Sects. 3 and 4 we provide, respectively, the necessary mathematical background and formulation of our method, in Sect. 5 we provide a thorough experimental evaluation of our method and finally in Sect. 6 we discuss our results and conclude the paper.

## 2 Related Literature

**Fairness in Classification.** Classifiers exhibiting discriminatory behaviour towards certain demographic groups have been a concern for some time now [36]. Over the years, many metrics have been proposed to measure discrimination in these settings. These include *group metrics*, such as demographic parity and equal odds, that compare how classifiers behave over different population groups in the data. Particularly, demographic parity compares a classifier’s output ratios and equal odds its error ratios across demographics [36]. Next to *group metrics*, there are *individual metrics* to identify for one instance at a time whether they are affected by discrimination. These metrics operate on the principle of *treating likes alike* and check if similar individuals receive similar decision

outcomes [36]. When it comes to mitigating bias in classification tasks, a common approach is to choose one of the available metrics and build a classifier to satisfy the associated fairness goal while maintaining its predictive accuracy [7, 24, 48]. Recently, however, the simplicity of these approaches has been criticized: optimizing for group metrics comes with the risk of *cherry-picking*, the practice of arbitrarily changing prediction labels in pursuit of some “superficial” fairness goal, without further attention to whether the decisions make sense on an individual level [18]. Contrarily, only paying attention to individual fairness does not ensure that discrimination does not still happen globally, and certain demographic groups are not systematically excluded from receiving favourable decision outcomes [18]. Hence, researchers have argued that instead of fixating on one fairness goal in an automated manner, any efforts to detect and mitigate discrimination should be guided by domain experts, who can take a more holistic approach to fairness, and make nuanced considerations about the nature of bias and how to address it [25, 41, 45]. Related to this, researchers have also pointed out the importance of addressing intersectional discrimination [12]. This describes the unique discrimination that people from a combination of marginalized groups (e.g., black women) face, which cannot be solely explained by the “sum” of discrimination faced by each marginalized group in isolation (e.g., being black and being female) [13]. Currently, many works on fair classification only focus on discrimination experienced by demographic groups as defined by a single binary-sensitive feature. Recognizing that algorithmic harms can only be combated when understanding how they uniquely unfold, some studies like [6, 19, 46] have started incorporating intersectionality in their research.

**Prediction with a Reject Option.** The idea to allow a machine learning model to abstain in the prediction stage dates back to the 1970s, when it was introduced for classification tasks [9]. Two main frameworks allow one to learn abstaining models, i.e. ambiguity rejection and novelty rejection [26]. The former focuses on abstaining from instances where mistakes are more likely; the latter builds methods that abstain on instances that are largely dissimilar from the training data distribution [30, 35, 47]. Within ambiguity rejection, we can further distinguish between Learning to Reject (LtR) [9] and Selective Prediction (SP) [15]. The former (LtR) requires one to define a class-wise cost function that penalizes mispredictions and rejections [10, 11]. The latter (SP) requires instead one to either pre-define a target coverage  $c$  to achieve and minimize the risk (*bounded-abstention*) [23, 28, 38, 39], or fix a target risk  $e$  to guarantee and maximize the coverage (*bounded-improvement*) [21, 22].

**Fairness and Reject Option.** There are a few works that analyze the effects on fairness caused by a reject option. [29] show that even if abstaining can improve the overall accuracy, some demographic groups can be negatively impacted by the reject option. [31] propose a surrogate loss for the classification task considering performance on different subgroups of instances. The proposed loss allows enforcing a sufficiency condition to avoid unfair results. A similar approach for the regression task is proposed by [42]. [40] provide a theoretical analysis of the

selective classification framework when introducing a fairness constraint in the bounded-abstention problem.

**Explainability and Reject Option.** The study of explainable AI (XAI) methods in the context of abstaining classifiers is limited. [17] propose a reject option for natively interpretable models such as prototype-based ones. [2] consider counterfactual techniques to explain reject options of learning vector quantization classifiers. [3] introduce semi-factual explanations for the reject option, yielding a model-agnostic approach at the expense of potentially high complexity. Finally, [4] propose a model-agnostic framework to explain the abstention mechanism, including counterfactual, semi-factual, and factual approaches.

### 3 Background

#### 3.1 Selective Classification

Consider the triplet  $(\mathbf{L}, \mathbf{S}, Y)$ :  $\mathbf{L}$  represents the legally-grounded features and takes values in  $\mathcal{L} \subseteq \mathbb{R}^{d_l}$ ;  $\mathbf{S}$  refers to the sensitive attributes and takes values in  $\mathcal{S} \subseteq \mathbb{R}^{d_s}$ ;  $Y$  is the (binary) target variable, whose domain is  $\mathcal{Y} = \{0, 1\}$ . For example, if  $Y$  encodes being rich and our goal is to predict  $Y$  given some set of features,  $\mathbf{L}$  could include educational level and employment status, while  $\mathbf{S}$  could refer to gender or race. We denote with  $\mathcal{X} = \mathcal{L} \times \mathcal{S}$  the whole feature space and with  $\mathbf{X} = (\mathbf{L}, \mathbf{S})$  the pair of both legally grounded and sensitive features.

Given the hypothesis space  $\mathcal{H}$  of functions (classification models) mapping  $\mathcal{X}$  to  $\mathcal{Y}$ , a learning algorithm aims to find a hypothesis  $h \in \mathcal{H}$  such that it minimizes some risk measure  $R(h) = \mathbb{E}[l(h(\mathbf{X}), Y)]$ , where  $l : \mathcal{Y} \times \mathcal{Y} \rightarrow \mathbb{R}$  is a *loss function* and  $\mathbb{E}$  is computed over the joint probability distribution  $P(\mathbf{X}, Y)$ .

To reduce the classifier’s error rates, one can add a selection mechanism that allows the model to abstain from predicting over more difficult-to-classify instances. More formally, we can define a selective classifier<sup>1</sup> as:

$$(h, g)(\mathbf{x}) = \begin{cases} h(\mathbf{x}) & \text{if } g(\mathbf{x}) = 1 \\ \text{abstain} & \text{otherwise,} \end{cases} \quad (1)$$

where  $g : \mathcal{X} \rightarrow \{0, 1\}$  is the so-called *selection function* or *rejector*<sup>2</sup>.

In practice, the selection function is often obtained by setting a threshold  $\tau$  on a confidence function  $v : \mathcal{X} \rightarrow \mathbb{R}$ , which determines the portion of the data on which the classifier is more likely to misclassify. In such a case, the selection function can be defined as  $g(\mathbf{x}) = \mathbb{1}\{v(\mathbf{x}) \geq \tau\}$ .

To avoid rejecting too many instances, the selective classification framework introduces *the coverage*, i.e. the percentage of instances for which the selective classifier must provide a prediction. Coverage is denoted as  $\phi(g) = \mathbb{E}[g(\mathbf{X})]$  and can be traded off for performance improvements. In this case, performance

<sup>1</sup> In this work, we use the terms *abstaining* and *selective* interchangeably.

<sup>2</sup> We use the term *abstain* and *reject* when  $g(\mathbf{x}) = 0$  and *accept* or *selects* when  $g(\mathbf{x}) = 1$ .

is measured through the risk over the accepted region, commonly called the *selective risk* and defined as  $R(h, g) = \frac{\mathbb{E}[l(h(\mathbf{X}), Y)g(\mathbf{X})]}{\phi(g)}$ .

To find a selective classifier that minimizes selective risk, it is necessary to select a lower bound  $c$  as a *target coverage* [23]. Given a target coverage  $c$ , an optimal selective predictor  $(h, g)$  (parameterized by  $\theta^*$ ,  $\psi^*$ ) is defined as:

$$\arg \min_{\theta \in \Theta, \psi \in \Psi} R(h_\theta, g_\psi) \quad \text{s.t.} \quad \phi(g_\psi) \geq c \quad (2)$$

We learn the optimal parameters using an empirical counterpart of selective risk and coverage, using an i.i.d. dataset  $\mathcal{D} = \{(\mathbf{x}_i, y_i)\}_{i=1}^n$  drawn from  $P$ .

Finally, we call *coverage-calibration* the post-training procedure of estimating the threshold  $\tau$  for the target coverage  $c$  specified in Eq. 2. This is generally done by estimating the  $(1-c) \cdot 100$ -th percentile of the confidence function over a held-out calibration dataset.

### 3.2 Measuring Fairness With Association Rules and Situation Testing

**Association Rules:** In our methodology, we make use of association rules to identify discriminatory behaviour of a base classifier  $h$ , upon which  $g$  can decide to reject its predictions. Let us assume we have access to a dataset of realizations  $\mathcal{D}$ . We recall  $\mathbf{x}_i = (\mathbf{l}_i, \mathbf{s}_i) = (l_i^1, \dots, l_i^{d_l}, s_i^1, \dots, s_i^{d_s})$ , where  $l_i^j$  refers to the value taken by the  $j^{\text{th}}$  legally grounded feature of instance  $i$  and  $s_i^j$  to the  $j^{\text{th}}$  sensitive feature of instance  $i$ .

We call a specific realization of a single variable within  $\mathbf{x}_i$  an *item*, e.g. if we consider the variable `race`, `race=White` is an item. Let  $\mathcal{I}$  be the set of all possible items. A subset  $I$  of  $\mathcal{I}$  is called an *itemset*.

We can decompose  $I$  into its legally grounded and sensitive parts,  $I = (I_L, I_S)$ , where  $I_L$  is an itemset containing only legally grounded features and  $I_S$  is an itemset that contains only sensitive ones. A transaction  $T$  is a subset of  $I$  with exactly one item for every feature in  $\mathbf{x}$ . In other words, a sampled instance's features  $\mathbf{x}_i$  can be seen as a transaction  $T$ . For a transaction  $T$ , we say  $T$  *verifies* itemset  $(I_L, I_S)$  if  $(I_L, I_S) \subseteq T$ . The support of itemset  $(I_L, I_S)$  with respect to the dataset  $\mathcal{D}$  is denoted as  $\text{supp}_{\mathcal{D}}((I_L, I_S)) = \frac{|\{T \in \mathcal{D} : (I_L, I_S) \subseteq T\}|}{|\mathcal{D}|}$ .

A decision rule is an expression  $(I_L, I_S) \rightarrow Y$ . The support of a decision rule is  $\text{supp}_{\mathcal{D}}((I_L, I_S) \rightarrow Y) = \text{supp}_{\mathcal{D}}((I_L, I_S), Y)$ . The confidence of the rule is then defined as  $\text{conf}_{\mathcal{D}}((I_L, I_S) \rightarrow Y) = \frac{\text{supp}_{\mathcal{D}}((I_L, I_S), Y)}{\text{supp}_{\mathcal{D}}((I_L, I_S))}$ .

To measure the impact of the sensitive features of a decision rule, the Selective Lift (*slift*) measure introduced by [34] can be used. In this paper we use the definition by *difference* of *slift*, which is detailed as follows:

$$\text{slift}_{\mathcal{D}}((I_L, I_S) \rightarrow Y) = \text{conf}_{\mathcal{D}}((I_L, I_S) \rightarrow Y) - \text{conf}_{\mathcal{D}}((I_L, \neg I_S) \rightarrow Y) \quad (3)$$

Computing  $\text{conf}_{\mathcal{D}}(I_L, \neg I_S) \rightarrow Y$  requires one to take the confidence of all the transactions that verify  $I_L$  but do not verify  $I_S$ .

*Example.* Consider an association rule  $\text{race} = \text{Black}, \text{education} = \text{Masters} \rightarrow \text{income} = \text{low}$ , with  $\text{race} \subseteq \mathbf{S}$  and  $\text{education} \subseteq \mathbf{L}$  and  $\text{income} = Y$ . Imagine the confidence of this rule is 0.90 and its slift is 0.50. This means that the confidence of  $\text{race} \neq \text{Black}, \text{education} = \text{Masters} \rightarrow \text{income} = \text{low}$  is  $0.90 - 0.50 = 0.40$ . Because of this high difference  $\text{race} \neq \text{Black}, \text{education} = \text{Masters}$  could be seen as a subgroup at risk of discrimination.

As indicated by [33], decision rules can be learned on the original data using algorithms like Apriori [1] and then filtered according to fairness-based policies.

**Situation Testing:** Since association rules only detect global discrimination patterns, one can use the Situation Testing algorithm to further analyse fairness on a local level [44]: To check whether instance  $\mathbf{x}_i$  receives a fair outcome  $Y$ , we use a distance function to search  $\mathcal{D}$  for  $\mathbf{x}_i$ 's  $k$ -nearest neighbors from a reference group and a non-reference group, meaning we obtain two sets of instances  $\mathcal{K}_{tr}^r$  and  $\mathcal{K}_{tr}^{nr}$ . A reference group is defined by sensitive feature values of those instances from the data we assume to be treated favorably, for instance,  $\text{race} = \text{White}, \text{sex} = \text{Male}$ . All instances not belonging to this group are seen as the non-reference group. To define instance  $\mathbf{x}_i$ 's individual discrimination score we calculate the ratio of positive decision ratio for  $\mathcal{K}_{tr}^r$  and  $\mathcal{K}_{tr}^{nr}$ :  $dec_r = \frac{|\{j \in \mathcal{K}_{tr}^r : y_j = 1\}|}{k}$ ,  $dec_{nr} = \frac{|\{j \in \mathcal{K}_{tr}^{nr} : y_j = 1\}|}{k}$  and take the difference between both ( $dec_r - dec_{nr}$ ). If this score exceeds the user-defined individual discrimination threshold  $t$ , it indicates that the treatment reserved to instance  $i$  depends on its sensitive characteristics.

## 4 Methodology

We propose to learn a selective classifier that does not only reject instances based on the uncertainty of their predictions but also their unfairness. In doing so we can decrease unfairness over all non-rejected instances. Further, by providing explanations for why some predictions are marked as unfair, we aid human reviewers in understanding whether the fairness concerns are indeed justified and enable a more informed decision process over them. We call our approach IFAC (Interpretable and Fair Abstaining Classifier). The intuition behind IFAC is visualized in Fig. 1: on top of the base classifier  $h$  we have our rejector  $g$ , which takes an instance's features  $\mathbf{x}_i$  and the classifier  $h$ 's prediction as its input. The rejector first executes a global fairness analysis on this instance, checking if it falls under any subgroups at risk of discrimination, as identified by discriminatory association rules (Sect. 3.2). If it does, it performs a local fairness check using Situation Testing [44], evaluating how the prediction for  $h(\mathbf{x}_i)$  compares to the labels of similar instances in the data. After this, a *certainty assessment* is performed. Depending on the outcome of the assessment and the former fairness analysis there are four possibilities for our rejector: in case the prediction is deemed as fair and it exceeds a dedicated confidence threshold, the prediction is kept. Contrary, fair predictions that fall below this threshold are rejected. If we are dealing with an unfair prediction exceeding a separate confidence threshold for unfair data, it also gets rejected: though the prediction is certain, we have

reasons to doubt it, because it is unfair. Finally, on predictions that are both unfair and uncertain, IFAC flips the original classifier  $h(\mathbf{x}_i)$  prediction. The reasoning behind these interventions is that predictions that are neither fair nor certain are probably inaccurate, to begin with, and it is safe to alter them. This flipping mechanism is also added in case the user-defined *coverage* for IFAC does not allow to reject *all* unfair predictions. A complete walk-through example of how IFAC makes rejections is provided in Appendix A.

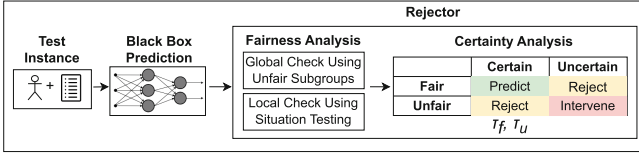


Fig. 1. Intuition behind IFAC

Now that we have described the basic intuition behind how IFAC is applied, we outline how it is learned. Given some data  $\mathcal{D}$ , we split it into a training set  $\mathcal{D}_{tr}$  and two validation sets  $\mathcal{D}_{val_1}$ ,  $\mathcal{D}_{val_2}$ . Then, given the target coverage  $c$  and the unfair reject weight  $w_u$ <sup>3</sup>, IFAC is devised as follows:

1. **Learn a classifier:** we train classifier  $h$  from  $\mathcal{D}_{tr}$ . We highlight that any off-the-shelf probabilistic classifier can be considered, making our approach model-agnostic;
2. **Learn at-risk subgroups:** we extract association rules from validation set  $\mathcal{D}_{val_1}$ . The rules allow us to understand if there are correlations between sensitive features  $\mathbf{S}$  and predictions of  $h$ , and, consequently, identify at-risk subgroups [33];
3. **Situation Testing:** we prepare the hyperparameters and distance function to run Situation Testing.
4. **Calibration:** we use the second validation set  $\mathcal{D}_{val_2}$  to calibrate the rejection strategy, considering both *unfairness* and *uncertainty*:
  - (i) the learned association rules are applied on  $\mathcal{D}_{val_2}$ ;
  - (ii) situation testing is performed for those instances falling under discriminatory patterns. This allows one to split the sample into a *fair* part  $\mathcal{D}_{val_2^f}$  and an *unfair* one  $\mathcal{D}_{val_2^u}$ ;
  - (iii) depending on  $c$  and  $w_u$ , we estimate two different rejection thresholds, i.e.  $\tau_f$  and  $\tau_u$ . These thresholds are computed following the *coverage-calibration* procedure described in Sect. 3, ranking instances w.r.t. the confidence function over samples  $\mathcal{D}_{val_2^f}$  and  $\mathcal{D}_{val_2^u}$  respectively.

Fig. 2 summarizes the steps needed to learn IFAC. In the rest of this section, we further detail steps 2, 3, and 4.

<sup>3</sup> The unfair reject weight  $w_u$  determines how many rejections can be made based on unfairness concerns.

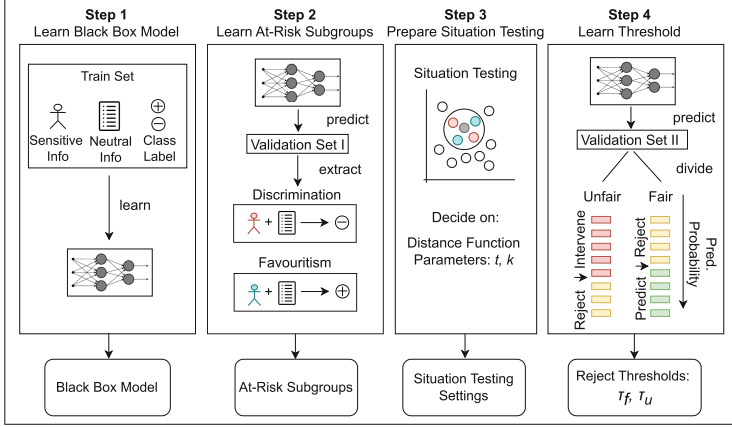


Fig. 2. The four steps for learning IFAC.

#### 4.1 Step 2: Learn At-Risk Subgroups

To learn global patterns of unfairness, we use discriminatory association rules, as described in Sect. 3.2. To do so we apply  $h$  on the first validation set  $\mathcal{D}_{val_1}$  and extract the association rules for the data and  $h$ s predictions with the apriori algorithm. We do so separately for each sensitive feature value and their combination. For example, let us have two sensitive attributes **sex** and **race** with two possible values, F, M and W, B respectively. We apply apriori and extract rules for each of the itemsets:  $\{\text{sex}=\text{M}\}$ ,  $\{\text{sex}=\text{F}\}$ ,  $\{\text{race}=\text{W}\}$ ,  $\{\text{race}=\text{B}\}$ ,  $\{\text{sex}=\text{M} \wedge \text{race}=\text{B}\}$ ,  $\{\text{sex}=\text{M} \wedge \text{race}=\text{W}\}$ ,  $\{\text{sex}=\text{F} \wedge \text{race}=\text{W}\}$ ,  $\{\text{sex}=\text{F} \wedge \text{race}=\text{B}\}$ . Thus, the number of rules found meeting minimum support is not biased towards the largest demographic groups in the data.

As per our previous notation, we extract rules in the form of  $(I_L, I_S) \rightarrow Y$ , for some prediction outcome  $h(\mathbf{x})$  in a binary classification setting  $Y \in \mathcal{Y} = \{0, 1\}$ . We say that rules with  $Y = 0$  describe potentially discriminated subgroups, while rules with  $Y = 1$  describe potentially favored ones. We extract favoring associations only for fixed reference groups defined for our data, e.g. white men (as described in Sect. 3.2). After extracting both favoring and discriminatory associations, we filter out statistically significant rules meeting an *sift* threshold. We calculate statistical significance using Z-test, testing if the proportion of some decision outcome  $Y$  is significantly different for the groups  $(I_L, I_S)$  and  $(I_L, \neg I_S)$  [8]. We only select rules with  $p < 0.01$ . Further, we filter out *high-sift* rules by checking for which ones the following holds:

$$conf_{\mathcal{D}_{val_1}}((I_L, I_S) \rightarrow Y_v) - sift_{\mathcal{D}_{val_1}}((I_L, I_S) \rightarrow Y_v) < 0.5 \quad (4)$$

Which in the context of binary classification is true *iff*:

$$conf_{\mathcal{D}_{val_1}}((I_L, \neg I_S) \rightarrow Y_v) < conf_{\mathcal{D}_{val_1}}((I_L, \neg I_S) \rightarrow \neg Y_v) \quad (5)$$



Intuitively, this means that we only select the groups  $\{I_L, I_S\}$  for which negating the sensitive part of the group ( $\{I_L, \neg I_S\}$ ) yields higher confidence for value  $Y_v$  w.r.t. the opposite value  $\neg Y_v$  (brief proof in Appendix Section B).

## 4.2 Step 3: Situation Testing

Part of the abstention mechanism of IFAC is based on a local fairness check for instances that are covered by global discrimination patterns. The aim is to use the global check to identify larger subgroups at risk of unfair treatment, while the local check allows us to execute a more fine-grained analysis taking all of an instance’s characteristics into account. Our local fairness check is performed via Situation Testing, comparing a prediction  $h(\mathbf{x}_i)$  for instance  $i$  with the decision labels of similar instances from  $\mathcal{D}_{tr}$  (see Sect. 3.2). For the algorithm, a suitable distance function must be chosen e.g. we can consider the one used by Luong et al. [44] or one learned from the data [32]. We follow Luong’s suggestion of a context-dependent approach and let an expert choose hyperparameters  $t$  and  $k$  depending on the decision task [44].

## 4.3 Step 4: Calibrate Rejection Strategy

Whether the rejector keeps, rejects, or intervenes on the original prediction for  $\mathbf{x}$ , depends on the (un)certainty of the base classifier. To evaluate the confidence of the classifier, we resort to the softmax response  $v(\mathbf{x}) = \max_{y \in \mathcal{Y}} s_y$  [20, 22], where  $s_y(\mathbf{x}) \approx P(Y = y | \mathbf{X} = \mathbf{x})$  is an estimate of the conditional probability. We then estimate two thresholds  $\tau_f$  and  $\tau_u$  to choose between prediction, intervention, and abstention. The final selective classifier is in the form:

$$(h, g)(\mathbf{x}) = \begin{cases} h(\mathbf{x}) & \text{if } Fair(\mathbf{x}) \text{ and } v(\mathbf{x}) \geq \tau_f \\ \text{abstain} & \text{if } Fair(\mathbf{x}) \text{ and } v(\mathbf{x}) < \tau_f \\ 1 - h(\mathbf{x}) & \text{if } \neg Fair(\mathbf{x}) \text{ and } v(\mathbf{x}) < \tau_u \\ \text{abstain} & \text{if } \neg Fair(\mathbf{x}) \text{ and } v(\mathbf{x}) \geq \tau_u \end{cases}$$

To learn  $\tau_f$  and  $\tau_u$ ,  $h$  is applied on our second validation dataset  $\mathcal{D}_{val_2}$  and its predictions are extracted. We then first extract those predictions that fall under discriminatory associations as learned in Step 2. After, we apply the Situation Testing algorithm as set up in Step 3 on those instances, and extract all that fail this individual fairness test. We consider those as the unfair fraction of the validation data ( $\mathcal{D}_{val_2^u}$ ) and the remaining ones as the fair fraction  $\mathcal{D}_{val_2^f}$ . The number of rejections that can be made for both groups is determined by two parameters given by the user, namely the target coverage  $c$  and the unfair reject weight  $w_u$ . Given that the  $\mathcal{D}_{val_2}$  consists of  $N$  instances of which  $N_u$  belong to  $\mathcal{D}_{val_2^u}$  and  $N_f$  belong to  $\mathcal{D}_{val_2^f}$ , we calculate the number of total rejections ( $N_{rej}$ ), the number of unfairness-based rejections ( $N_{ufr}$ ) and the number of uncertainty-based rejections ( $N_{ucr}$ ) as follows:

$$N_{\text{rej}} = \lceil (1 - c) \cdot N \rceil; \quad N_{\text{ufr}} = \min(\lceil N_{\text{rej}} \cdot w_u \rceil, N_u); \quad N_{\text{ucr}} = N_{\text{rej}} - N_{\text{ufr}} \quad (6)$$

We then proceed by separately ordering the fair and unfair instances of the validation data according to the confidence function  $v(\mathbf{x})$ . On the fair instances, we determine the threshold  $\tau_f$  such that  $N_{\text{ucr}}$  instances fall below this threshold, and on the unfair sample such that  $N_{\text{ufr}}$  instances exceed  $\tau_u$ .

## 5 Experimental Evaluation

The goal of our experimental section aims to address the following questions:

- Q1:** Does IFAC achieve comparable results to state-of-the-art selective classifiers in terms of predictive performance and fairness?
- Q2:** How does IFAC explain the drivers behind unfairness-based rejections, and how could these explanations be utilized?
- Q3:** How do *coverage*  $c$  and the *unfair-reject weight*  $w_u$  affect our results?

### 5.1 Experimental Settings

**Data and Baselines.** We run experiments considering two real datasets, namely ACSINCOME [14] and WISCONSINRECIDIVISM [5]. The former is about predicting high or low income based on instances’ education, occupation etc. We define **sex** (male vs. female) and **race** (white vs. black vs. other) as sensitive attributes and take the group of white men as our reference group. We compare their treatment to each intersectional group based on race and sex.

WISCONSINRECIDIVISM contains information about criminal defendants, like their type of offense, number of prior offenses, etc. The task is to predict if they will not recidivate. We take **race** as the sensitive attribute (white vs. black vs. other). Because of a base classifiers’ lower False Negative and higher False Positive rates on white people, we define this as the reference group<sup>4</sup>.

We use different classification algorithms, namely a Random Forest, a Neural Network, and an XGBoost Classifier. We fitted all models with the default parameters of the corresponding Python libraries. Starting from these base classifiers, we compare IFAC with the following model-agnostic methods:

- *Full Coverage* (FC): the classifier itself when predicting on all the instances ( $c = 1.00$ )
- *Uncertainty Based Abstaining Classifier* (UBAC): The plug-in algorithm by [27]. This is the most well-known model-agnostic method and achieves state-of-the-art performance [37]. As for IFAC, we consider  $v(\mathbf{x}) = \max_{y \in \mathcal{Y}} s_y(\mathbf{x})$  as the confidence function. The rejection threshold is computed according to the *coverage-calibration* procedure.

<sup>4</sup> For full details on the preprocessing steps executed on both datasets we refer to our [github repository](#).

Because we consider discrimination based on non-binary sensitive attributes (and in the case of ACSINCOME even intersectional discrimination), we do not compare with the fair abstention mechanism of Schreuder et al. [40] as a baseline, which only works on a single binary sensitive feature.

**Hyperparameters.** For **Q1** and **Q2**, we set  $c = .80$  for the abstaining classifiers. Further, for IFAC we set the *unfair reject weight* ( $w_u$ ) equal to 1.0. The intuition behind this is that if the coverage is large enough, IFAC should abstain from predicting any unfair instance, and only if not, fairness interventions should be performed. For the Situation Testing algorithm used by IFAC we set  $k$ , i.e. the number of neighbors used for the fairness comparisons to 10, and  $t$  to 0.3. For extracting discriminatory association rules we use the apriori algorithm of `apriori` with min. support of 0.01 and min. confidence of 0.85.

**Metrics.** For **Q1**, we evaluate predictive performance in terms of accuracy, precision, and recall on all non-rejected instances. Concerning fairness measures, we report the False Negative, False Positive, and Positive Decision Rates for the different demographic groups of each dataset. Further, we report the range and the standard deviation across demographic groups over these measures. Note, that we define these measures regarding the desirable label of each dataset. Hence, the positive decision ratio for ACSINCOME is the ratio of *high* income prediction, and for WISCONSINRECIDIVISM it is the ratio of *non-recidivism* predictions.

**Experimental Setup.** We split each dataset into training, two validation, and a test part (40% for train, 15% for each validation, and 30% for test) and train the classifiers on the former. For IFAC we learn the discriminatory associations on the first validation set. The reject thresholds for both IFAC and UBAC are calibrated based on the second. Finally, we randomly split the test set into 10 samples [32] and compute the final metrics on each of these samples. We provide results as averages and standard errors over these 10 test set samples.

## 5.2 Results

**Q1: Performance and Fairness.** We describe the predictive performance on each dataset and each classifier-methodology combination in Table 1. As can be seen, both selective classification methods improve upon the performance of FC, however, for UBAC this improvement is slightly larger, especially for the income prediction task.

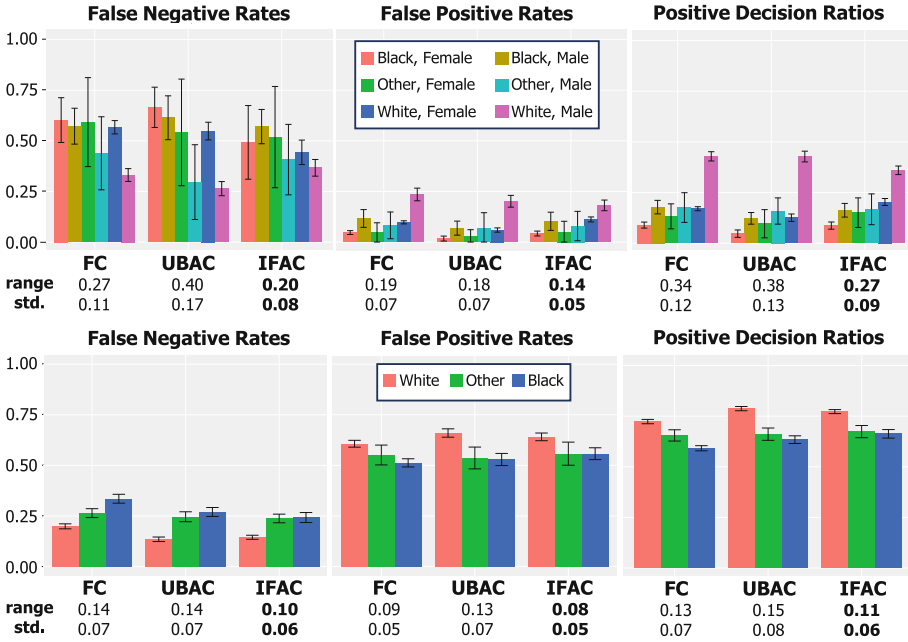
In Fig. 3 we can see how the increased performance of UBAC comes at the cost of its fairness. In this Figure, we highlight the results of a Random Forest classifier combined with different selective classification methods, showing the average False Negative -, False Positive, and Positive Decision Rates (FNR, FPR, and PDR) over demographic groups (the results for Neural Networks and XGBoost follow the same patterns and are included in the Appendix). We also highlight the range of these metrics across demographics (i.e. the performance difference between the highest- and lowest performing group) and the standard deviation. Fairer classifiers should score lower on both metrics, to ensure that there are no big performance differences across groups.

**Table 1.** Performance Results ACSINCOME and WISCONSINRECIDIVISM

		ACSINCOME			WISCONSINRECIDIVISM		
		Acc.	Rec.	Prec.	Acc.	Rec.	Prec.
<b>RF</b>	FC	.78 ± .01	.57 ± .02	.65 ± .03	.62±.01	.77±.01	.65±.01
	UBAC	<b>.83 ± .01</b>	<b>.62 ± .02</b>	<b>.69 ± .03</b>	<b>.65±.01</b>	<b>.83±.01</b>	<b>.66±.01</b>
	IFAC	.80 ± .01	.59 ± .04	.64 ± .03	<b>.65±.01</b>	<b>.83±.01</b>	<b>.66±.01</b>
<b>NN</b>	FC	.80 ± .01	.58 ± .03	.71 ± .03	.63±.01	0.74±.01	.65±.01
	UBAC	<b>.86 ± .01</b>	<b>.62 ± .03</b>	<b>.77 ± .03</b>	<b>.66±.02</b>	<b>.77±.01</b>	<b>.68±.02</b>
	IFAC	.83 ± .01	.58 ± .03	.73 ± .02	<b>.66±.02</b>	.76±.01	<b>.68±.02</b>
<b>XGB</b>	FC	.81 ± .01	.60 ± .03	.73 ± .03	.63±.01	.77±.01	.65±.01
	UBAC	<b>.87 ± .01</b>	<b>.64 ± .03</b>	<b>.78 ± .03</b>	<b>.66±.01</b>	<b>.83±.01</b>	<b>.68±.01</b>
	IFAC	.84 ± .01	.59 ± .03	.75 ± .03	<b>.66±.01</b>	.82±.01	<b>.68±.01</b>

Starting with ACSINCOME, we see that for UBAC this is not the case: we observe an especially unequal distribution of FNR across demographic groups, with the highest difference being 0.4 (between white men and black women). This difference is even higher than for the FC classifier, as the UBAC selection mechanism only decreases the FNR for white men while increasing it for others. With using IFAC this effect does not occur: through rejecting predictions that are at high risk of unfairness, FNRs decrease for minority groups like women or black people, and overall the rates become more equal across demographics, bringing the range down to 0.2 and the std. to 0.08. The patterns are slightly less strong when considering the FPR and PDR across demographics, but still hold. Similar patterns occur for WISCONSINRECIDIVISM: the range and standard deviation for FNR, FPR, and PDR across demographics decrease when using IFAC, while they increase with UBAC. We acknowledge that the effect is less strong here, but attribute this to IFACs selection criteria for unfair instances being too strict. In Appendix D we show results with a lower threshold  $t$  for situation testing (meaning that more instances can get rejected out of unfairness concern), where IFAC makes FNR, FPR, and PDR nearly equal across groups. Further, we highlight how equalizing error rates across demographics is only the first step towards improving the fairness of the decision task. As we illustrate in the next section, enabling humans to review rejected instances and the explanation behind them, is the most crucial contribution of our method.

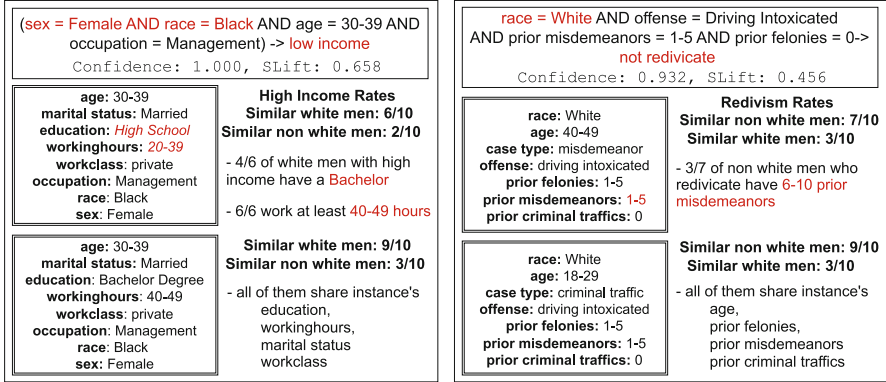
**Q2: Explaining Unfair Rejections.** One of the main advantages of IFAC is that it can explain why rejected predictions are seen as unfair. In Fig. 4 we show some explanations behind rejected instances for both of our datasets, and we use the ACSINCOME case to highlight how a human expert can utilize them. We see two instances that were both rejected based on the same global pattern of unfairness: the classifier predicting “low income” ratios for black women, aged between 30 and 39 working in management, than for people with the same age



**Fig. 3.** Performance measures over demographic groups when applying a Random Forest in combination with various selective classifiers on ACSINCOME (above) and WISCONSINRECIDIVISM (below). A regular UBAC increases differences in error- as well as positive decision rates among groups. Using IFAC, and rejecting instances based on unfairness, diminishes these differences.

and occupation, but different demographics. While an algorithm only analyses such patterns statistically, human experts can examine them with sensitivity surrounding their historical context. For instance, it is well known that racism and sexism contribute to hostile work environments for black women. Hence, a human expert can reason how these dynamics may hinder fair compensation in roles like management, that are normally associated with high salaries.

The results of situation testing provide further insight into the unfairness of the classifier: For both instances, a high ratio of the 10 most similar white men have a high income; explaining why their own low income predictions are marked as unfair. However, for the first instance, many of the white men considered for the comparison have a higher education level and amount of working hours than her. Since it makes sense, that people working part-time do not get the same compensation as people working full-time, the low income prediction could be seen as justified and a human reviewer could decide to keep it. For the second case, all similar white men do share the instances' education level, working hours, etc. Hence, there is no justification for why she would be the only one receiving a low income prediction, and a human expert could decide to override this decision.



**Fig. 4.** Examples for ACSINCOME (left) and WISCONSINRECIDIVISM (right) of two rejected instances, and the explanation behind their rejections.

To conclude, these examples show how IFAC's interpretable-by-design rejector can have a large impact in increasing the fairness of a decision process. In particular, our approach goes beyond a rough statistical analysis of discriminatory patterns and allow for the integration of human domain knowledge to achieve a much deeper fairness assessment.

*Q3: Effects of  $c$  and  $w_u$ .* In this section, we explore the effect of parameters  $c$  and  $w_u$  on IFAC's performance. Out of space constraints, we only report the results with a Random Forest as a base-classifier on ACSINCOME. The results for the other classifiers and the other dataset follow the same pattern and are included in the Appendix. In Fig. 5 we visualize how the accuracy, the range in positive decision ratio across demographics, and the standard deviation change as a function of the coverage and the  $w_u$ . Unsurprisingly, for both UBAC and IFAC the accuracy drops as the coverage increases. Regardless of the coverage and the  $w_u$  UBAC outperforms IFAC. Further, we see that a lower  $w_u$  comes at the cost of accuracy, especially when the coverage is high. Intuitively this makes sense:  $w_u$  determines how many of the unfair predictions are rejected, and for how many an intervention is performed. With the low weight of 0.25, the majority of unfair prediction labels are simply flipped, and only the ones with very high prediction probability are abstained from. With an increase in coverage, this pattern is more extreme, as the general number of instances that can be abstained from is lower. When observing the effect of differing coverages and  $w_u$  on the fairness of the predictions, we observe that performing more interventions (as a result of a lower  $w_u$ ) has a desirable effect: both the range and standard deviation of positive decision ratios decreases across demographics. The effect is again larger for higher coverages because fewer allowed rejections mean more interventions, which bring the positive decision ratios across demographic groups closer together.

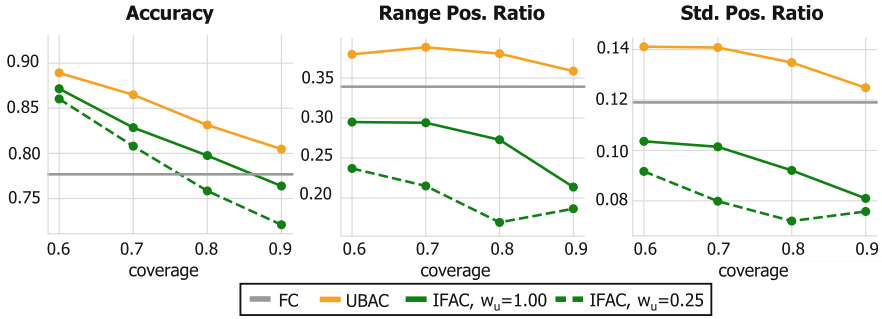


Fig. 5. Effects of  $c$  and  $w_u$  parameters in our selective classification settings.

## 6 Discussion and Conclusion

In this paper, we have introduced IFAC, an Interpretable and Fair Abstaining Classifier. This classifier rejects predictions from a base classifier, both in cases of uncertainty and unfairness. Unfairness rejections are based on the interpretable-by-design methods of unfair association patterns and situation testing. Through our experiments, we have shown how using our abstention mechanism yields satisfying overall performance, while improving fairness across demographic groups over all non-rejection instances. This stands in contrast to a regular uncertainty-based abstaining classifier, that does not take the fairness of predictions into account. We have also shown how the explanations behind our abstention mechanism, can empower human decision-makers to review the rejected instances and make fairer decisions for them. This holds immense potential for complying with recent AI regulations, which require automated decision-making processes to be supervised by humans to mitigate the risks of discrimination. By only having to review instances at high risk of unfairness, our framework can make this process more practical and time-efficient. To further empower human users, further research could involve human experts in the selection of *at-risk* subgroups and in choosing distance function and parameters for Situation Testing. Also, user studies can help in understanding how humans engage with such a system. For this, one should consider adding explanations for all non-rejected instances, so that humans can still explore the base classifier in the accepted cases.

**Acknowledgments.** D. Lenders and T. Calders were funded by Digitax Centre of Excellence UAntwerp and by Research Foundation Flanders under FWO file number: V467123N. A. Pugnana and R. Pellungrini and D. Pedreschi and F. Giannotti have received funding by PNRR - M4C2 - Investimento 1.3, Partenariato Esteso PE00000013 - “FAIR - Future Artificial Intelligence Research” - Spoke 1 “Human-centered AI”, funded by the European Commission under the NextGeneration EU programme, ERC-2018-ADG G.A. 834756 “XAI: Science and technology for the eXplanation of AI decision making” and Prot. IR0000013. This work was also funded by the European Union under Grant Agreement no. 101120763 - TANGO. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European

Union or the European Health and Digital Executive Agency (HaDEA). Neither the European Union nor the granting authority can be held responsible for them. The work has also been realised thanks to NextGenerationEU - National Recovery and Resilience Plan, PNRR) - Project: “SoBigData.it - Strengthening the Italian RI for Social Mining and Big Data Analytics” - Prot. IR000001 3 - Notice n. 3264 of 12/28/2021.

## References

1. Agrawal, R., Srikant, R.: Fast algorithms for mining association rules in large databases. In: VLDB, pp. 487–499. Morgan Kaufmann (1994)
2. Artelt, A., Brinkrolf, J., Visser, R., Hammer, B.: Explaining reject options of learning vector quantization classifiers. In: IJCCI, pp. 249–261. SCITEPRESS (2022)
3. Artelt, A., Hammer, B.: “even if ...” - diverse semifactual explanations of reject. In: SSCI, pp. 854–859. IEEE (2022)
4. Artelt, A., Visser, R., Hammer, B.: “i do not know! but why?” - local model-agnostic example-based explanations of reject. *Neurocomputing* **558**, 126722 (2023)
5. Ash, E., Goel, N., Li, N., Marangon, C., Sun, P.: WCLD: curated large dataset of criminal cases from wisconsin circuit courts (2023)
6. Cabrera, Á.A., Epperson, W., Hohman, F., Kahng, M., Morgenstern, J., Chau, D.H.: Fairvis: visual analytics for discovering intersectional bias in machine learning. In: 2019 IEEE Conference on Visual Analytics Science and Technology (VAST), pp. 46–56. IEEE (2019)
7. Calmon, F., Wei, D., Vinzamuri, B., Natesan Ramamurthy, K., Varshney, K.R.: Optimized pre-processing for discrimination prevention. In: *Advances in Neural Information Processing Systems*, vol. 30 (2017)
8. Casella, G., Berger, R.L.: *Statistical Inference*, Duxbury Press. Pacific Grove, CA (2002)
9. Chow, C.K.: On optimum recognition error and reject tradeoff. *IEEE Trans. Inf. Theory* **16**(1), 41–46 (1970)
10. Condessa, F., Bioucas-Dias, J.M., Castro, C.A., Ozolek, J.A., Kovacevic, J.: Classification with reject option using contextual information. In: ISBI, pp. 1340–1343. IEEE (2013)
11. Cortes, C., DeSalvo, G., Mohri, M.: Theory and algorithms for learning with rejection in binary classification. *Ann. Math. Artif. Intell.* **92**, 1–39 (2023)
12. Costanza-Chock, S., Raji, I.D., Buolamwini, J.: Who audits the auditors? recommendations from a field scan of the algorithmic auditing ecosystem. In: FAccT, pp. 1571–1583. ACM (2022)
13. Crenshaw, K.: Demarginalizing the intersection of race and sex: a black feminist critique of antidiscrimination doctrine, feminist theory and antiracist politics. In: *University of Chicago Legal Forum*, vol. 1989 (1989)
14. Ding, F., Hardt, M., Miller, J., Schmidt, L.: Retiring adult: new datasets for fair machine learning, pp. 6478–6490 (2021)
15. El-Yaniv, R., Wiener, Y.: On the foundations of noise-free selective classification. *J. Mach. Learn. Res.* **11**, 1605–1641 (2010)
16. Enqvist, L.: Human oversight’ in the EU artificial intelligence act: what, when and by whom? *Law Innov. Technol.* **15**(2), 508–535 (2023)
17. Fischer, L., Hammer, B., Wersing, H.: Optimal local rejection for classifiers. *Neurocomputing* **214**, 445–457 (2016)
18. Fleisher, W.: What’s fair about individual fairness? In: *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, pp. 480–490 (2021)



19. Foulds, J.R., Islam, R., Keya, K.N., Pan, S.: An intersectional definition of fairness. In: 2020 IEEE 36th International Conference on Data Engineering (ICDE), pp. 1918–1921. IEEE (2020)
20. Franc, V., Prusa, D., Voracek, V.: Optimal strategies for reject option classifiers. *J. Mach. Learn. Res.* **24**(11), 1–49 (2023)
21. Gangrade, A., Kag, A., Saligrama, V.: Selective classification via one-sided prediction. In: AISTATS, vol. 130, pp. 2179–2187. PMLR (2021)
22. Geifman, Y., El-Yaniv, R.: Selective classification for deep neural networks. In: NIPS, pp. 4878–4887 (2017)
23. Geifman, Y., El-Yaniv, R.: Selectivenet: a deep neural network with an integrated reject option. In: ICML, vol. 97, pp. 2151–2159. PMLR (2019)
24. Goel, N., Yaghini, M., Faltings, B.: Non-discriminatory machine learning through convex fairness criteria. In: Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society, pp. 116–116 (2018)
25. Goethals, S., Martens, D., Calders, T.: PreCoF: counterfactual explanations for fairness. *Mach. Learn.* **113**, 1–32 (2023)
26. Hendrickx, K., Perini, L., der Plas, D.V., Meert, W., Davis, J.: Machine learning with a reject option: a survey. ArXiv [abs/2107.11277](https://api.semanticscholar.org/CorpusID:236318084) (2021). <https://api.semanticscholar.org/CorpusID:236318084>
27. Herbei, R., Wegkamp, M.H.: Classification with reject option. *Can. J. Stat.* **34**(4), 709–721 (2006)
28. Huang, L., Zhang, C., Zhang, H.: Self-adaptive training: beyond empirical risk minimization. In: NeurIPS (2020)
29. Jones, E., Sagawa, S., Koh, P.W., Kumar, A., Liang, P.: Selective classification can magnify disparities across groups. In: ICLR (2021)
30. Kühne, J., März, C., et al.: Securing deep learning models with autoencoder based anomaly detection. In: PHM Society European Conference, vol. 6, pp. 13–13 (2021)
31. Lee, J.K., et al.: Fair selective classification via sufficiency. In: ICML. Proceedings of Machine Learning Research, vol. 139, pp. 6076–6086. PMLR (2021)
32. Lenders, D., Calders, T.: Learning a fair distance function for situation testing. In: Kamp, M., et al. Machine Learning and Principles and Practice of Knowledge Discovery in Databases, ECML PKDD 2021, Communications in Computer and Information Science, vol. 1524, pp. 631–646. Springer, Cham (2021). [https://doi.org/10.1007/978-3-030-93736-2\\_45](https://doi.org/10.1007/978-3-030-93736-2_45)
33. Pedreschi, D., Ruggieri, S., Turini, F.: Discrimination-aware data mining. In: KDD, pp. 560–568. ACM (2008)
34. Pedreschi, D., Ruggieri, S., Turini, F.: Measuring discrimination in socially-sensitive decision records, pp. 581–592. SIAM (2009)
35. Perini, L., Davis, J.: Unsupervised anomaly detection with rejection. In: NeurIPS (2023)
36. Pessach, D., Shmueli, E.: A review on fairness in machine learning. *ACM Comput. Surv. (CSUR)* **55**(3), 1–44 (2022)
37. Pugnana, A., Perini, L., Davis, J., Ruggieri, S.: Deep neural network benchmarks for selective classification. arXiv preprint [arXiv:2401.12708](https://arxiv.org/abs/2401.12708) (2024)
38. Pugnana, A., Ruggieri, S.: AUC-based selective classification. In: AISTATS, vol. 206, pp. 2494–2514. PMLR (2023)
39. Pugnana, A., Ruggieri, S.: A model-agnostic heuristics for selective classification. In: AAAI, pp. 9461–9469. AAAI Press (2023)
40. Schreuder, N., Chzhen, E.: Classification with abstention but without disparities. In: UAI. Proceedings of Machine Learning Research, vol. 161, pp. 1227–1236. AUAI Press (2021)

41. Selbst, A.D., Boyd, D., Friedler, S.A., Venkatasubramanian, S., Vertesi, J.: Fairness and abstraction in sociotechnical systems. In: Proceedings of the Conference on Fairness, Accountability, and Transparency, pp. 59–68 (2019)
42. Shah, A., et al.: Selective regression under fairness criteria. In: ICML. Proceedings of Machine Learning Research, vol. 162, pp. 19598–19615. PMLR (2022)
43. Stevens, A., Deruyck, P., Veldhoven, Z.V., Vanthienen, J.: Explainability and fairness in machine learning: improve fair end-to-end lending for kiva. In: SSCI, pp. 1241–1248. IEEE (2020)
44. Thanh, B.L., Ruggieri, S., Turini, F.: k-nn as an implementation of situation testing for discrimination discovery and prevention. In: KDD, pp. 502–510. ACM (2011)
45. Wachter, S., Mittelstadt, B.D., Russell, C.: Why fairness cannot be automated: bridging the gap between EU non-discrimination law and AI. *Comput. Law Secur. Rev.* **41**, 105567 (2021)
46. Wang, A., Ramaswamy, V.V., Russakovsky, O.: Towards intersectionality in machine learning: Including more identities, handling underrepresentation, and performing evaluation. In: Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency, pp. 336–349 (2022)
47. Wang, X., Yiu, S.: Classification with rejection: scaling generative classifiers with supervised deep infomax. In: IJCAI, pp. 2980–2986. [ijcai.org](http://ijcai.org) (2020)
48. Zafar, M.B., Valera, I., Rogríguez, M.G., Gummadi, K.P.: Fairness constraints: mechanisms for fair classification. In: Artificial intelligence and statistics, pp. 962–970. PMLR (2017)