An Explainable Feature Selection Approach for Fair Machine Learning

Zhi Yang[1], Ziming Wang[1], Changwu Huang[1(✉)], and Xin Yao[1,2]

[1] Guangdong Provincial Key Laboratory of Brain-inspired Intelligent Computation, Department of Computer Science and Engineering, Southern University of Science and Technology, Shenzhen, China
[2] Research Institute of Trustworthy Autonomous Systems, Southern University of Science and Technology, Shenzhen, China
huangcw3@sustech.edu.cn

Abstract. As machine learning (ML) algorithms are extensively adopted in various fields to make decisions of importance to human beings and our society, the fairness issue in algorithm decision-making has been widely studied. To mitigate unfairness in ML, many techniques have been proposed, including pre-processing, in-processing, and post-processing approaches. In this work, we propose an explainable feature selection (ExFS) method to improve the fairness of ML by recursively eliminating features that contribute to unfairness based on the feature attribution explanations of the model's predictions. To validate the effectiveness of our proposed ExFS method, we compare our approach with other fairness-aware feature selection methods on several commonly used datasets. The experimental results show that ExFS can effectively improve fairness by recursively dropping some features that contribute to unfairness. The ExFS method generally outperforms the compared filter-based feature selection methods in terms of fairness and achieves comparable results to the compared wrapper-based feature selection methods. In addition, our method can provide explanations for the rationale underlying this fairness-aware feature selection mechanism.

Keywords: Fairness in machine learning · Group fairness · Feature selection · Feature attribution explanation · Ethics of AI.

1 Introduction

Machine learning (ML) algorithms are increasingly adopted in more and more fields and have brought significant impact on our daily lives and society. However, despite the advantages brought by adopting ML models, there is plenty of evidence of discriminatory behavior in algorithmic decision-making. For instance, the software product COMPAS used to predict future criminals was found to be biased against blacks [1]. Many other similar behaviors and findings have also been exposed in other areas and applications [15]. Thus, fairness in ML has received considerable attention and discussions in the last decades [9].

The widespread concerns about algorithmic fairness have led to growing interest in fairness-aware ML. Hence, many different measurements of fairness have been formalized [5], and different approaches have been proposed to mitigate the unfairness of ML models. According to the model development stage in which the mitigation techniques are adopted, the existing approaches are usually categorized into pre-processing, in-processing, and post-processing methods. Pre-processing approaches try to adjust or transform the training data for removing the underlying bias in the data before feeding it to an ML algorithm [15]. In-processing methods directly account for fairness during the model design stage usually by modifying ML algorithms to address discrimination during the model training phase [22]. Post-processing techniques dedicate to calibrating the predictions of a model after model training to make decisions fairer [17].

Although there are many different fairness-enhancing techniques, each type of method shows its advantages and limitations and there was no conclusively dominating method [17]. The existing pre-processing methods usually either do not consider the fairness measurements explicitly or are limited to the type of bias they can handle. In-processing mechanisms need to modify the downstream ML algorithms, which is nontrivial and requires rich knowledge and experience. Since post-processing approaches are applied to the relatively late stage of the ML process, these methods typically obtain inferior results [23]. Additionally, the existing methods lack explainability for their fairness enhancement mechanisms.

This work focuses on fairness-aware feature selection (FS), which is a type of pre-processing method that aims at mitigating unfairness by selecting a suitable subset of features to train models. We proposed an explainable feature selection (ExFS) approach to mitigate the unfairness by dropping or eliminating the features that contribute most to the unfairness of the model's prediction in an iterative manner, based on an explainable artificial intelligence (XAI) approach. The experiments on several commonly used datasets and ML models show that our proposed method can enhance the fairness of the used model effectively and efficiently. The main contributions of this work are: 1) We implement a method to explain the prediction of the black-box model by constructing an explainable boosting machine (EBM) surrogate model; 2) We propose an approach to explain which features contribute to the unfairness of a model; 3) We design an explainable feature selection (ExFS) method for mitigating unfairness.

The remainder of this paper is organized as follows. Section 2 introduces the commonly used fairness measurements and some related approaches for mitigating unfairness. The proposed explainable feature selection (ExFS) method is described in Section 3. Section 4 presents the experimental studies. Finally, the paper is briefly concluded in Section 5.

2 Related Work

In this section, we first give the problem setting investigated in this work. Then, some commonly used group fairness measurements and fairness-aware feature selection approaches are introduced under this setting.

2.1 Problem Setting

We consider the most commonly investigated problem in fairness-aware ML literature [11], that is, the binary classification problem that aims to learn a mapping function between user feature vectors $\mathbf{x} \in \mathbb{R}^d$ and class labels $y \in \{0, 1\}$. This task is often achieved by finding a model or classifier $f : \mathbb{R}^d \mapsto \mathbb{R}$ based on the training set $\mathcal{D} = \{(\mathbf{x}^{(i)}, y^{(i)})\}_{i=1}^N$ (where $\mathbf{x}^{(i)} = [x_1^{(i)}, \cdots, x_d^{(i)}] \in \mathbb{R}^d$ are feature vectors and $y^{(i)} \in \{0, 1\}$ are the corresponding labels) such that given a feature vector $\mathbf{x}$ with unknown label y, the classifier can predict its label $\hat{y} = f(\mathbf{x})$. In the context of fairness-aware ML, each $\mathbf{x}$ also has an associated sensitive attribute $s \in \mathcal{S}$ (e.g., sex, race) that indicates the group membership of a user and the model f also needs to be fair with respect to the sensitive attribute. Actually, there can be multiple sensitive attributes. Here we consider a single sensitive attribute case (e.g., the gender of each user $s = \{male, female\}$) and use s_a and s_b to denote two different groups associated with the sensitive attribute. That is, each training data instance $\left(\mathbf{x}^{(i)}, y^{(i)}\right) \in \mathcal{D}$ has an associate sensitive feature value $s^{(i)} \in \{s_a, s_b\}$. The goal of fairness-ware ML is to learn a model f that can provide accurate predictions while satisfying fairness requirements.

We introduce some additional notations used in work. The subsets of training dataset $\mathcal{D}$ with values $s = s_a$ and $s = s_b$ are denoted as $\mathcal{D}_a = \{(\mathbf{x}^{(i)}, y^{(i)}) \in \mathcal{D} | s^{(i)} = s_a\}$ and $\mathcal{D}_b = \{(\mathbf{x}^{(i)}, y^{(i)}) \in \mathcal{D} | s^{(i)} = s_b\}$, respectively. Let $X_k = [x_k^{(1)}, \cdots, x_k^{(N)}]^T$ $(k = 1, \cdots, d)$ be the k-th feature of the training dataset $\mathcal{D}$, and $S = [s^{(1)}, \cdots, s^{(N)}]^T$ denotes the sensitive attribute associated with $\mathcal{D}$.

2.2 Fairness Measurements

Generally, fairness means the absence of any bias towards individuals or groups based on their inherent or acquired characteristics [15]. Different types of measures for fairness have been proposed, including group and individual fairness [15]. Below, we introduce some widely used group fairness measures.

- Demographic Parity (DP) [4] requires the positive prediction rates across different sensitive groups should be the same, which is evaluated as:

$$m_{DP} = \left| P(\hat{y} = 1 | s = s_a) - P(\hat{y} = 1 | s = s_b) \right|. \tag{1}$$

- Equal Opportunity (EOp) [8] requires the true-positive rates across different groups should be the same, which is computed as:

$$m_{EOp} = \left| P(\hat{y} = 1 | s = s_a, y = 1) - P(\hat{y} = 1 | s = s_b, y = 1) \right|. \tag{2}$$

- Equalized Odds (EOd) [8] requires both the false-positive and true-positive rates across different groups should be the same, which is assessed as:

$$\begin{aligned} m_{EOd} = &\left| P(\hat{y} = 1 | s = s_a, y = 0) - P(\hat{y} = 1 | s = s_b, y = 0) \right| \\ &+ \left| P(\hat{y} = 1 | s = s_a, y = 1) - P(\hat{y} = 1 | s = s_b, y = 1) \right|. \end{aligned} \tag{3}$$

2.3 Fairness-Aware Feature Selection (FS)

Feature selection (FS) is an important method to optimize the performance of ML by selecting a suitable feature subset. FS methods are usually categorized into filter, wrapper, and embedded approaches [24]. Recently, there is a growing body of work that uses FS to improve the fairness of ML [6, 18], which is referred to as fairness-aware FS [10]. Both filter and wrapper methods have been adopted for mitigating unfairness in ML.

Fairness-Aware Filter FS It is well known that bias caused by proxy features of the sensitive attribute is one of the main causes of unfairness in ML [17]. Hence, fairness-aware filter approaches intend to identify features that are highly related to the sensitive attribute (i.e., the proxy features) and then drop these features before training a model. Based on this idea, the Pearson correlation coefficient (PCC) and mutual information (MI) [7] can be used to measure the correlation between each feature X_k $(k = 1, \cdots, d)$ and the sensitive attribute S.

Fairness-Aware Wrapper FS This category of methods directly incorporates fairness measures into its objective when evaluating the goodness of the selected subset features. According to the number of objectives, fairness-aware wrapper approaches can be divided into single and multiple objective methods. In single-objective wrapper approaches, the performance (e.g., accuracy, F1-score) and fairness measures (e.g., DP, EOp) of the model are combined to form a single objective that guides the FS process [3], or only the fairness measure is taken as the objective to be optimized [18]. As for multi-objective wrapper methods, both fairness measures and performance metrics are considered as different objectives to be optimized during the FS [18] so as to obtain a set of Pareto optimal solutions. In [18], both single and multi-objective wrapper approaches have been investigated.

However, the above-described fairness-aware filter and wrapper FS approaches both suffer from their drawbacks. On the one hand, a filter method is computationally efficient but its performance may be inferior to a wrapper method due to not considering the adopted model. On the other hand, wrapper methods usually can provide good results but involve high computational costs. Furthermore, neither filter nor wrapper fairness-aware FS approaches can offer the rationale or cause why removing some features can lead to fairness enhancement.

This has motivated us to design the ExFS method which eliminates features based on the feature attribution explanations to the fairness measure. Our proposed method not only utilize the information or knowledge learned by the model but also provides explanations for the underlying reason why dropping the identified features can improve the fairness of the model.

3 Explainable Feature Selection for Mitigating Unfairness

In this section, we propose an explainable feature selection (ExFS) method for mitigating unfairness. Firstly, the used feature attribution explanation method is introduced. Then, we describe the procedure of the ExFS method.

3.1 Feature Attribution Explanation Method

Explainable artificial intelligence (XAI) is an attractive and rapidly developing research area since AI models are increasingly applied in high-stake domains [9]. During the past decades, numerous XAI methods have been proposed to explain the decisions of ML models [2]. With the goal of explaining which feature(s) contribute to the unfairness, we focus on a kind of XAI techniques known as feature attribution explanation (FAE) methods. FAE methods are popular XAI approaches that compute the attribution of input features to the model's output and provide a per-feature attribution score to represent its importance [25].

There are many popular FAE methods, including SHAP [14], LIME [19], EBM [16], etc. SHAP [14] is a post-hoc approach that can provide both global and local explanations. SHAP computes feature importances by removing features in a game-theoretic framework which leads to expensive computational costs. LIME [19] is also a post-hoc approach. It provides local explanations through the perturbation method to identify the importance of each input feature. While EBM [16] is an intrinsic approach that can provide both global and local explanations, it requires a cheap computational cost than others. Thus, we choose EBM as the FAE method used in this work.

EBM belongs to the family of generalized additive models (GAMs) and can be formulated in the following form [16]:

$$g(\mathbf{x}) = \beta_0 + \sum f_k\left(x_k\right), \tag{4}$$

where f_k is the shape function of k-th feature that EBM learns through modern ML techniques such as bagging and gradient boosting. EBM is highly intelligible and explainable because the contribution of each feature to a prediction can be revealed by $f_k\left(x_k\right)$ and the term contribution of each feature can be sorted and visualized to show which features had the most impact on the prediction [12].

3.2 Explainable Feature Selection (ExFS) Method

Using EBM to Explain Black-box Model's Predictions : To leverage the explainability of EBM, we use an EBM as a surrogate model to explain the predictions of other black-box models, such as random forest and neural network models. The goal is to use an EBM g to simulate or mimic the input-output mapping of the trained model f. The procedure for constructing an EBM surrogate model is illustrated in Fig. 1. The only difference between building an EBM surrogate and training an EBM directly on the training dataset is that the output (the predicted probability) $\hat{Y}$ of the trained black-box model f are taken

as targets rather than the ground-truth labels Y in the dataset when constructing the EBM surrogate. Then, the EBM surrogate model g can be used to provide explanations for the prediction of model f at any input.

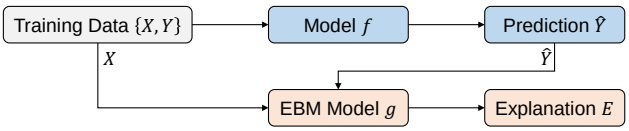


Fig. 1. Illustration of constructing an EBM surrogate model to explain another model.

Calculating the Feature Attributions for Fairness Measurement : Let $e(\mathbf{x}) \in \mathbb{R}^d$ denotes the explanation of prediction provided by EBM, which is a vector of contribution score or importance score of each input feature, at input vector $\mathbf{x}$. We use $E_a = \{e\left(\mathbf{x}^{(i)}\right) \mid \mathbf{x}^{(i)} \in \mathcal{D}_a\}$, $E_b = \{e\left(\mathbf{x}^{(j)}\right) \mid \mathbf{x}^{(j)} \in \mathcal{D}_b\}$ represent the explanation sets for the two subsets (or groups) $\mathcal{D}_a$ and $\mathcal{D}_b$ of dataset $\mathcal{D}$ associated with the sensitive attribute. Based on the individual prediction's explanations, we can attribute DP fairness measurement m_{DP} (in Eq. 1) back to each of the input features [13, 21]. We calculate how each feature contribute to the m_{DP} by,

$$FA_{DP} = mean\left(E_a\right) - mean\left(E_b\right) = \frac{\sum_{\mathbf{x}^{(i)} \in \mathcal{D}_a} e\left(\mathbf{x}^{(i)}\right)}{|\mathcal{D}_a|} - \frac{\sum_{\mathbf{x}^{(j)} \in \mathcal{D}_b} e\left(\mathbf{x}^{(j)}\right)}{|\mathcal{D}_b|}. \quad (5)$$

The feature attribution for other group fairness measurements (EOp and EOd) can be derived in a similar way as that for DP described above.

Eliminating Features based on Explanations : The achieved FA_{DP} is a vector that includes the contributions of each feature to the DP measure, and ΣFA_{DP} indicates the DP value. The larger value of items in FA_{DP} vector indicates the corresponding features contribute more to the fairness measure, i.e., causing unfairness. Hence, we can eliminate features that have large contribution scores to fairness measure for reducing unfairness. In our ExFS method, we recursively eliminate the feature that contributes mostly to the computed fairness measure. The procedure of the ExFS approach is described in Algorithm 1.

4 Experimental Study

In this section, we first present the setup of our experiments and then demonstrate the effectiveness of our method by comparing it with four state-of-the-art methods on three datasets with three different ML models.

4.1 Experimental Setting

Compared Approaches : We compare our approach with four fairness-aware FS methods, as described in Section 2.3, namely FS based on Mutual Information (FS-MI), FS based on Pearson Correlation Coefficient (FS-PCC), FS using Genetic Algorithm (FS-GA) [18], and FS using NSGA-II (FS-NSGA-II) [18].

Algorithm 1 The Procedure of Explainable Feature Selection (ExFS) Method

Input: Training dataset $\mathcal{D} = \{(\mathbf{x}^{(i)}, y^{(i)})\}_{i=1}^{N}$, Classification model f, fairness measure m (e.g., m_{DP}, m_{EOp}, or m_{EOd}), Unfairness tolerance ϵ.
Output: Trained model f that satisfying the fairness requirement.
1: Train the initial model f on dataset $\mathcal{D}$ with the initial set of features.
2: Evaluate the fairness measure m of the trained model f on dataset $\mathcal{D}$.
3: **while** $m > \epsilon$ and $\#Features > 1$ **do**
4: Construct an EBM surrogate model g for explaining the predictions of model f.
5: Calculate the feature attributions FA_m for fairness measure m.
6: Eliminate the feature that has largest contribution to m according to FA_m.
7: Retrain the model f on the dataset with remaining features.
8: Evaluate the fairness measure m of the retrained model f.
9: **end while**
10: Return the model f.

Datasets : We validate the proposed method on three commonly used datasets of binary classification tasks [11]: Adult, Dutch, and Compas, whose sensitive attributes are sex, sex, and race, respectively.

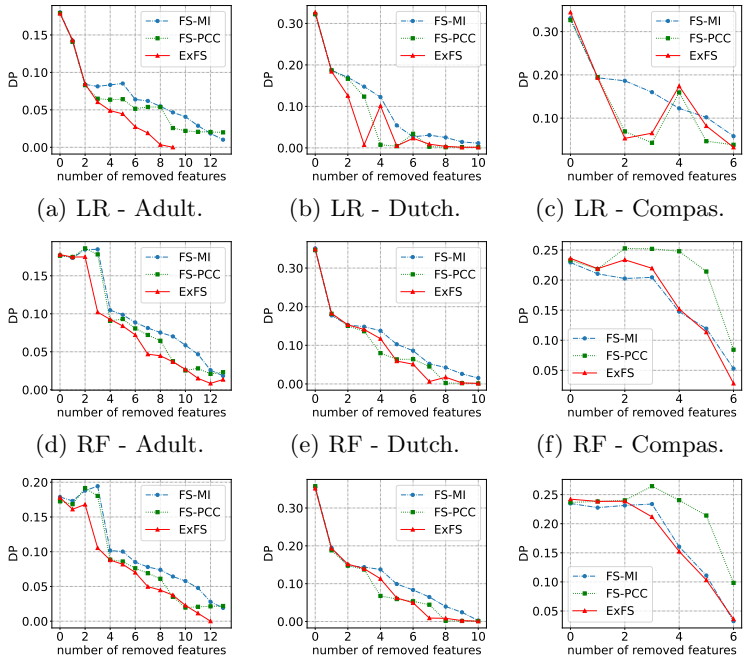
Models : Experiments were performed on three models: Logistic Regression (LR) with maximum number of iterations of 1000, Random Forest (RF) with 10 estimators and max-depth of 20 [21], and Multi-layer Perceptron (MLP) with two hidden layers of size 64 and 32 and maximum number of iterations of 200 [20].

Evaluation Metrics : We used three widely used group fairness metrics as evaluation criteria, which have been described in Section 2.2. We randomly split each dataset into training and test sets with a ratio of 7:3. All reported results are the average results on the test set obtained from 15 different random splits.

4.2 Experimental Results

Firstly, we conduct a comparative analysis between our approach and the other two filter-based methods, as they all gradually drop features until the fairness measure reaches the threshold. Due to space limitation, we only present the comparison results of FS process for improving the DP metric in Fig. 2. To standardize the comparison, all methods drop the sensitive attribute at the beginning and we set the unfairness tolerance $\epsilon = 0.0$. From Fig. 2, we can see that all methods are effective in reducing the DP value. However, the ExFS method tends to be the most efficient method for improving the DP metric, especially on the Adult dataset. In addition, the ExFS method is ultimately capable of achieving extremely low DP values, which makes it capable of satisfying diverse fairness requirements, including more stringent ones.

Then, we provide explanations to demonstrate the operational mechanism of our method and to explain the reason of fairness measure changes during the ExFS procedure. The FAE graphs from the ExFS method for each step

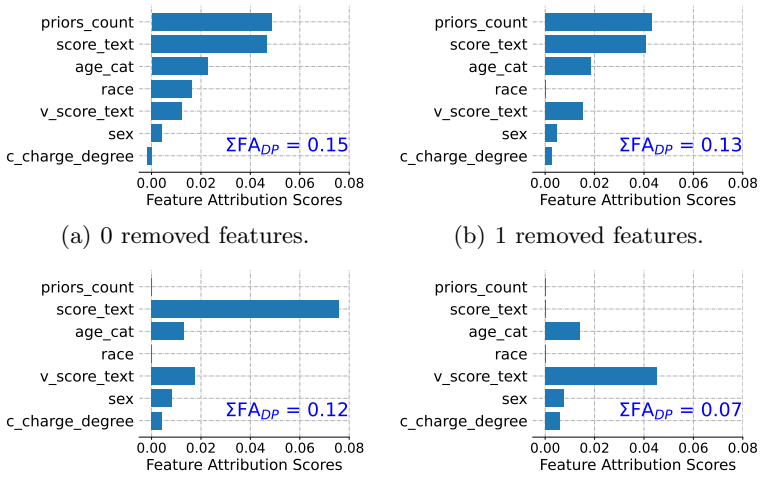


(a) LR - Adult. (b) LR - Dutch. (c) LR - Compas.

(d) RF - Adult. (e) RF - Dutch. (f) RF - Compas.

(g) MLP - Adult. (h) MLP - Dutch. (i) MLP - Compas.

Fig. 2. The comparison results of three filter approaches to enhance DP fairness metric on different datasets using different models.



(a) 0 removed features. (b) 1 removed features.

(c) 2 removed features. (d) 3 removed features.

Fig. 3. Feature attribution explanations for DP on Compas dataset using MLP model. (Firstly, the sensitive attribute 'race' is eliminated, then 'priors_count' is removed, and 'score_text' is eliminated subsequently.)

of the recursive deletion of the first 3 features for the case of Fig. 2 (i) are presented in Fig. 3, showcasing part of the attribution process of the ExFS method. As can be seen in Fig. 3 (b), when removing the second feature, we selected the 'priors_count' feature that contributed the most to the DP based on the attribution scores provided by FA_{DP}. However, after removing this feature, we can see from Fig. 3 (c) that there is no significant change in ΣFA_{DP}. This is due to the sudden increase in the contribution score of the feature 'score_text' (see Fig. 3 (c)) following the removal of 'priors_count'. After eliminating the feature 'score_text', the DP value decrease significantly. This demonstrates that FAE of fairness measure can explain why DP value decreases or not during the feature elimination process. Apparently, the ExFS method not only makes the selection process transparent and understandable but also helps us to analyze the reasons for the results generated by this selection. Furthermore, it can be observed from Fig. 3 that the feature ranking based on FAE is not constant. Hence, depending on the ML model used and other hyperparameter settings, the ranking of features that lead to model unfairness may vary. As mentioned in Section 2.3, the existing filter-based methods are limited to analyzing the relationship between features in the dataset and are unable to observe or adapt to such changes. Fortunately, the proposed ExFS method can improve the fairness of the model by gaining insight into the features that cause unfairness in the model through FAE and removing the corresponding features purposefully.

Lastly, we conduct a comprehensive comparison between the ExFS method and all the compared methods on DP, EOp, and EOd metrics, respectively, and the results are listed in Tab. 1. Specifically, for FS-MI and FS-PCC methods, we trained the models and gradually removed the feature with the largest score with sensitive attribute calculated by MI or PCC, and report the corresponding results of the models with the best DP, EOp and EOd metrics obtained in this process, respectively. For FS-GA, FS-NSGA-II, and ExFS methods, we optimize or attribute DP, EOp, and EOd metrics, respectively. Based on the results presented in Tab. 1, it can be observed that our ExFS approach generally performs better (achieves smaller fairness measurement values) than the two filter-based methods (FS-MI and FS-PCC) on three fairness measurements. At the same time, ExFS achieves comparable results to the two wrapper-based approaches (FS-GA and FS-NSGA-II).

In summary, it can be concluded that our ExFS method generally outperforms the compared filter-based methods in terms of fairness enhancement, and achieves comparable results to the wrapper-based methods. But the wrapper-based methods are somewhat brute-force and black-box approaches. While our ExFs method is transparent and able to provide explanations for the rationale behind removing certain features to achieve fairness enhancement. Furthermore, our method is computationally efficient, which involves a lower computational cost compared to the wrapper-based methods. Since a wrapper-based approach usually requires to evaluate a large number of feature subsets by training a model on each feature subset, while our ExFS method only needs to retrain the model after each feature elimination.

Table 1. The comparison results of all investigated fairness-aware feature selection approaches to enhance different fairness measurements. The number in bold face means the corresponding method achieves the best fairness result.

Dataset	Model	Method	Fairness Measurement		
			DP	EOp	EOd
Adult	LR	FS-MI	0.010 ± 0.011	0.012 ± 0.011	0.015 ± 0.008
		FS-PCC	0.020 ± 0.004	0.015 ± 0.013	0.017 ± 0.014
		FS-GA	**0.000 ± 0.000**	**0.000 ± 0.000**	0.021 ± 0.012
		FS-NSGA-II	**0.000 ± 0.000**	**0.000 ± 0.000**	**0.000 ± 0.000**
		ExFS	**0.000 ± 0.000**	**0.000 ± 0.000**	**0.000 ± 0.000**
	RF	FS-MI	0.018 ± 0.007	0.019 ± 0.016	0.024 ± 0.019
		FS-PCC	0.021 ± 0.003	0.014 ± 0.011	0.019 ± 0.008
		FS-GA	**0.000 ± 0.000**	0.054 ± 0.021	0.022 ± 0.016
		FS-NSGA-II	**0.000 ± 0.000**	0.121 ± 0.027	**0.000 ± 0.000**
		ExFS	0.008 ± 0.003	**0.012 ± 0.008**	0.017 ± 0.007
	MLP	FS-MI	0.019 ± 0.008	0.022 ± 0.015	0.021 ± 0.013
		FS-PCC	0.020 ± 0.005	0.016 ± 0.009	0.020 ± 0.010
		FS-GA	**0.000 ± 0.000**	**0.000 ± 0.000**	0.022 ± 0.013
		FS-NSGA-II	**0.000 ± 0.000**	0.111 ± 0.221	0.023 ± 0.011
		ExFS	**0.000 ± 0.000**	**0.000 ± 0.000**	**0.000 ± 0.000**
Dutch	LR	FS-MI	0.011 ± 0.019	0.001 ± 0.002	0.001 ± 0.003
		FS-PCC	0.001 ± 0.001	0.006 ± 0.004	0.009 ± 0.004
		FS-GA	0.003 ± 0.001	**0.000 ± 0.000**	**0.000 ± 0.000**
		FS-NSGA-II	**0.000 ± 0.000**	0.010 ± 0.014	**0.000 ± 0.000**
		ExFS	0.001 ± 0.002	0.002 ± 0.005	0.011 ± 0.005
	RF	FS-MI	0.015 ± 0.015	0.023 ± 0.024	0.043 ± 0.032
		FS-PCC	0.001 ± 0.001	0.008 ± 0.004	0.009 ± 0..004
		FS-GA	**0.000 ± 0.000**	**0.000 ± 0.000**	0.001 ± 0.001
		FS-NSGA-II	**0.000 ± 0.000**	**0.000 ± 0.000**	**0.000 ± 0.000**
		ExFS	0.001 ± 0.001	0.002 ± 0.004	0.010 ± 0.004
	MLP	FS-MI	0.002 ± 0.004	0.018 ± 0.025	0.039 ± 0.028
		FS-PCC	**0.001 ± 0.001**	**0.003 ± 0.003**	**0.004 ± 0.006**
		FS-GA	**0.001 ± 0.002**	0.025 ± 0.034	0.046 ± 0.025
		FS-NSGA-II	**0.001 ± 0.002**	0.045 ± 0.017	0.044 ± 0.026
		ExFS	**0.001 ± 0.001**	**0.003 ± 0.004**	0.005 ± 0.005
Compas	LR	FS-MI	0.058 ± 0.047	0.044 ± 0.051	0.104 ± 0.075
		FS-PCC	0.038 ± 0.049	0.038 ± 0.029	0.063 ± 0.055
		FS-GA	**0.000 ± 0.000**	0.030 ± 0.042	**0.000 ± 0.000**
		FS-NSGA-II	**0.000 ± 0.000**	0.053 ± 0.033	0.070 ± 0.087
		ExFS	0.033 ± 0.047	**0.018 ± 0.037**	0.018 ± 0.045
	RF	FS-MI	0.053 ± 0.043	0.053 ± 0.038	0.082 ± 0.085
		FS-PCC	0.084 ± 0.014	0.102 ± 0.028	0.153 ± 0.030
		FS-GA	**0.005 ± 0.013**	0.077 ± 0.023	**0.023 ± 0.042**
		FS-NSGA-II	0.014 ± 0.029	0.074 ± 0.037	0.055 ± 0.064
		ExFS	0.028 ± 0.036	**0.047 ± 0.046**	0.074 ± 0.066
	MLP	FS-MI	0.032 ± 0.044	0.035 ± 0.039	0.090 ± 0.083
		FS-PCC	0.098 ± 0.022	0.093 ± 0.017	0.159 ± 0.042
		FS-GA	**0.005 ± 0.017**	**0.020 ± 0.035**	0.048 ± 0.061
		FS-NSGA-II	0.029 ± 0.029	0.056 ± 0.035	**0.024 ± 0.040**
		ExFS	0.036 ± 0.041	0.036 ± 0.044	0.062 ± 0.082

5 Conclusion

In this paper, we proposed an explainable feature selection (ExFS) approach that is capable of explaining and mitigating unfairness in ML models. The results of our experiments demonstrate the effectiveness of our approach in improving the fairness of ML models. Our proposed ExFS method is transparent and is able to provide explanations for the rationale of why removing some features can lead to fairness enhancement. Furthermore, ExFS is computationally efficient, which requires a lower computational cost compared to wrapper-based methods. The ExFS method fills the gap of the lack of explainability in the investigated fairness-aware feature selection approaches. But, it should be noted that there are still challenges and limitations to the trade-off between fairness and performance of the ML models. In future work, we will focus on how to achieve a trade-off between performance and fairness based on the explanations of the predictions and the fairness measures.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (Grant No. 62250710682), the Guangdong Provincial Key Laboratory (Grant No. 2020B121201001), the Program for Guangdong Introducing Innovative and Entrepreneurial Teams (Grant No.2017ZT07X386), the Shenzhen Science and Technology Program (Grant No.KQTD2016112514355531), and the Research Institute of Trustworthy Autonomous Systems.

References

1. Angwin, J., Larson, J., Mattu, S., Kirchner, L.: Machine bias (2016), https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing
2. Arrieta, A.B., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., García, S., Gil-López, S., Molina, D., Benjamins, R., et al.: Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. Information Fusion **58**, 82–115 (2020)
3. Dorleon, G., Megdiche, I., Bricon-Souf, N., Teste, O.: Feature selection under fairness constraints. In: Proceedings of the 37th ACM/SIGAPP Symposium on Applied Computing. pp. 1125–1127 (2022)
4. 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)
5. Gajane, P., Pechenizkiy, M.: On formalizing fairness in prediction with machine learning. arXiv preprint arXiv:1710.03184 (2017)
6. 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. 11. Barcelona, Spain (2016)
7. Guyon, I., Elisseeff, A.: An introduction to variable and feature selection. Journal of Machine Learning Research **3**, 1157–1182 (2003)

8. Hardt, M., Price, E., Srebro, N.: Equality of opportunity in supervised learning. Advances in Neural Information Processing Systems **29** (2016)
9. Huang, C., Zhang, Z., Mao, B., Yao, X.: An overview of artificial intelligence ethics. IEEE Transactions on Artificial Intelligence pp. 1–21 (2022). https://doi.org/10.1109/TAI.2022.3194503
10. Khodadadian, S., Nafea, M., Ghassami, A., Kiyavash, N.: Information theoretic measures for fairness-aware feature selection. arXiv preprint arXiv:2106.00772 (2021)
11. Le Quy, T., Roy, A., Iosifidis, V., Zhang, W., Ntoutsi, E.: A survey on datasets for fairness-aware machine learning. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery **12**(3), 1–59 (2022)
12. Lou, Y., Caruana, R., Gehrke, J., Hooker, G.: Accurate intelligible models with pairwise interactions. In: Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. p. 623–631. KDD '13, Association for Computing Machinery, New York, NY, USA (2013)
13. Lundberg, S.M.: Explaining quantitative measures of fairness. In: Fair & Responsible AI Workshop@ CHI2020 (2020)
14. Lundberg, S.M., Lee, S.I.: A unified approach to interpreting model predictions. Advances in Neural Information Processing Systems **30** (2017)
15. 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)
16. Nori, H., Jenkins, S., Koch, P., Caruana, R.: Interpretml: A unified framework for machine learning interpretability. arXiv preprint arXiv:1909.09223 (2019)
17. Pessach, D., Shmueli, E.: A review on fairness in machine learning. ACM Computing Surveys (CSUR) **55**(3), 1–44 (2022)
18. Rehman, A.U., Nadeem, A., Malik, M.Z.: Fair feature subset selection using multiobjective genetic algorithm. In: Proceedings of the Genetic and Evolutionary Computation Conference Companion. pp. 360–363 (2022)
19. Ribeiro, M.T., Singh, S., Guestrin, C.: "Why should I trust you?": Explaining the predictions of any classifier. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. pp. 1135–1144. Association for Computing Machinery, New York, NY, USA (2016)
20. Singh, M.: Fair classification under covariate shift and missing protected attribute–an investigation using related features. arXiv preprint arXiv:2204.07987 (2022)
21. Thampi, A.: Interpretable AI: Building explainable machine learning systems. Manning Publications Co. (2022)
22. Wan, M., Zha, D., Liu, N., Zou, N.: In-processing modeling techniques for machine learning fairness: A survey. ACM Transactions on Knowledge Discovery from Data **17**(3), 1–27 (2023)
23. Woodworth, B., Gunasekar, S., Ohannessian, M.I., Srebro, N.: Learning non-discriminatory predictors. In: Conference on Learning Theory. pp. 1920–1953. PMLR (2017)
24. Xue, B., Zhang, M., Browne, W.N., Yao, X.: A survey on evolutionary computation approaches to feature selection. IEEE Transactions on Evolutionary Computation **20**(4), 606–626 (2015)
25. Zhou, Y., Booth, S., Ribeiro, M.T., Shah, J.: Do feature attribution methods correctly attribute features? In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 36, pp. 9623–9633 (2022)