ICANN23
32nd International Conference on Artificial Neural Networks

An Explainable Feature Selection Approach for Fair Machine Learning

Zhi Yang, Ziming Wang, Changwu Huang, and Xin Yao

Department of Computer Science and Engineering,
Southern University of Science and Technology, Shenzhen, China

Sep. 2023

Outline

1. Introduction

2. Related Work

3. Explainable Feature Selection (ExFS) for Mitigating Unfairness

4. Experimental Study

5. Conclusion

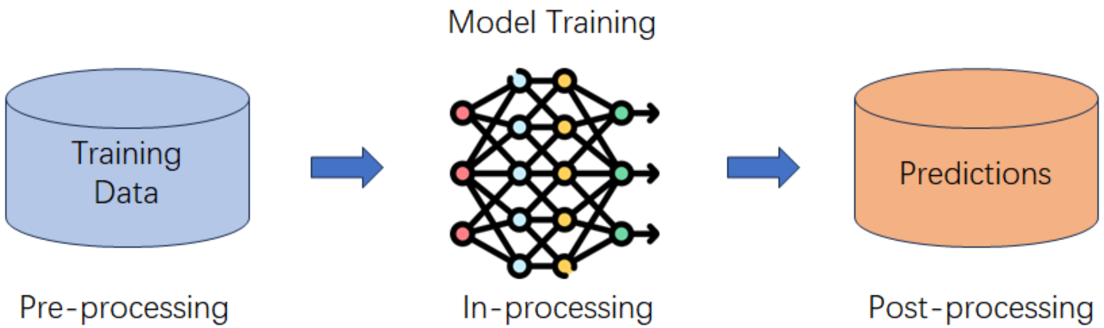
1. Introduction

- Machine learning (ML) algorithms are increasingly adopted in more and more fields and have brought significant impact on our daily lives and society.
- However, discriminatory behavior in algorithmic decision-making hinders the widespread adoption of machine learning. For instance, the software product COMPAS used to predict future criminals was found to be biased against blacks.
- Thus, **fairness in machine learning(ML) has received considerable attention and discussions in the last decades [1].**

[1] 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

1. Introduction

- There are many fairness-enhancing methods. Each type of method shows its advantages and limitations and there was no conclusively dominating method.



- The existing methods all lack **explainability** for fairness-enhancement mechanisms.
- We proposed an **explainable feature selection (ExFS)** approach to mitigate the unfairness based on an explainable artificial intelligence (XAI) approach.

2. Related Work

- Three widely used fairness measurements:
  - Demographic Parity(DP) [2]:

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

  - Equal Opportunity(EOp) [3]:

$$m_{EOp} = |P(\hat{y} = 1|s = s_a, y = 1) - P(\hat{y} = 1|s = s_b, y = 1)|$$

  - Equalized Odds(EOd) [3]:

$$m_{EOd} = |P(\hat{y} = 1|s = s_a, y = 0) - P(\hat{y} = 1|s = s_b, y = 0)| + |P(\hat{y} = 1|s = s_a, y = 1) - P(\hat{y} = 1|s = s_b, y = 1)|$$

Note: s_a and s_b represent the different group between sensitive attribute.

[2] 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)
[3] 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)

2. Related Work

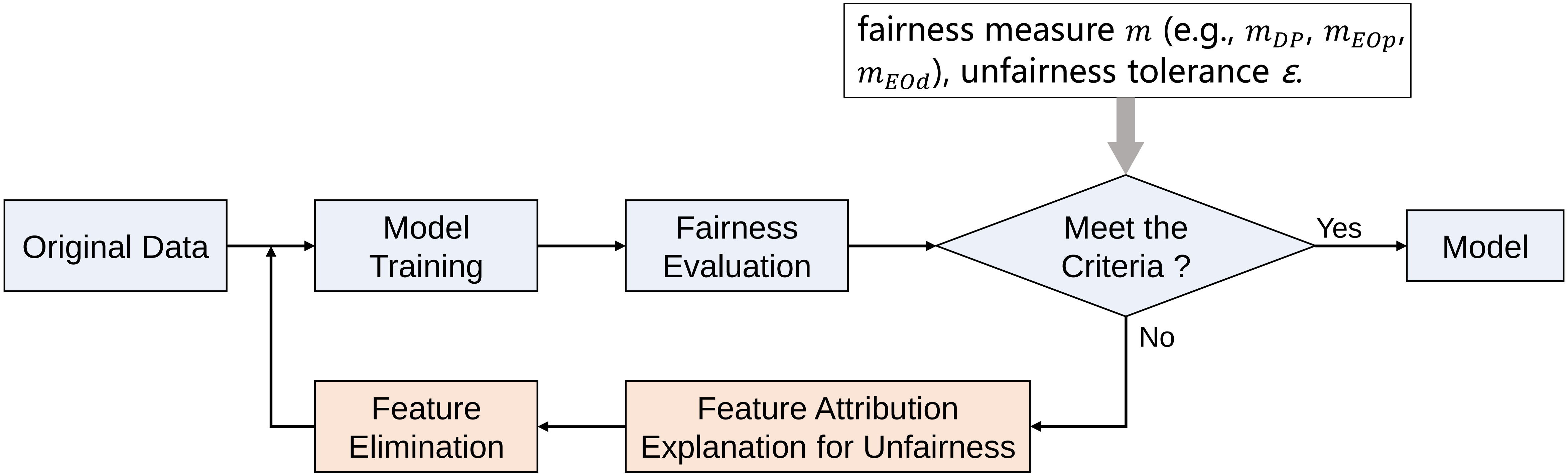
- Recently, there is a growing body of work that uses feature selection(FS) to improve the fairness of ML [4], which is referred to as fairness-aware FS [5].
  - **Fairness-Aware Filter FS:** filter method is computationally efficient, but its performance may be inferior to a wrapper method due to not considering the adopted model.
  - **Fairness-Aware Wrapper FS:** wrapper methods usually can provide good results but involve high computational costs.

Neither filter nor wrapper fairness-aware FS approaches can offer the rationale or cause why removing some features can lead to fairness enhancement.

[4] 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)
[5] Khodadadian, S., Nafea, M., Ghassami, A., Kiyavash, N.: Information theoretic measures for fairness-aware feature selection. arXiv preprint arXiv:2106.00772 (2021)

3. Explainable Feature Selection

■ The Overall Procedure of ExFS



Key Steps:

- Calculate the feature attribution for unfairness, i.e., the contribution of each feature to the unfairness.
- Eliminate the feature that has largest contribution to unfairness, so as to reduce the unfairness.

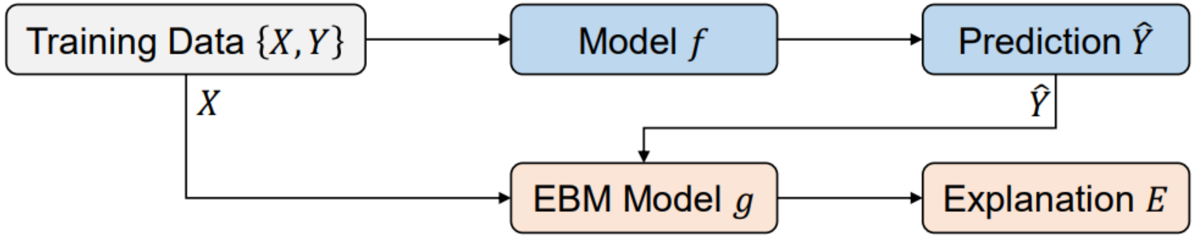
3. Explainable Feature Selection

■ Feature Attribution Explanation Method

- SHAP (Shapley Additive Explanations) is a post-hoc approach that can provide both global and local explanations. It has an expensive computational cost.
- LIME (Local Interpretable Model-Agnostic Explanation) is a post-hoc approach that can provide local explanations.
- EBM (Explainable Boosting Machine) is an intrinsic approach that can provide both global and local explanations, it requires a cheap computational cost than others.

Method	Intrinsic vs. Post-hoc	Computing Cost	Scope
SHAP	Post-hoc	High	Global and Local
LIME	Post-hoc	Medium	Local
EBM	Intrinsic	Low	Global and Local

■ Using EBM to Explain Black-box Model’s Predictions



3. Explainable Feature Selection

■ Calculating the Feature Attributions for Fairness Measurement

- Let $e(\mathbf{x}) \in \mathbb{R}^{\mathrm{d}}$ denotes the explanation of prediction provided by EBM, we use $E_a = \{e(\mathbf{x}^{(i)}) \mid \mathbf{x}^{(i)} \in \mathcal{D}_a\}$, $E_b = \{e(\mathbf{x}^{(j)}) \mid \mathbf{x}^{(j)} \in \mathcal{D}_b\}$ 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 [6, 7], We calculate how each feature contribute to the m_{DP} by,

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

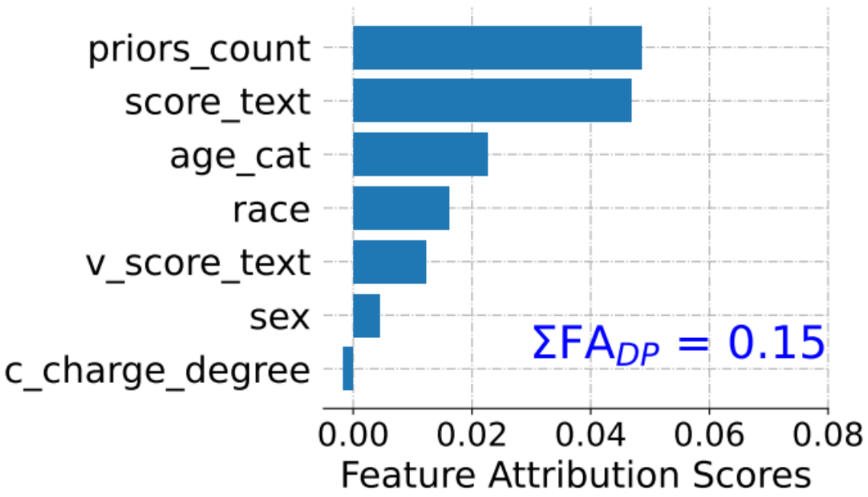
FA_{DP} is a vector that includes the contributions of each feature to the DP measure, and $\sum FA_{DP}$ indicates the DP value. The feature attribution for other group fairness measurements (EOp and EOd) can be derived in a similar way as that for DP described above.

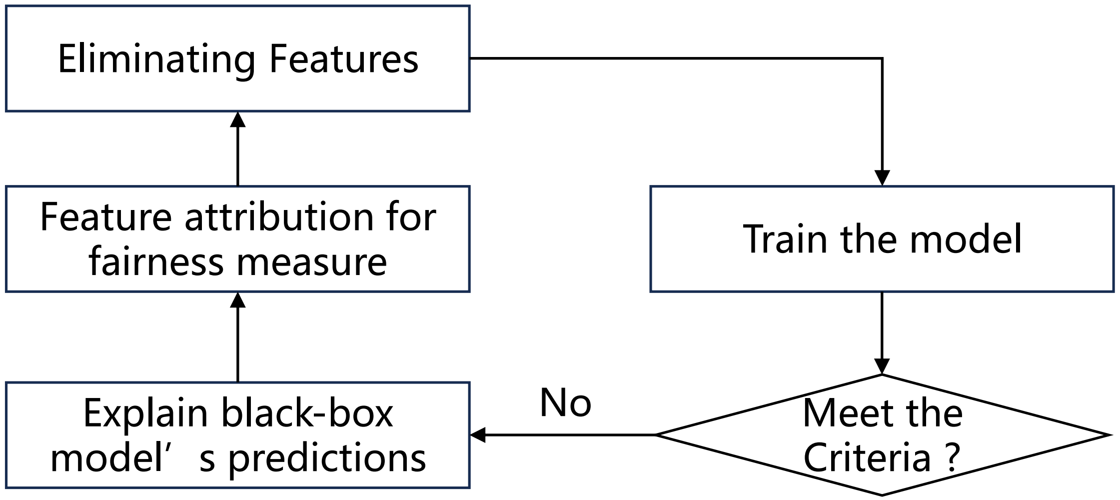
[6] Thampi, A.: Interpretable AI: Building explainable machine learning systems. Manning Publications Co. (2022)
[7] Lundberg, S.M.: Explaining quantitative measures of fairness. In: Fair & Responsible AI Workshop@ CHI2020 (2020)

3. Explainable Feature Selection

■ Eliminating Features based on Explanations

- 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 the feature that have largest contribution score to fairness measure for reducing unfairness.
- We recursively eliminate the feature that contributes mostly to the computed fairness measure.





4. Experimental Study

■ 4.1 Experimental Setting

- **Compared Approaches :**
  - Feature Selection based on Mutual Information (FS-MI)
  - Feature Selection based on Pearson Correlation Coefficient (FS-PCC)
  - Feature Selection using Genetic Algorithm (FS-GA)
  - Feature Selection using NSGA-II (FS-NSGA-II)
- **Datasets :** Adult, Compas, Dutch.
- **Models :** Logistic Regression (LR), Random Forest(RF), Multi-layer Perceptron(MLP).
- **Evaluation Metrics :** DP, EOp and EOd.

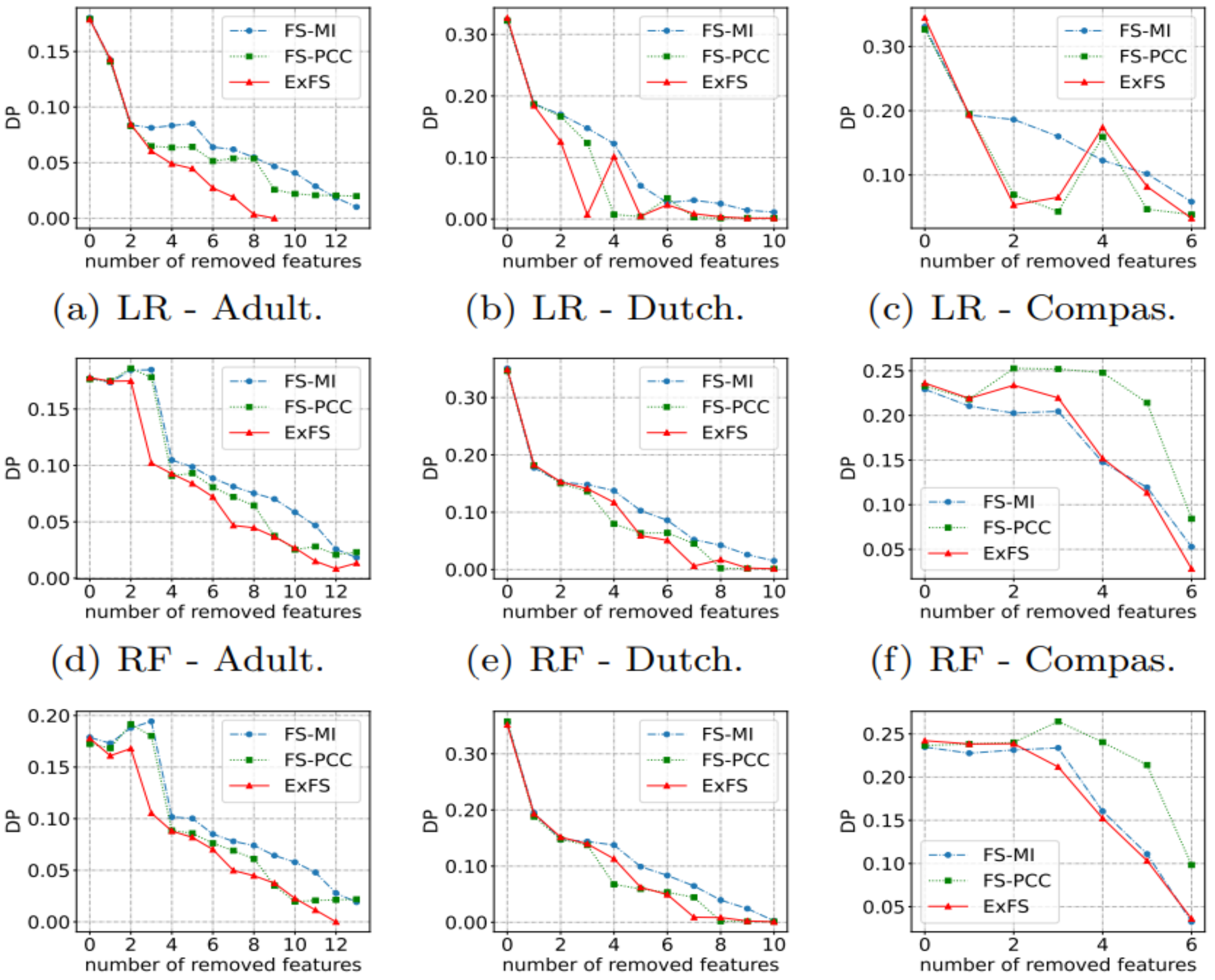
The train and test dataset split ratio is 7:3.

All reported results are the average results on the test set obtained from 15 different random splits.

4. Experimental Study

■ 4.2 Experimental results

We can see that the ExFS method tends to be **the most efficient method** for improving the DP metric, especially on the Adult dataset.



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

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

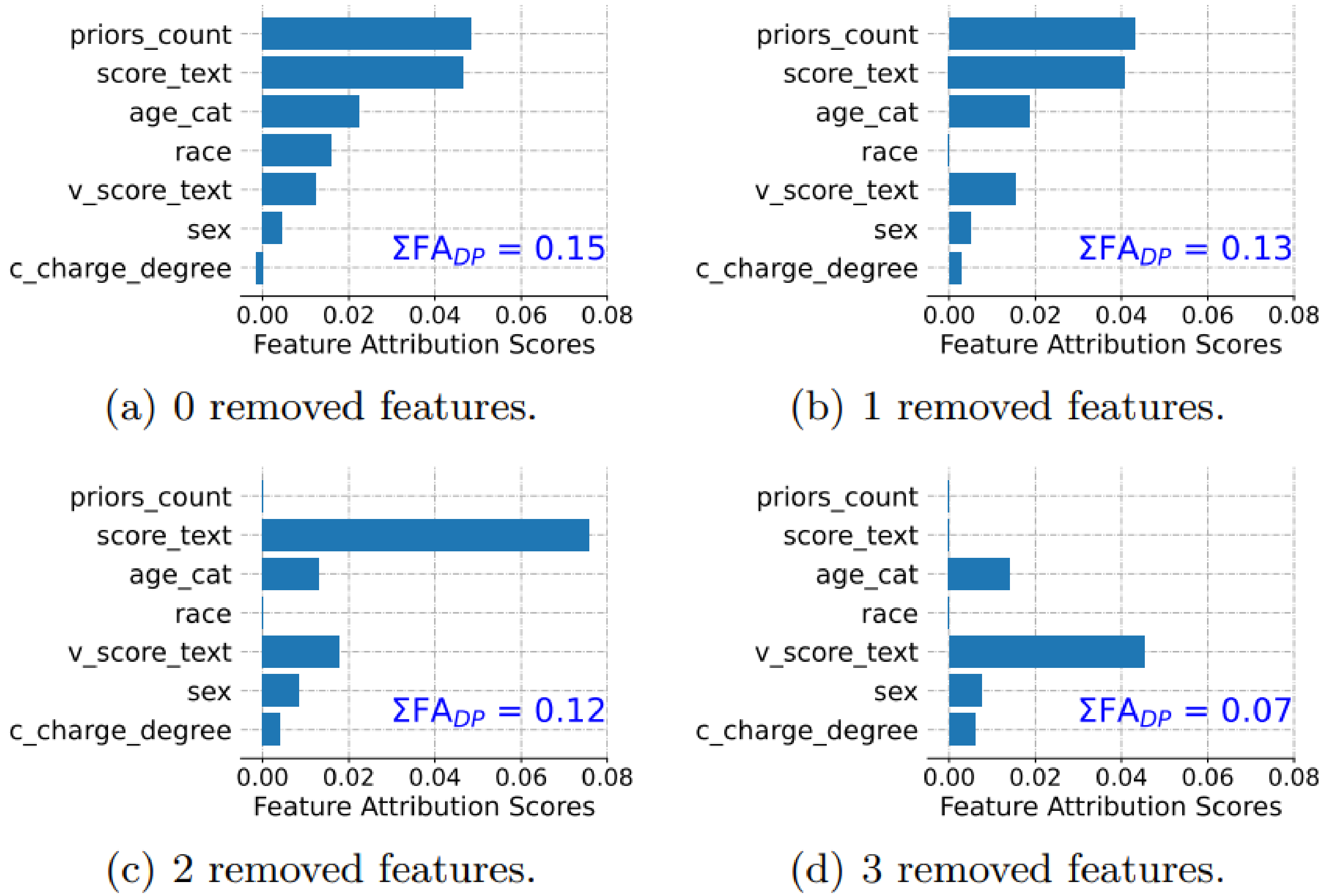
(g) MLP - Adult. (h) MLP - Dutch. (i) MLP - Compas.

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

4. Experimental Study

■ 4.2 Experimental results

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.



Feature attribution explanations for DP on Compas dataset using MLP model.

4. Experimental Study

■ 4.2 Experimental results

ExFS approach generally performs **better** than the two filter-based methods (FS-MI and FS-PCC) on three fairness measurements and achieves **comparable** results to the two wrapper-based approaches (FS-GA and FS-NSGA-II).

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	$\mathbf{0.000 \pm 0.000}$	$\mathbf{0.000 \pm 0.000}$	0.021 ± 0.012
		FS-NSGA-II	$\mathbf{0.000 \pm 0.000}$	$\mathbf{0.000 \pm 0.000}$	$\mathbf{0.000 \pm 0.000}$
		ExFS	$\mathbf{0.000 \pm 0.000}$	$\mathbf{0.000 \pm 0.000}$	$\mathbf{0.000 \pm 0.000}$
	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	$\mathbf{0.000 \pm 0.000}$	0.054 ± 0.021	0.022 ± 0.016
		FS-NSGA-II	$\mathbf{0.000 \pm 0.000}$	0.121 ± 0.027	$\mathbf{0.000 \pm 0.000}$
		ExFS	0.008 ± 0.003	$\mathbf{0.012 \pm 0.008}$	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	$\mathbf{0.000 \pm 0.000}$	$\mathbf{0.000 \pm 0.000}$	0.022 ± 0.013
		FS-NSGA-II	$\mathbf{0.000 \pm 0.000}$	0.111 ± 0.221	0.023 ± 0.011
		ExFS	$\mathbf{0.000 \pm 0.000}$	$\mathbf{0.000 \pm 0.000}$	$\mathbf{0.000 \pm 0.000}$

The comparison results of all investigated fairness-aware feature selection approaches to enhance different fairness measurements.

5. Conclusion

- In summary, we proposed an ExFS approach that is capable of explaining and mitigating unfairness in ML models. The results of our experiments demonstrate that:
  - The effectiveness of our approach in improving the fairness of ML models.
  - 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.

ICANN23

32nd International Conference on Artificial Neural Networks

Thank You !